OCMAE: OCEAN MASKED AUTOENCODERS FOR BIOGEOCHEMISTRY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from sparse labeled data remains a central challenge in machine learning. While large-scale pretraining has driven major advances in domains such as computer vision, its potential for scientific data remains elusive. We introduce Ocean Masked Autoencoders (OcMAE), a self-supervised framework for learning representations from large-scale satellite data to improve downstream prediction tasks in ocean biogeochemistry. OcMAE introduces two key components. First, it discretizes continuous variables into bins, improving representation learning while naturally handling missing values. Second, it leverages large-scale satellite data for self-supervised pretraining, enabling the model to learn prior over biogeochemical processes. We evaluate OcMAE on nine downstream tasks and show consistent improvements over strong baselines, including random forests (RF), extreme gradient boosting (XGB), and tabular prior-data fitted networks (TabPFN). Our results demonstrate that large-scale self-supervised pretraining on satellite data can improve generalization in ocean biogeochemistry.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved strong performance across many domains, including natural language processing (Devlin et al., 2019; Brown et al., 2020), computer vision (He et al., 2016; Dosovitskiy et al., 2021), numerical weather forecasting (Ravuri et al., 2021; Bi et al., 2023; Kochkov et al., 2024), and computational biology (Jumper et al., 2021). Despite these successes, DNNs still struggle in many other domains. In particular, DNNs are often inferior to tree-based models on tabular data (Gorishniy et al., 2021; Shwartz-Ziv & Armon, 2022). This inferiority has been attributed to ensemble learning employed by tree-based models and to the characteristics of tabular data, including limited labeled data, missing values, noise, and heterogeneous variables with high nonlinearity.

Missing values are pervasive in scientific observations. For example, satellite observations often contain gaps due to cloud cover. While tree-based models can naturally accommodate missing values, DNNs typically rely on imputation, which may introduce bias and thereby degrade the quality of learned representations. Moreover, tree-based models efficiently capture highly nonlinear relationships by iterative binary partitioning which is relatively robust to noise in datasets. Inspired by this mechanism, recent studies have shown that feature discretization can also improve the performance of DNNs on tabular data (Müller et al., 2022; Lee et al., 2024).

Directly training DNNs with limited labeled data often leads to poor generalization. Representation learning via large-scale pretraining has emerged as an effective approach to address this challenge. For example, pretraining has achieved remarkable success in natural language processing (Devlin et al., 2019; Brown et al., 2020) and computer vision (Dosovitskiy et al., 2021; Bao et al., 2021; He et al., 2022; Feichtenhofer et al., 2022), and has recently been explored for tabular datasets (Gorishniy et al., 2021; Huang et al., 2020; Somepalli et al., 2021; Jaegle et al., 2021; Müller et al., 2022; Hollmann et al., 2025). In particular, masked autoencoders (MAE), a self-supervised learning (SSL) framework originally developed for images, have recently been applied to satellite observations (Cong et al., 2022; Hong et al., 2024; Tarasiou et al., 2023; Szwarcman et al., 2025; Dawson et al., 2025). These approaches assume that models learn representations (latent) that capture data structures (distributions) from large amounts of unlabeled data (Vincent et al., 2008; 2010). The representations are informative and transferable to downstream tasks through fine-tuning on limited

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

labeled data. However, it remains unclear why reconstructing masked inputs leads to representations that generalize well to downstream tasks, especially for scientific datasets with strong physical constraints and structured dependencies.

In this work, we introduce Ocean Masked Autoencoders (OcMAE; Figure 1), a SSL framework designed for oceanic datasets with missing observations and highly nonlinear relationships between inputs and outputs. Our approach integrates quantile-binning-based feature tokenization with MAE pretrained on large-scale satellite observations. We derive the mathematical foundation underlying the OcMAE, providing theoretical insights into masked reconstruction for ocean biogeochemistry. We evaluate OcMAE on nine prediction tasks spanning carbon, oxygen, organic matter (i.e., metabolism), and nitrogen cycles: surface seawater carbon dioxide (CO_2) partial pressure (pCO_2), surface ocean oxygen (O_2) concentration, air-sea O_2 flux, net community production (NCP), net primary production (NPP), gross primary production (GPP), community respiration (CR), nitrogen (N_2) fixation, and nitrification. We benchmark OcMAE against strong baselines, including random forests (RF) (Breiman, 2001), extreme gradient boosting (XGB) (Chen & Guestrin, 2016), and tabular prior-data fitted networks (TabPFN) (Hollmann et al., 2025).

Our main contributions lie at the intersection of machine learning and ocean science:

- **Tokenized representations for continuous variable:** We introduce a quantile-binning-based feature tokenization strategy that improves representation learning for continuous variables while naturally accommodating missing values.
- **Physical interpretation of masked autoencoders:** We provide a theoretical analysis showing how masked reconstruction can implicitly learn the ocean tracer transport equation.
- **Improved predictive performance for ocean biogeochemical datasets:** OcMAE achieves superior performance compared to strong baselines on four of nine downstream tasks.

2 Related Works

Machine Learning for Oceanic Biogeochemistry. Machine learning algorithms have been widely adopted to scale up key marine biogeochemical variables, including N_2 fixation rate (Tang et al., 2019), NCP (Li & Cassar, 2016), GPP (Huang et al., 2021), O_2 concentration (Lu et al., 2024; Giglio et al., 2018; Ito et al., 2024; Sharp et al., 2023), and air-sea CO_2 flux (Gloege et al., 2022; Landschützer et al., 2015). The performance of these approaches depends on the availability of training data. However, observations are typically sparse and unevenly distributed across the ocean. In addition, separate models are often developed for individual variables despite their shared underlying biogeochemical processes. In contrast, we develop a model pretrained on large-scale satellite observations, capturing shared representations of biogeochemical dynamics in the world's oceans.

DNNs for Tabular Data. Our model is related to attention-based architectures for tabular data. TabTransformer improves representations for categorical variables by encoding them with transformer layers (Huang et al., 2020). Subsequent work extends this approach to continuous variables (Gorishniy et al., 2021). To better capture input-output relationships, recent methods discretize continuous variables and encode both discretized and categorical features before applying transformer layers (Lee et al., 2024). Quantile binning is also employed by XGB to efficiently approximate optimal split points for continuous variables (Chen & Guestrin, 2016). While our framework shares similarities with these approaches (Chen & Guestrin, 2016; Lee et al., 2024), it differs in its objective. It aims to learn representations that generalize across the global ocean through large-scale pretraining on satellite observations.

Physics-Informed DNNs. Imposing structures on machine learning frameworks often improves their learnability and generalization. One such structures is differential equations that describe physical processes (e.g., Navier-Stokes equations). For example, previous studies have shown that integrating parts of differential equations into diffusion models can improve model performance (Xu & Li, 2025). DNNs can also be used to solve differential equations, or systems of differential equations, by incorporating them into the loss function (Raissi et al., 2019; Lu et al., 2021). Existing studies have further demonstrated the promise of integrating DNNs with state estimation (i.e., data assimilation in geoscience) (Reichstein et al., 2019; Cheng et al., 2023). In contrast to these approaches, we aim to learn implicit physical structures from massive unlabeled data, capturing the underlying data-generating processes and improving downstream task performance.

Unsupervised Pretraining. Our approach aligns with a line of research that learns robust data representations through large-scale pretraining (Devlin et al., 2019; Brown et al., 2020). Pretrained models are then adapted to various downstream applications via fine-tuning on labeled data, enhancing model performance (Erhan et al., 2010; Larochelle & Murray, 2011; Germain et al., 2015). Representation learning in self-supervised learning frameworks is commonly achieved by designing loss functions (e.g., contrastive losses) (Hadsell et al., 2006; He et al., 2020) or by constructing pretext tasks (e.g., denoising autoencoders) (Vincent et al., 2008). Our approach belongs to the latter category.

Our approach builds on the MAE, a SSL algorithm inspired by masked language modeling (MLM) (He et al., 2022). MAE has been extended to learn spatiotemporal representations of videos (Feichtenhofer et al., 2022), satellite data (Cong et al., 2022; Szwarcman et al., 2025; Hong et al., 2024), and ocean color (Dawson et al., 2025). Our work is related to BEIT (Bao et al., 2021), which reconstructs tokenized image patches to capture long-range dependencies while discarding fine-grained details. Similar ideas of reconstructing simplified data (i.e., latent variables learned from variational autoencoders (Kingma & Welling, 2014)) have been explored in earlier studies (Ramesh et al., 2021; Chen et al., 2020). Our work is particularly related to Szwarcman et al. (2025); Dawson et al. (2025), which adapt pretrained MAE models to predict terrestrial GPP, chlorophyll (Chl), and NPP. In contrast to Dawson et al. (2025), our model learns representations of spatiotemporal processes by reconstructing tiers of discretized data on pixel level. We derive mathematical foundation underlying our framework.

3 Background

3.1 Transformers

The architecture of transformers consists of an encoder-decoder structure, with both components composed of stacked layers built upon multiheaded self-attention (MSA) (Radford et al., 2019):

$$n^l = LN(h^l), \tag{1}$$

$$a^l = MSA(n^l) + h^l, \tag{2}$$

$$h^{l+1} = MLP(a^l) + a^l, \tag{3}$$

where l is the index of stacked layers, MLP represents multi-layer perception, and LN denotes layer normalization (Ba et al., 2016). More details on transformers can be found in Vaswani et al. (2017).

3.2 Masked Autoencoders

MAE is a variant of the MLM framework, a SSL approach rooted in denoising autoencoders (He et al., 2022; Vincent et al., 2008; 2010). Denoising autoencoders are mathematically equivalent to score matching, which learns the gradient of the log-density with respect to data (Hyvarinen, 2005; Vincent et al., 2008; Vincent, 2011; Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021):

$$\min_{\theta} \mathbb{E}_{q_\sigma(\mathbf{x},\hat{\mathbf{x}})} \left[\frac{1}{2} \left\| \psi(\hat{\mathbf{x}};\theta) - \nabla_{\hat{\mathbf{x}}} \log q_\sigma(\hat{\mathbf{x}} \mid \mathbf{x}) \right\|_2^2 \right], \tag{4}$$

where $\mathbf{x}$ denotes the original data, $\hat{\mathbf{x}}$ is the corrupted data, $\psi(.;\theta)$ is a parametric model, and $q_\sigma(\hat{\mathbf{x}}|\mathbf{x})$ defines the noising process. In MAE, $q_\sigma(\hat{\mathbf{x}}|\mathbf{x})$ corresponds to randomly masking a subset of input values, which are subsequently reconstructed from unmasked (visible) values:

$$\min_{\theta} \mathbb{E}_{q_\sigma(\mathbf{x},\mathbf{M}\circ\mathbf{x})} \left[-\log p_\theta(\mathbf{M} \circ \mathbf{x} \mid (\mathbf{1} - \mathbf{M}) \circ \mathbf{x}) \right], \tag{5}$$

where $\mathbf{M}$ represents masking vector and $\circ$ denotes element-wise multiplication. Therefore, MAE is closely linked to the pseudolikelihood estimation (Besag, 1975).

4 Ocean Masked Autoencoders

4.1 Feature Discretization

Consider a data set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where each observation $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ represents a vector with p entries (i.e., $\mathbf{x}_i \in \mathbb{R}^p$). Feature discretization partitions the range of a continuous variable (e.g., $x(., j)$) into T non-overlapped tiers as follows:

$$B_{t,j} = [b_{t,j}, b_{t+1,j}),\ t = 0, ..., T-1, \tag{6}$$

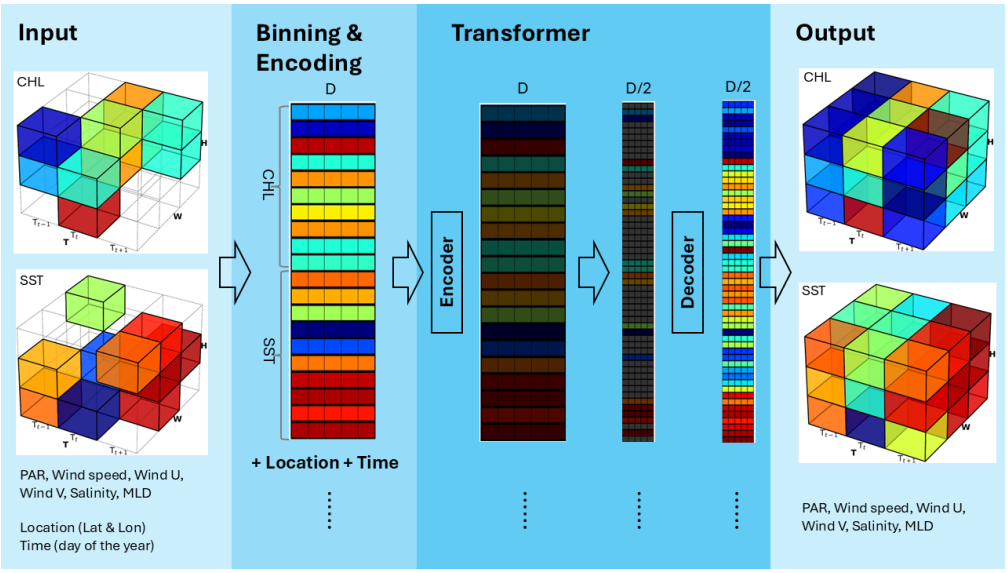

Figure 1: Our proposed framework for learning the robust representations of oceanic biogeochemical processes from large-scale satellite observations. The input variables are discretized to enhance representation capacity and to accommodate missing values. By reconstructing randomly masked values from visible ones, the model is enabled to learn physical processes (e.g., advection and diffusion) and biological dynamics such as the sensitivity of source-sink term to temporal variations in environmental factors.

where t is the index of tiers, and $b_{t,j}$ and $b_{t+1,j}$ denote the lower and upper boundaries, respectively. Collectively, the set of tiers $B_j = \{B_{0,j}, B_{1,j}, ..., B_{T-1,j}\}$ covers the range of $x_{.,j}$. A new observational value $x_{i,j}$ is assigned to the tier $B_{t,j}$ if $b_{t,j} \leq x_{i,j} < b_{t+1,j}$.

One challenge in feature discretization is the selection of tier boundaries. Quantile binning approach partitions a continuous feature into T tiers of equal size:

$$\frac{1}{T} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}(x_{k,j} \in B_{t,j}),\ \forall t \in \{0, ..., T-1\}, \tag{7}$$

where $\mathbf{1}(.)$ is the indicator function. Missing values are assigned to a new tier $B_{T,j}$. A discretized representation with sufficiently many tiers can theoretically approximate a continuous variable to arbitrary precision. In practice, however, the optimal discretization is constrained by a trade-off between bias (coarse bins) and variance (fine bins), further influenced by observational noise and the effective sample size of the training dataset.

4.2 Embedding

Each tier of discretized features is represented as a D-dimensional embedding vector $\mathbf{y}_{t,j} \in \mathbb{R}^D$, obtained from a learnable codebook. Because self-attention discards order information, the embedding vector is further encoded with order information through adding a bias term that is a combination of local (not learnable) and global (learnable) positional encodings. The local positional encoding follows the standard transformer-based approach to encode spatial (i.e., row and column of the pixel in the image cube), temporal (i.e., the order of consecutive 8-day averages), and spectral (i.e., variable) information. It is represented as $\mathbf{e}_{\text{loc}} = [\mathbf{e}_l, \mathbf{e}_c] \in \mathbb{R}^D$, where $\mathbf{e}_{\text{loc}}$ is a concatenation of spatial location $\mathbf{e}_l \in \mathbb{R}^{D/2}$ and a joint spectral-temporal encoding $\mathbf{e}_c \in \mathbb{R}^{D/2}$.

The global positional encoding captures seasonality and geographical location. Seasonality is represented as the day of year based on 8-day composites, treated as a categorical variable, and embedded via a learnable codebook: $\mathbf{e}_s \in \{\mathbf{s}_1, \ldots, \mathbf{s}_{46}\}$, where $\mathbf{s}_i \in \mathbb{R}^D$. Geographical location is encoded using latitude and longitude. Due to computational constraints, we do not learn a separate embedding for each location. Instead, we project them into a D-dimensional vector space:

$\mathbf{e}_g = f(\text{latitude}, \text{longitude})$, where $\mathbf{e}_g \in \mathbb{R}^D$. Finally, the local positional encoding, seasonal embedding, and geographical embedding are added to $\mathbf{y}_{t,j}$:

$$\mathbf{z}_{t,j} = \mathbf{y}_{t,j} + \mathbf{e}_{\text{loc}} + \mathbf{e}_s + \mathbf{e}_g. \tag{8}$$

4.3 Masked Reconstruction for Learning Spatiotemporal Dynamics

Masking can be interpreted as introducing structured corruption to the data, while reconstruction corresponds to learning conditional distributions that approximate the joint data distribution (Germain et al., 2015; Chen et al., 2020; van den Oord et al., 2016) or equivalently, score functions with respect to the data (Vincent et al., 2008; Vincent, 2011; Hyvarinen, 2005). This perspective is conceptually analogous to inferring oceanic biogeochemical processes governed by the tracer transport equation:

$$\frac{\partial C}{\partial t} = -\nabla \cdot (UC) + \nabla \cdot (K \nabla C) + S, \tag{9}$$

where C denotes tracer concentration, U denotes the velocity field, K is the diffusivity, S denotes the source-sink term, $\partial C/\partial t$ represents the local rate of change, $\nabla \cdot (UC)$ denotes the advection, and $\nabla \cdot (K \nabla C)$ denotes the diffusion.

When trained on spatiotemporal fields, masked reconstruction encourages the model to learn statistical dependencies that reflect a combination of transport and biogeochemical processes. In this sense, pretraining can be viewed as implicitly capturing a discrete approximation to the dynamics governing tracer evolution and inducing a data-driven prior over spatiotemporal dynamics. For example, local variability in Chl can be expressed as:

$$\mathbf{A}^{n+1}\mathbf{C}^{n+1} + \mathbf{A}^{n}\mathbf{C}^{n} + \mathbf{B}_I^n \left(\mathbf{I}^n - \mathbf{I}^{n+1}\right) + \mathbf{B}_T^n \left(\mathbf{T}^n - \mathbf{T}^{n+1}\right) + \mathbf{B}_N^n \left(\mathbf{N}^n - \mathbf{N}^{n+1}\right) + \mathbf{S}^n = 0, \tag{10}$$

where $\mathbf{C}^n$, $\mathbf{I}^n$, $\mathbf{T}^n$, and $\mathbf{N}^n$ denote spatiotemporal fields of tracer concentration and environmental drivers, and $\mathbf{S}^n$ denotes source-sink term. $\mathbf{A}^n$ and $\mathbf{B}_x^n$, where $x \in I, T, N$, parameterize physical and biological interactions across space and time and are learned implicitly by the model.

In practice, strong correlations in spatiotemporal fields can hinder representation learning. Masking alleviates this issue by reducing redundancy and acting as an implicit feature selection mechanism. For example, MAE typically mask up to 75% of input pixels, which is also used in this study. Accordingly, masked reconstruction can implicitly learn discrete transport operators from large-scale satellite observations, without requiring explicit physical constraints or supervision. The derivation of Equation (10) is provided in Appendix F.

4.4 Objective Function

We consider two reconstruction targets during pretraining: the original continuous values and discretized tiers. Because most biological parameters in our dataset follow log-normal distributions (Campbell, 1995; Koch, 1966), optimizing mean squared error (MSE) on continuous targets biases predictions toward high values. Therefore, we reconstruct discretized tiers following Lee et al. (2024):

$$\mathcal{L} = \frac{1}{\sum \mathbf{M}_i} \sum_{\substack{i=1 \\ \mathbf{M}_i=1}}^{p} \|g(x_{.,i}) - f((\mathbf{1} - \mathbf{M}) \circ \mathbf{z}; \theta)\|_2^2, \tag{11}$$

where g represents a mapping from raw input values to reconstructed targets (e.g., tier index of masked value), $f(.;\theta)$ is the model, and $\mathbf{z}$ denotes the embedded representations of input data. We formulate the learning objective (Equation 11) as regression to preserve the ordinal structure induced by discretization. Missing labels are excluded from the reconstruction loss (i.e., equation 11) while their corresponding inputs remain visible to the model, as reconstructing unobserved labels does not have a clear physical interpretation in our data.

5 Experiments

We evaluate several model variants described in Appendix C. Models are pretrained using large-scale satellite observations described in Appendix A and the loss function of Equation (11), and are fine-tuned on tasks described in Appendix B. We benchmark the model against linear regression (LR),

Table 1: The prediction accuracies (R^2) of $p\mathrm{CO}_2$, O_2 concentration, air-sea O_2 flux (O_2FLUX), NCP, and CR derived by different model variants. Model configuration is provided in Table 5. The uncertainties are quantified as one standard deviation (see Appendix D).

MODEL	$p\mathrm{CO}_2$	O_2	O_2FLUX	NCP	CR
BASE	0.662 ± 0.009	0.971 ± 0.000	0.379 ± 0.004	0.258 ± 0.036	0.260 ± 0.026
BIN32	0.637 ± 0.016	0.970 ± 0.000	0.335 ± 0.012	0.145 ± 0.036	0.244 ± 0.048
BIN256	0.655 ± 0.013	0.971 ± 0.001	0.371 ± 0.007	0.265 ± 0.024	0.297 ± 0.018
BIN512	0.668 ± 0.011	0.971 ± 0.001	0.366 ± 0.020	0.174 ± 0.050	0.290 ± 0.020
BIN2048	0.663 ± 0.013	0.971 ± 0.001	0.367 ± 0.014	0.189 ± 0.023	0.201 ± 0.029
L4	0.640 ± 0.013	0.969 ± 0.002	0.353 ± 0.010	0.282 ± 0.023	0.289 ± 0.010
L8	0.641 ± 0.024	0.970 ± 0.001	0.369 ± 0.011	0.305 ± 0.018	0.218 ± 0.027
L32	0.669 ± 0.005	0.972 ± 0.000	0.387 ± 0.007	0.297 ± 0.022	0.198 ± 0.020
D32	0.629 ± 0.018	0.970 ± 0.000	0.361 ± 0.006	0.293 ± 0.012	0.228 ± 0.020
D128	0.649 ± 0.019	0.971 ± 0.001	0.360 ± 0.011	0.258 ± 0.019	0.246 ± 0.009
S5	0.622 ± 0.008	0.970 ± 0.001	0.351 ± 0.008	0.332 ± 0.008	0.238 ± 0.030
S7	0.617 ± 0.009	0.969 ± 0.001	0.348 ± 0.012	0.289 ± 0.040	0.258 ± 0.017
T2	0.631 ± 0.017	0.970 ± 0.001	0.360 ± 0.023	0.128 ± 0.092	0.265 ± 0.026
T5	0.638 ± 0.023	0.970 ± 0.001	0.351 ± 0.010	0.212 ± 0.041	0.232 ± 0.029
PD_FEW	0.634 ± 0.017	0.962 ± 0.000	0.262 ± 0.007	0.310 ± 0.020	0.245 ± 0.026
PD_0.3	0.643 ± 0.015	0.972 ± 0.000	0.372 ± 0.006	0.273 ± 0.010	0.244 ± 0.040
PD_REG	0.641 ± 0.010	0.970 ± 0.001	0.327 ± 0.012	0.245 ± 0.009	0.126 ± 0.008

RF, XGB, multi-layer perceptron (MLP), and TabPFN, using the coefficient of determination (R^2) as the metric. Details of the training procedure, hyperparameters, and experimental setup are provided in Appendices D and E.

6 RESULTS

6.1 ABLATION STUDIES

Table 1 summarizes the prediction accuracies of the model variants across the $p\mathrm{CO}_2$, O_2, air-sea O_2 flux, NCP, and CR tasks. Overall, the models achieve higher accuracy for the inventory variables (O_2 and $p\mathrm{CO}_2$) than for the rate variables (air-sea O_2 flux, NCP, and CR). The models perform best on the O_2 task and worst on the GPP, NPP, N_2 fixation, and nitrification tasks (Table 8 in Appendix). These differences likely reflect variations in data availability (Table 4 in Appendix) and differences in the complexity of the input–output relationships.

Model size. Increasing the number of encoder layers from 4 to 32 (i.e., L4, L8, and L32) improves prediction accuracies for the $p\mathrm{CO}_2$, O_2, air-sea O_2 flux, and NCP tasks, while providing no statistically significant improvement, or even slight degradation, for other tasks. Expanding the hidden dimension (i.e., D32 and D128) leads to moderate improvements for the $p\mathrm{CO}_2$, O_2, and CR tasks, though these changes are not statistically significant. Overall, scaling model capacity does not uniformly improve performance, and optimal model size is task-dependent.

Discretization. Model performance generally improves as the number of discretized bins increases, then plateaus or slightly declines. The initial gains likely reflect a finer representation of input variability. The subsequent saturation and degradation may arise from four factors. First, ocean color observations carry substantial uncertainties in low-value ranges due to imperfect atmospheric correction and limited sensor signal-to-noise ratios (Hu et al., 2012). Second, quantile-based binning with many bins produces very narrow intervals in low-value ranges due to log-normal distributions (Campbell, 1995; Koch, 1966). Third, input-output relationships in our tasks are inherently noisy, limiting the benefit of improved representation in variables. Fourth, larger bin vocabularies require more pretraining data to adequately learn embedding representations.

Image size. The prediction accuracy varies across tasks as the input image size increases. For example, the R^2 for NCP increases to 0.332 when input image size is enlarged to 5×5, but it slightly decreases to 0.289 at 7×7. In contrast, enlarging the input image size degrades performance

for most tasks, particularly $p\mathrm{CO}_2$. We hypothesize that these differences may reflect mismatches in characteristic spatiotemporal scales between inputs and outputs across tasks. NCP varies on timescales of weeks, comparable to those of input variables such as Chl, and therefore benefits from additional local spatial context. By contrast, $p\mathrm{CO}_2$ varies on longer timescales (around 10 months), as a result incorporating larger spatial neighborhoods may introduce variability that acts as noise rather than signal.

Input time series length. Extending the input to three consecutive 8-day averages yields significant improvements for most tasks, but not for O_2 and N_2 fixation, suggesting importance of temporal variability in predicting most variables. Increasing the temporal window to five consecutive averages leads to plateaued or reduced performance, likely due to the inclusion of redundant information.

Scaling pretraining data volume. Prediction accuracies increase with the spatial coverage and sampling rate of the pretraining data. However, they exhibit no statistically significant improvements when increasing the sampling rate from 10% to 30%, except of O_2 and NCP, likely due to strong spatiotemporal correlations in satellite remote sensing observations.

6.2 Benchmark Comparison

Table 2 summarizes predictive performance (R^2) across $p\mathrm{CO}_2$, O_2, air-sea O_2 flux, NCP, and CR estimated by LR, RF, XGB, MLP, TabPFN, and OcMAE (the model variants achieving the highest performance are selected). All models achieve high prediction accuracy for the O_2 task. The superior performance arises from the robust relationship between temperature and O_2 solubility, which dominates surface ocean O_2 (Garcia & Gordon, 1992). This relationship is relatively linear and stable across space and time, as evidenced in the high prediction accuracy of LR model ($R^2 = 0.957$). Also, sea surface temperature (SST) can be accurately derived from satellite observations (Minnett et al., 2019). OcMAE and TabPFN models achieve slightly higher prediction accuracy than the other models, likely because they can better capture biological O_2, which becomes important during the growing season (Li & Cassar, 2017; 2018; Li et al., 2024).

For the $p\mathrm{CO}_2$ task, most models achieve $R^2 \in [0.5, 0.7]$. Despite $> 2,000,000$ training data (Table 4 in Appendix), OcMAE still outperform the baselines, due to complex input-output relationship that varies across space and time. This is partly supported by the poor performance of LR ($R^2 = 0.015$).

For the air-sea O_2 flux task, most models achieve $R^2 \approx 0.3$, with OcMAE consistently outperforming baselines with moderate gains. The low prediction accuracy for all models is partly attributable to uncertainties in the target data, which largely arise from estimating air-sea O_2 flux using gas exchange parameterizations (Wanninkhof et al., 2009).

For the NCP, CR, and NPP tasks, model performance diverges (Tables 2 and 8). LR achieves $R^2 = 0.170$ (NCP) and 0.151 (CR), while RF and XGB improve upon this by capturing nonlinear input-output relationships. MLP and TabPFN underperform relative to LR on NCP, suggesting overfitting under limited training data (Table 4 in Appendix). OcMAE achieves the best performance on NCP and CR; however it underperforms RF, XGB, and TabPFN on NPP. The poor performance may be attributed to a temporal resolution mismatch between the inputs and NPP measurements, which have a 24-hour temporal resolution. Pretraining on 8-day inputs may fail to capture the temporal variability relevant to NPP, resulting in less informative representations for NPP.

For the GPP, N_2 fixation, and nitrification tasks, all models perform poorly (Table 8 in Appendix). This is likely due to limited training data and input features that are insufficient to capture spatiotemporal variability in these processes (Tang et al., 2019). These results highlight the challenges of predicting ocean biogeochemical variables from satellite observations and motivate future work.

Overall, OcMAE improves performance on four of nine downstream tasks, indicating that large-scale self-supervised pretraining on satellite observations enhances generalization in ocean biogeochemical prediction. Tree-based models (i.e., RF and XGB) remain strong baselines. TabPFN performs well when input-output relationship is relatively simple (e.g., approximately linear), but it is prone to overfitting in complex tasks (e.g., NCP and CR).

Table 2: The prediction accuracy (R^2) of pCO_2, O_2 concentration, air-sea O_2 flux (O_2FLUX), NCP, and CR derived by linear regression (LR), random forests (RF), extreme gradient boosting (XGB), multi-layer perceptron (MLP), tabular prior-data fitted networks (TabPFN), and OcMAE (our) models. The uncertainties are quantified as one standard deviation (see Appendix D for details).

MODEL	pCO_2	O_2	O_2FLUX	NCP	CR
LR	0.015 ± 0.002	0.957 ± 0.000	0.249 ± 0.001	0.170 ± 0.009	0.151 ± 0.087
RF	0.604 ± 0.005	0.970 ± 0.001	0.372 ± 0.004	0.252 ± 0.019	0.194 ± 0.049
XGB	0.650 ± 0.003	0.970 ± 0.001	0.344 ± 0.007	0.225 ± 0.026	0.220 ± 0.080
MLP	0.565 ± 0.011	0.955 ± 0.003	0.234 ± 0.018	0.041 ± 0.109	-0.644 ± 0.738
TABPFN	0.647 ± 0.003	$\mathbf{0.973} \pm 0.000$	0.367 ± 0.006	0.090 ± 0.027	0.128 ± 0.089
OCMAE	$\mathbf{0.669} \pm 0.005$	0.972 ± 0.000	$\mathbf{0.387} \pm 0.007$	$\mathbf{0.297} \pm 0.022$	$\mathbf{0.297} \pm 0.018$

6.3 GLOBAL DISTRIBUTIONS

Figure 2 shows the spatial distribution of the first 8-day NCP in 2022 estimated by OcMAE. The spatial distributions for pCO_2, O_2, air–sea O_2 flux, and CR can be found in Appendix H. We do not report the distributions for the other tasks due to poor model performance. The spatial patterns of these estimates are broadly consistent with current understanding (e.g., pCO_2 by Fay et al. (2024)), while also exhibiting notable discrepancies relative to benchmark models (Appendix I). For example, models predict different NCP patterns in the eastern equatorial Pacific. These discrepancies highlight the need for continued improvements in predictive models to better constrain the spatiotemporal dynamics of these variables. We defer a detailed analysis of spatiotemporal patterns to future work.

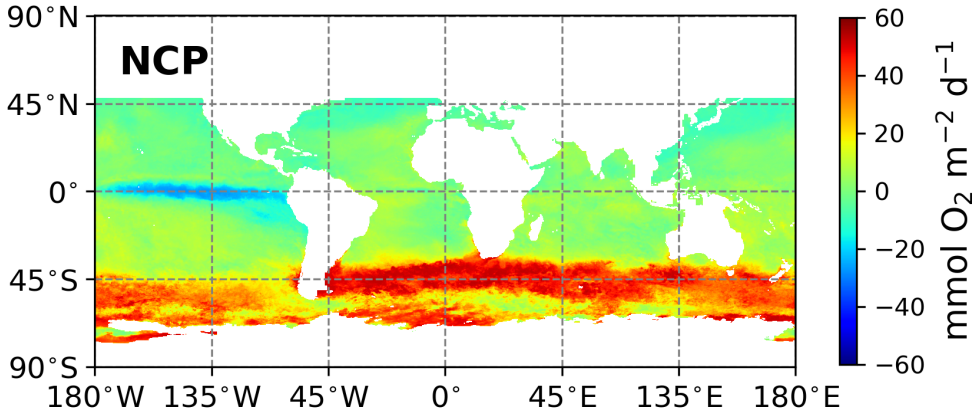

Figure 2: Global distribution of the first 8-day NCP in 2022 estimated by OcMAE.

7 DISCUSSION

Binning and embedding. By discretizing continuous variables into tiers via a quantile-based binning strategy, OcMAE approximates the input-output mapping using a piecewise-constant function. This design can effectively capture high nonlinearities, reduce the impact of noise, and naturally accommodate missing values. The discretization strategy is analogous to the quantile-based binning used in tree methods such as XGB to approximate candidate split points (Chen & Guestrin, 2016). This approximation enables scalable learning on large datasets by reducing the search space for splits, while accurately approximating complex input-output relationships through fine-grained partitioning of the input space (Chen & Guestrin, 2016).

However, classical impurity-based tree methods are known to favor features with many categories or split points, leading to selection bias and potential overfitting (Hothorn et al., 2006; Strobl et al., 2007). While tree-based models address this challenge through techniques such as regularization

(e.g., XGB) and bagging (e.g., RF) (Breiman, 2001; Chen & Guestrin, 2016), OcMAE provides a complementary approach by embedding discretized bins into a learnable representation space, which can be used for incorporating knowledge from large-scale satellite observations, allowing the model to capture higher-order nonlinearities and feature interactions.

Self-supervised learning using large-scale satellite observations. By factorizing a joint distribution as the product of conditional probability distributions, as commonly done in natural language modeling (Bengio et al., 2003; Radford et al., 2019), OcMAE can be interpreted as learning the representations that capture the spatiotemporal dynamics of oceanic tracers (e.g., Chl and SST) and the sensitivity of source-sink terms to the forcing variables. The dynamics is described by the tracer transport equation (see Appendix F), which encapsulates physical and biological (e.g., photosynthesis encoded in the source-sink term) processes. From this perspective, OcMAE also can be interpreted as learning an operator that approximates the underlying data-generating processes. The mathematical foundation of this approach is supported by the universal approximation (Lu et al., 2021).

The representations learned from large-scale satellite observations can be viewed as a prior distribution for ocean biogeochemical processes from Bayesian perspective (Grant et al., 2018; Gordon et al., 2019; Müller et al., 2022; Müller et al., 2024), with downstream applications interpreted as inferring posterior distributions conditioned on the prior and likelihood from labeled datasets. This knowledge transfer through representations is critical for the global ocean biogeochemistry, where satellites provide massive global observations with a high spatiotemporal resolution over the past more than three decades, while most labeled datasets remain sparse (see Figures 3-10).

Ocean biogeochemistry prediction. While prior knowledge improves the prediction across four of nine downstream tasks, it is most beneficial for tasks with a moderate number of labeled data and nonlinear input-output relationships. This observation is broadly consistent with Bayesian inference, where informative priors improve predictions when data are limited but still sufficient to constrain the posterior. The poor performance of OcMAE on NPP may reflect an uninformative or misaligned prior for this task. Meanwhile, RF and XGB remain strong baselines, demonstrating robust performance across downstream tasks, consistent with prior studies (Hollmann et al., 2025; Gorishniy et al., 2021; Shwartz-Ziv & Armon, 2022).

Satellite data are confined to the surface, missing subsurface information. For tasks with critical dependent variables not in the inputs, the system is analogous to a dynamical system with missing states, making predictions more dependent on initial states (Brajard et al., 2020; Cheng et al., 2023). This is consistent with poor performance on the nitrogen tasks (Tang et al., 2019; 2023). Some tasks (e.g., NPP) have sparse spatial coverage, warranting additional data collection and further investigation. Future work could improve model performance by incorporating additional input variables relevant to ocean biogeochemistry and key biogeochemical models (e.g., physics-informed neural networks (Raissi et al., 2019; Cranmer et al., 2020)). Finally, global distributions estimated by different algorithms exhibit substantial discrepancies, highlighting the need for improved understanding of ocean biogeochemical processes that are critical to the changing climate.

8 Conclusion

OcMAE improves prediction accuracy on four of nine ocean biogeochemistry regression tasks by learning representations of ocean biogeochemical processes from large-scale satellite observations through two key components. First, the quantile binning strategy allows the model to better capture nonlinearities and naturally address missing values, while embedding these discretized bins into a learnable representation space enables incorporation of large-scale satellite observations. Second, masked reconstruction can be interpreted as learning the tracer transport equation, including advection, diffusion, and the sensitivity of the source-sink term to the temporal variability in forcing factors. Finally, the OcMAE consistently outperforms strong baselines in four ocean biogeochemistry prediction tasks. While developed for ocean biogeochemistry, the proposed paradigm is broadly applicable to other spatiotemporal prediction problems. Future work could further improve performance by integrating explicit biogeochemical process models and additional data modalities that better characterize the underlying data-generating system.

AI USE STATEMENT

In this work, we used generative AI tools for editing the grammar. We have carefully reviewed the grammar and syntax provided by the tools. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Details on datasets and models can be found in the supplementary materials A, B, C, D, and E. Detailed derivation of advection-diffusion-reaction equation can be found in the supplementary materials F. Upon acceptance, we will release the Python scripts and in-situ measurements to facilitate reproducibility and future research. We will also make a time series of global 8-day data products since 2003 publicly available online to support the other studies.

ACKNOWLEDGMENTS

We acknowledge SOCAT for the pCO_2 data, BGC-Argo program for the O_2 observations, NASA for the satellite observations, and the community for sharing their datasets including O_2/Ar, GPP and CR estimated from light-dark bottle observations, and marine N_2 fixation rate.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization, 2016.

D. C. E. Bakker, B. Pfeil, C. S. Landa, N. Metzl, K. M. O'Brien, A. Olsen, K. Smith, C. Cosca, S. Harasawa, S. D. Jones, S. Nakaoka, Y. Nojiri, U. Schuster, T. Steinhoff, C. Sweeney, T. Takahashi, B. Tilbrook, C. Wada, R. Wanninkhof, S. R. Alin, C. F. Balestrini, L. Barbero, N. R. Bates, A. A. Bianchi, F. Bonou, J. Boutin, Y. Bozec, E. F. Burger, W.-J. Cai, R. D. Castle, L. Chen, M. Chierici, K. Currie, W. Evans, C. Featherstone, R. A. Feely, A. Fransson, C. Goyet, N. Greenwood, L. Gregor, S. Hankin, N. J. Hardman-Mountford, J. Harlay, J. Hauck, M. Hoppema, M. P. Humphreys, C. W. Hunt, B. Huss, J. S. P. Ibánhez, T. Johannessen, R. Keeling, V. Kitidis, A. Körtzinger, A. Kozyr, E. Krasakopoulou, A. Kuwata, P. Landschützer, S. K. Lauvset, N. Lefèvre, C. Lo Monaco, A. Manke, J. T. Mathis, L. Merlivat, F. J. Millero, P. M. S. Monteiro, D. R. Munro, A. Murata, T. Newberger, A. M. Omar, T. Ono, K. Paterson, D. Pearce, D. Pierrot, L. L. Robbins, S. Saito, J. Salisbury, R. Schlitzer, B. Schneider, R. Schweitzer, R. Sieger, I. Skjelvan, K. F. Sullivan, S. C. Sutherland, A. J. Sutton, K. Tadokoro, M. Telszewski, M. Tuma, S. M. A. C. van Heuven, D. Vandemark, B. Ward, A. J. Watson, and S. Xu. A multi-decade record of high-quality fco_2 data in version 3 of the surface ocean co_2 atlas (socat). *Earth System Science Data*, 8(2):383–413, 2016. doi:10.5194/essd-8-383-2016. URL https://essd.copernicus.org/articles/8/383/2016/.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers, 2021.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975. doi:10.2307/2987782.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023. doi:10.1038/s41586-023-06185-3.

Julien Brajard, Alberto Carrassi, Marc Bocquet, and Laurent Bertino. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the lorenz 96 model. *Journal of Computational Science*, 44:101171, 2020. doi:10.1016/j.jocs.2020.101171.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi:10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Janet W. Campbell. The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans*, 100(C7):13237–13254, 1995. doi:10.1029/95JC00458.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative Pretraining from Pixels. In *International Conference on Machine Learning*, volume 119, pp. 1691–1703. PMLR, 2020.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. ISBN 978-1-4503-4232-2. doi:10.1145/2939672.2939785.

Xiao Chen, Yibin Huang, Haoran Liu, Nicolas Cassar, Xin Liu, Weilei Wang, Fei Chai, Jianhua Kang, and Bangqin Huang. Reduced estimate of global marine primary productivity and hemispheric redistribution over the satellite era. *Communications Earth & Environment*, 2026. doi:10.1038/s43247-026-03828-x.

Sibo Cheng, César Quilodrán-Casas, Said Ouala, Alban Farchi, Che Liu, Pierre Tandeo, Ronan Fablet, Didier Lucor, Bertrand Iooss, Julien Brajard, Dunhui Xiao, Tijana Janjic, Weiping Ding, Yike Guo, Alberto Carrassi, Marc Bocquet, and Rossella Arcucci. Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA Journal of Automatica Sinica*, 10(6):1361–1387, 2023. doi:10.1109/JAS.2023.123537.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. doi:10.48550/arXiv.2207.08051.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi:10.1073/pnas.1912789117.

Geoffrey Dawson, Remy Vandaele, Andrew Taylor, David Moffat, Helen Tamura-Wicks, Sarah Jackson, Rosie Lickorish, Paolo Fraccaro, Hywel Williams, Chunbo Luo, and Anne Jones. A sentinel-3 foundation model for ocean colour, 2025. URL `https://arxiv.org/abs/2509.21273`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 4171–4186, Minneapolis, Minnesota, 2019. doi:10.18653/v1/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. doi:10.48550/arXiv.2010.11929.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why Does Unsupervised Pre-training Help Deep Learning? In *International Conference on Artificial Intelligence and Statistics*, volume 9 of *JMLR Workshop and Conference Proceedings*, pp. 201–208, 2010.

Amanda R. Fay, David R. Munro, Galen A. McKinley, Denis Pierrot, Stewart C. Sutherland, Colm Sweeney, and Rik Wanninkhof. Updated climatological mean ΔfCO_2 and net sea–air co_2 flux over the global open ocean regions. *Earth System Science Data*, 16(4):2123–2139, 2024. doi:10.5194/essd-16-2123-2024.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners, 2022.

Hernan E. Garcia and Louis I. Gordon. Oxygen solubility in seawater: Better fitting equations. *Limnology and Oceanography*, 37(6):1307–1312, 1992. doi:10.4319/lo.1992.37.6.1307.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked Autoencoder for Distribution Estimation. In *International Conference on Machine Learning*, volume 37, pp. 881–889. PMLR, 2015.

D. Giglio, V. Lyubchich, and M. R. Mazloff. Estimating Oxygen in the Southern Ocean Using Argo Temperature and Salinity. *Journal of Geophysical Research: Oceans*, 123:4280–4297, 2018. doi:10.1029/2017JC013404.

L. Gloege, M. Yan, T. Zheng, and G. A. McKinley. Improved Quantification of Ocean Carbon Uptake by Using Machine Learning to Merge Global Models and pCO2 Data. *Journal of Advances in Modeling Earth Systems*, 14, 2022. doi:10.1029/2021MS002620.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://arxiv.org/abs/1805.09921`.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, pp. 18932–18943, 2021.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas L. Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations (ICLR)*, 2018.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1735–1742. IEEE, 2006. ISBN 978-0-7695-2597-6. doi:10.1109/CVPR.2006.100.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi:10.48550/arXiv.1512.03385.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735. IEEE, 2020. ISBN 978-1-7281-7168-5. doi:10.1109/CVPR42600.2020.00975.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988. IEEE, 2022. ISBN 978-1-6654-6946-3. doi:10.1109/CVPR52688.2022.01553.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi:10.1038/s41586-024-08328-6.

James Holte, Lynne D. Talley, John Gilson, and Dean Roemmich. An Argo mixed layer climatology and database. *Geophysical Research Letters*, 44:5618–5626, 2017. doi:10.1002/2017GL073426.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5227–5244, 2024. doi:10.1109/TPAMI.2024.3362475.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

Chuanmin Hu, Lian Feng, Zhongping Lee, Curtiss O. Davis, Antonio Mannino, Charles R. McClain, and Bryan A. Franz. Dynamic range and sensitivity requirements of satellite ocean color sensors: Learning from the past. *Applied Optics*, 51:6045–6062, 2012. doi:10.1364/AO.51.006045.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. TabTransformer: Tabular Data Modeling Using Contextual Embeddings, 2020.

Yibin Huang, David Nicholson, Bangqin Huang, and Nicolas Cassar. Global estimates of marine gross primary production based on machine learning upscaling of field observations. *Global Biogeochemical Cycles*, 35(3):e2020GB006718, 2021. doi:10.1029/2020GB006718.

Aapo Hyvarinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Takamitsu Ito, Ahron Cervania, Kaylin Cross, Sanika Ainchwar, and Sara Delawalla. Mapping Dissolved Oxygen Concentrations by Combining Shipboard and Argo Observations Using Machine Learning Algorithms. *Journal of Geophysical Research: Machine Learning and Computation*, 1, 2024. doi:10.1029/2024JH000272.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General Perception with Iterative Attention, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. doi:10.1038/s41586-021-03819-2.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6114.

Arthur L. Koch. The logarithm in biology: I. mechanisms generating the log-normal distribution exactly. *Journal of Theoretical Biology*, 12(2):276–290, 1966. doi:10.1016/0022-5193(66)90119-6.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024. doi:10.1038/s41586-024-07744-y.

Peter Landschützer, Nicolas Gruber, F. Alexander Haumann, Christian Rödenbeck, Dorothee C. E. Bakker, Steven Van Heuven, Mario Hoppema, Nicolas Metzl, Colm Sweeney, Taro Takahashi, Bronte Tilbrook, and Rik Wanninkhof. The reinvigoration of the Southern Ocean carbon sink. *Science*, 349:1221–1224, 2015. doi:10.1126/science.aab2620.

Hugo Larochelle and Iain Murray. The Neural Autoregressive Distribution Estimator. In *International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 29–37, 2011.

Kyungeun Lee, Ye Seul Sim, Hye-Seung Cho, Moonjung Eo, Suhee Yoon, Sanghyu Yoon, and Woohyung Lim. Binning as a Pretext Task: Improving Self-Supervised Learning in Tabular Domains, 2024.

Zuchuan Li and Nicolas Cassar. Satellite estimates of net community production based on O_2 /Ar observations and comparison to other estimates. *Global Biogeochemical Cycles*, 30:735–752, 2016. doi:10.1002/2015GB005314.

Zuchuan Li and Nicolas Cassar. A mechanistic model of an upper bound on oceanic carbon export as a function of mixed layer depth and temperature. *Biogeosciences*, 14(22):5015–5027, 2017. doi:10.5194/bg-14-5015-2017.

Zuchuan Li and Nicolas Cassar. Theoretical considerations on factors confounding the interpretation of the oceanic carbon export ratio. *Global Biogeochemical Cycles*, 32(11):1644–1658, 2018. doi:10.1029/2018GB006003.

Zuchuan Li, Yajuan Lin, and Nicolas Cassar. The influence of phytoplankton size fractions on the carbon export ratio in the surface ocean. *Ecological Modelling*, 495:110798, 2024. doi:10.1016/j.ecolmodel.2024.110798.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2019.

Bin Lu, Ze Zhao, Luyu Han, Xiaoying Gan, Yuntao Zhou, Lei Zhou, Luoyi Fu, Xinbing Wang, Chenghu Zhou, and Jing Zhang. OXYGENERATOR: Reconstructing Global Ocean Deoxygenation over a Century with Deep Learning. In *International Conference on Machine Learning*, pp. 33219–33242, 2024.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi:10.1038/s42256-021-00302-5.

P. J. Minnett, A. Alvera-Azcárate, T. M. Chin, G. K. Corlett, C. L. Gentemann, I. Karagali, X. Li, A. Marsouin, S. Marullo, E. Maturi, R. Santoleri, S. S. Picart, M. Steele, and J. Vazquez-Cuervo. Half a century of satellite remote sensing of sea-surface temperature. *Remote Sensing of Environment*, 233:111366, 2019. doi:10.1016/j.rse.2019.111366.

Samuel Müller, Noah Hollmann, and Frank Hutter. Bayes' power for explaining in-context learning generalizations. *arXiv preprint arXiv:2410.01565*, 2024. URL https://arxiv.org/abs/2410.01565.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers Can Do Bayesian Inference, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *Journal of Computational Physics*, pp. 686–707, 2019.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning*, volume 139, pp. 8821–8831. PMLR, 2021.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful Precipitation Nowcasting Using Deep Generative Models of Radar. *Nature*, 597:672–677, 2021.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566:195–204, 2019. doi:10.1038/s41586-019-0912-1.

Carol Robinson, Isabel Seguro, Giorgio Dall'Olmo, Gwenaelle Moncoiffe, Maria Aranguren-Gassis, Javier Arístegui, Maurizio Azzaro, Yong-Jae Baek, Federico Baltar, Melanie R. Cohn, et al. A global dataset of marine pelagic microbial respiration, 2026. URL https://www.bodc.ac.uk/data/published_data_library/catalogue/10.5285/4b2a5ac6-b6db-c98e-e063-7086abc040c6.

Z. Shao, Y. Xu, H. Wang, W. Luo, L. Wang, Y. Huang, N. S. R. Agawin, A. Ahmed, M. Benavides, M. Bentzon-Tilia, I. Berman-Frank, H. Berthelot, I. C. Biegala, M. B. Bif, A. Bode, S. Bonnet, D. A. Bronk, M. V. Brown, L. Campbell, D. G. Capone, E. J. Carpenter, N. Cassar, B. X. Chang, D. Chappell, Y.-L. Chen, M. J. Church, F. M. Cornejo-Castillo, A. M. S. Detoni, S. C. Doney, C. Dupouy, M. Estrada, C. Fernandez, B. Fernández-Castro, D. Fonseca-Batista, R. A. Foster, K. Furuya, N. Garcia, K. Goto, J. Gago, M. R. Gradoville, M. R. Hamersley, B. A. Henke, C. Hörstmann, A. Jayakumar, Z. Jiang, S.-J. Kao, D. M. Karl, L. R. Kittu, A. N. Knapp, S. Kumar, J. LaRoche, H. Liu, J. Liu, C. Lory, C. R. Löscher, E. Marañón, L. F. Messer, M. M. Mills, W. Mohr, P. H. Moisander, C. Mahaffey, R. Moore, B. Mouriño Carballido, M. R. Mulholland, S. Nakaoka, J. A. Needoba, E. J. Raes, E. Rahav, T. Ramírez-Cárdenas, C. F. Reeder, L. Riemann, V. Riou, J. C. Robidart, V. V. S. S. Sarma, T. Sato, H. Saxena, C. Selden, J. R. Seymour, D. Shi, T. Shiozaki, A. Singh, R. E. Sipler, J. Sun, K. Suzuki, K. Takahashi, Y. Tan, W. Tang, J.-É. Tremblay, K. Turk-Kubo, Z. Wen, A. E. White, S. T. Wilson, T. Yoshida, J. P. Zehr, R. Zhang, Y. Zhang, and Y.-W. Luo. Global oceanic diazotroph database version 2 and elevated estimate of global oceanic n_2 fixation. *Earth System Science Data*, 15(8):3673–3709, 2023. doi:10.5194/essd-15-3673-2023. URL `https://essd.copernicus.org/articles/15/3673/2023/`.

Jonathan D. Sharp, Andrea J. Fassbender, Brendan R. Carter, Gregory C. Johnson, Cristina Schultz, and John P. Dunne. GOBAI-O_2 : Temporally and spatially resolved fields of ocean interior dissolved oxygen over nearly 2 decades. *Earth System Science Data*, 15:4481–4518, 2023. doi:10.5194/essd-15-4481-2023.

Ravid Shwartz-Ziv and Amitai Armon. Tabular Data: Deep Learning is not All You Need. *Information Fusion*, pp. 84–90, 2022.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Carlos Gomes, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Disha Shidham, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, David Bell, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications, 2025.

Weiyi Tang, Zuchuan Li, and Nicolas Cassar. Machine Learning Estimates of Global Marine Nitrogen Fixation. *Journal of Geophysical Research: Biogeosciences*, 124:717–730, 2019. doi:10.1029/2018JG004828.

Weiyi Tang, Bess B. Ward, Michael Beman, Laura Bristow, Darren Clark, Sarah Fawcett, Claudia Frey, François Fripiat, Gerhard J. Herndl, Mhlangabezi Mdutyana, Fabien Paulot, Xuefeng Peng, Alyson E. Santoro, Takuhei Shiozaki, Eva Sintes, Charles Stock, Xin Sun, Xianhui S. Wan, Min N. Xu, and Yao Zhang. Database of nitrification and nitrifiers in the global ocean. *Earth System Science Data*, 15(11):5039–5077, 2023. doi:10.5194/essd-15-5039-2023.

Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. ViTs for SITS: Vision Transformers for satellite image time series. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10418–10428. IEEE, 2023. ISBN 979-8-3503-0129-8. doi:10.1109/CVPR52729.2023.01004.

Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional Image Generation with PixelCNN Decoders. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23:1661–1674, 2011. doi:10.1162/NECO_a_00142.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pp. 1096–1103. ACM Press, 2008. ISBN 978-1-60558-205-4. doi:10.1145/1390156.1390294.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

Rik Wanninkhof, William E. Asher, David T. Ho, Colm Sweeney, and Wade R. McGillis. Advances in quantifying air-sea gas exchange and environmental forcing. *Annual Review of Marine Science*, 1(1):213–244, 2009. doi:10.1146/annurev.marine.010908.163742.

Annie P. S. Wong, Susan E. Wijffels, Stephen C. Riser, Sylvie Pouliquen, Shigeki Hosoda, Dean Roemmich, John Gilson, Gregory C. Johnson, Kim Martini, David J. Murphy, Megan Scanderbeg, T. V. S. Udaya Bhaskar, Justin J. H. Buck, Frederic Merceur, Thierry Carval, Guillaume Maze, Cécile Cabanes, Xavier André, Noé Poffa, Igor Yashayaev, Paul M. Barker, Stéphanie Guinehut, Mathieu Belbéoch, Mark Ignaszewski, Molly O'Neil Baringer, Claudia Schmid, John M. Lyman, Kristene E. McTaggart, Sarah G. Purkey, Nathalie Zilberman, Matthew B. Alkire, Dana Swift, W. Brechner Owens, Steven R. Jayne, Cora Hersh, Pelle Robbins, Deb West-Mack, Frank Bahr, Sachiko Yoshida, Philip J. H. Sutton, Romain Cancouët, Christine Coatanoan, Delphine Dobbler, Andrea Garcia Juan, Jerôme Gourrion, Nicolas Kolodziejczyk, Vincent Bernard, Bernard Bourlès, Hervé Claustre, Fabrizio D'Ortenzio, Serge Le Reste, Pierre-Yve Le Traon, Jean-Philippe Rannou, Carole Saout-Grit, Sabrina Speich, Virginie Thierry, Nathalie Verbrugge, Ingrid M. Angel-Benavides, Birgit Klein, Giulio Notarstefano, Pierre-Marie Poulain, Pedro Vélez-Belchí, Toshio Suga, Kentaro Ando, Naoto Iwasaska, Taiyo Kobayashi, Shuhei Masuda, Eitarou Oka, Kanako Sato, Tomoaki Nakamura, Katsunari Sato, Yasushi Takatsuki, Takashi Yoshida, Rebecca Cowley, Jenny L. Lovell, Peter R. Oke, Esmee M. van Wijk, Fiona Carse, Matthew Donnelly, W. John Gould, Katie Gowers, Brian A. King, Stephen G. Loch, Mary Mowat, Jon Turton, E. Pattabhi Rama Rao, M. Ravichandran, Howard J. Freeland, Isabelle Gaboury, Denis Gilbert, Blair J. W. Greenan, Mathieu Ouellet, Tetjana Ross, Anh Tran, Mingmei Dong, Zenghong Liu, Jianping Xu, KiRyong Kang, HyeongJun Jo, Sung-Dae Kim, and Hyuk-Min Park. Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Frontiers in Marine Science*, Volume 7 - 2020, 2020. ISSN 2296-7745. doi:10.3389/fmars.2020.00700. URL `https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2020.00700`.

Qianxun Xu and Zuchuan Li. Partial Physics Informed Diffusion Model for Ocean Chlorophyll Concentration Reconstruction. In *Advances in Neural Information Processing Systems*, 2025.

A SATELLITE REMOTE SENSING DATA

We constructed a global oceanic satellite dataset for model pretraining. Table 3 shows satellite products included in the dataset: oceanic chlorophyll concentration (Chl), sea surface temperature (SST), photosynthetically active radiation (PAR), and wind vectors (including wind speed and horizontal and vertical components) spanning from 2003 to 2023. Daily L3 mapped Chl, SST, and PAR products from MODIS-Aqua were obtained from the NASA Ocean Color website (`https://oceancolor.gsfc.nasa.gov/`). Wind vectors were downloaded from the CCMP dataset (`https://www.remss.com/measurements/ccmp/`). NCEP sea level pressure reanalysis data, with a spatial resolution of $2.5^\circ \times 2.5^\circ$ and a temporal resolution of 6-hour, was used for calculating air-sea O_2 flux and downloaded from (`https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html`). Chl, SST, PAR, and Wind vectors were resampled to an 8-day temporal resolution and a spatial resolution of $0.083^\circ \times 0.083^\circ$. Mixed layer depth (MLD) cannot be directly measured by satellites. Although sea surface salinity (SSS) can now be estimated from satellite measurements, these estimates remain uncertain at high latitudes. Therefore, we used monthly SSS and MLD estimates derived from a previous study (Holte et al., 2017). For each image cube, monthly SSS and MLD values were matched to 8-day averages based on the closest location and date.

Due to computational constraints, we did not directly utilize the complete satellite observations in the world's oceans for model pretraining. Instead, we randomly sampled 10% of the data points from each 8-day average to construct a representative set of image cubes. This sampling rate leads to over 500,000 image cubes per 8-day period, which is comparable to the number obtained by uniformly sampling the data at a spatial interval of every 3 pixels. This sampling rate balances computational feasibility with data representativeness. Each image cube comprises eight sub-cubes corresponding to Chl, SST, PAR, Wind speed, Wind U, Wind V, SSS, and MLD. The Chl, SST, PAR, and wind vectors sub-cubes are represented as $3 \times 3 \times 3$ arrays (rows $\times$ columns $\times$ channels), where channels correspond to the number of consecutive 8-day average. The SSS and MLD sub-cubes, available at monthly temporal resolution, are represented as $3 \times 3 \times 1$ arrays. Each image cube is further associated with its geographical coordinates (i.e., latitude and longitude) to capture geographic variability and acquisition time (i.e., day of year) to represent seasonality. This sampling procedure results in over 500 million $3 \times 3 \times 20$ image cubes covering the world's oceans from 2003 to 2023.

Each feature is discretized into 1024 quantile-based bins using cut points derived from satellite observations in five representative years (2003, 2008, 2013, 2018, and 2023), selected based on the balance between temporal coverage and computational constraints. Each sub-cube (i.e., satellite product) is discretized and shares the same cutting points. The day-of-year variable contains 46 unique categories on an 8-day time scale. Missing values in individual features are treated as separate bins. In total, the pretraining dataset contains 8246 unique discretized bins (i.e., vocabulary) which will be embedded into D-dimensional feature vectors.

Notably, no data augmentation is applied to our input variables, as satellite observations have inherent physical meaning, and geometric transformations such as rotation or flipping could introduce artifacts. As we have massive large-scale satellite remote sensing observations, we can simulate data augmentation by increasing the sampling rate for each 8-day data (see session 6.1).

B DOWNSTREAM TASK

We constructed nine downstream prediction tasks spanning the carbon, oxygen, organic matter (i.e., metabolism), and nitrogen cycles. For the carbonate system, we used surface ocean pCO_2. For the oxygen cycle, we used surface ocean O_2 concentration and air-sea O_2 flux. For the ocean metabolism, we used NCP integrated over the mixed layer, as well as GPP, NPP, and CR integrated over the euphotic zone. For the nitrogen cycle, we used marine N_2 fixation rates and nitrification rates. The spatial distributions of sample locations are shown in Figures 3-11. The number of records for each task is reported in Table 4.

Surface ocean pCO_2 observations. The SOCATv2025 of pCO_2 was downloaded from a global dataset shared by the community online `https://socat.info/`. We used corrected estimates of the in-situ surface ocean CO_2 fugacity (fCO_2) that has very small difference from the pCO_2. We removed observations with values lower than 50 μatm or higher than 650 μatm. We gridded the

Table 3: Data products for model pretraining. *Ocean Color* dataset can be downloaded from `https://oceancolor.gsfc.nasa.gov/`. *CCMP* dataset can be downloaded from `https://www.remss.com/measurements/ccmp/`. *Chl* denotes chlorophyll concentration at the surface ocean. *SST* denotes sea surface temperature. *PAR* denotes photosynthetically active radiation. *MLD* denotes mixed layer depth that is defined based on the density threshold ($0.03\ kg\ m^{-3}$), and derived from temperature and salinity profiles measured by BGC-Argo floats. *SSS* denotes sea surface salinity. *Monthly climatology* is calculated as the average of all measurements in the corresponding month.

PRODUCT	SPATIAL RES.	TEMPORAL RES.	TIME LENGTH	SOURCE
CHL	$0.083° \times 0.083°$	1-DAY	3	OCEAN COLOR
SST	$0.083° \times 0.083°$	1-DAY	3	OCEAN COLOR
PAR	$0.083° \times 0.083°$	1-DAY	3	OCEAN COLOR
WIND SPEED	$0.25° \times 0.25°$	6-HOUR	3	CCMP
WIND U	$0.25° \times 0.25°$	6-HOUR	3	CCMP
WIND V	$0.25° \times 0.25°$	6-HOUR	3	CCMP
MLD	$1° \times 1°$	MONTHLY CLIMATOLOGY	1	HOLTE ET AL. (2017)
SSS	$1° \times 1°$	MONTHLY CLIMATOLOGY	1	HOLTE ET AL. (2017)

measurements to $0.083° \times 0.083°$ in space and 8-day average in time. The locations of pre-processed data points can be found in Figure 3. This resolution in space and time matches satellite remote sensing data. Details on the SOCAT dataset can be found in Bakker et al. (2016).

Surface ocean O_2 concentration. The observations of dissolved O_2 at the surface ocean were extracted from the global BGC-Argo dataset `ftp://usgodae.org/pub/outgoing/argo`. We only retained high-quality measurements with quality-control label of 1. As water in the mixed layer actively interacts with the atmosphere, we calculated surface ocean O_2 concentration as the average of observations within the MLD, where the density difference from the 10m depth is greater or equal to 0.03 $kg\,m^{-3}$. The vertical profiles of density were calculated from temperature and salinity profiles and sea-level pressure. Similarly, dissolved O_2 measurements were gridded to $0.083° \times 0.083°$ in space and 8-day average in time to match satellite remote sensing observations. The locations of O_2 concentrations are shown in Figure 4. Details on the BGC-Argo data can be found in Wong et al. (2020). While O_2 at the ocean surface is dominated by solubility and is influenced by temperature and salinity, O_2 at the ocean surface is usually well predicted by temperature. So, this task is easy among the downstream tasks.

Air-sea O_2 flux. Measurements of air-sea O_2 flux were calculated from the BGC-Argo dataset. We derived dissolved O_2 estimates by averaging measurements over the mixed layers from the BGC-Argo dataset. Then, we estimated O_2 supersaturation by removing O_2 solubility, which was estimated from temperature and salinity profiles and sea-level pressure, from dissolved O_2 observations. Finally, we estimated air-sea O_2 flux using O_2 supersaturation and air-sea gas exchange velocity parameterized by wind speed at 10 m above the sea surface. Air-sea O_2 flux estimates were further gridded to $0.083° \times 0.083°$ in space and 8-day average in time. The locations of air-sea O_2 fluxes are shown in Figure 5.

Oceanic GPP. Oceanic GPP profiles were extracted from a prior study (Huang et al., 2021). Using these profiles, we estimated oceanic GPP integrated over the euphotic depth, where the PAR is reduced to 1% of the surface value and is calculated from surface Chl. GPP observations were further gridded to $0.083° \times 0.083°$ in space and 8-day average in time. GPP observations have been transformed to a logarithmic space as the measurements approximately follow log-normal distribution. The locations of GPP are shown in Figure 6.

Oceanic CR. Oceanic CR profiles were extracted from Huang et al. (2021) and Robinson et al. (2026) (`https://www.bodc.ac.uk/data/published_data_library/catalogue/10.5285/4b2a5ac6-b6db-c98e-e063-7086abc040c6`). To ensure data consistency, We only retained profiles measured through the light-dark bottle approach in Robinson et al. (2026). Using these profiles, we estimated oceanic CR integrated over the euphotic depth as GPP. The preprocessed CR observations were further gridded to $0.083° \times 0.083°$ cells in space and 8-day

average in time. CR observations have been transformed to a logarithmic space as the measurements approximately follow log-normal distribution. The locations of CR are shown in Figure 7.

Oceanic NCP. Air-sea biological oxygen flux (i.e., NCP) is the balance between GPP and CR (i.e, NCP=GPP-CR). NCP measurements were calculated using dissolved O_2 and Argon (Ar) ratio (i.e., O_2/Ar) measurements. The key idea of O_2/Ar method is to extract O_2 due to biology by leveraging similar solubility properties between O_2 and Ar and removing physical component through their element ratio. The measurements were obtained from Li & Cassar (2016) and were recently updated for another independent research. There are over 2,000,000 NCP estimates calculated from raw O_2/Ar measurements which are highly correlated in space and time. They are gridded to $0.083^\circ \times 0.083^\circ$ in space and 8-day average in time to reduce correlations. More details on the NCP data can be found in Li & Cassar (2016). The locations of NCP are shown in Figure 8. NCP is influenced by factors including temperature, nutrient, light, and community. It is typically a nonlinear function of input variables, as documented in Li & Cassar (2016). Therefore, it is one of the difficult downstream tasks.

Oceanic NPP. Oceanic NPP observations integrated over the euphotic zone were obtained from an independent study (Chen et al., 2026). Based on the typical range of NPP in the world's oceans, we retained values within $[50, 2000]$ mg C m^{-2} day^{-1}, thereby removing some extreme values primarily from coastal regions (36 data points). The observations were gridded to a spatial resolution of $0.083^\circ \times 0.083^\circ$ and an 8-day average temporal resolution. The observations were further transformed to a logarithmic space as their distribution is approximately log-normal distribution. The locations of NPP observations are shown in Figure 9.

Marine N_2 fixation rate. We used the dataset of marine N_2 fixation rate compiled by a recent study (Shao et al., 2023). The dataset was downloaded online (`https://doi.org/10.6084/m9.figshare.21677687`). Among different measurements in this dataset, we used the 1,806 marine N_2 fixation measurements typically integrated over the euphotic zone. These measurements were gridded to $0.083^\circ \times 0.083^\circ$ in space and 8-day average in time. The locations of N_2 fixation rates are shown in Figure 10. More details on the marine N_2 fixation data set can be found in Shao et al. (2023). N_2 fixation rates have been transformed to a logarithmic space as the measurements approximately follow log-normal distribution. Marine N_2 fixation is poorly constrained and under sampled across the world's oceans. Therefore, it is a very difficult task.

Marine nitrification. We used the dataset of marine nitrification rate compiled by a recent study (Tang et al., 2023). The dataset was downloaded online (`https://doi.org/10.5281/zenodo.8355912`). We used the 2,417 observations of volumetric ammonia oxidation. We retained observations with complete date, valid observations, and the maximum for a given location and date. These measurements were further gridded to $0.083^\circ \times 0.083^\circ$ in space and 8-day average in time. The locations of nitrification are shown in Figure 11. More details on the marine nitrification data set can be found in Tang et al. (2023). Nitrification rates have been transformed to a logarithmic space as the measurements approximately follow log-normal distribution. Marine nitrification is poorly constrained and under sampled across the world's oceans. Therefore, it is also a very difficult task.

Each data point is aligned to input variables based on the center of image cube (i.e., data for pretraining) in space and time. We cleaned the matched data by removing records with missing data more than 20% in input variables. We normalized output variables using z-score. Then, we binned the input variables using the cutting points used in the pretraining and built the training datasets. We did not build our training, validation, and test datasets by randomly splitting the corresponding data. Instead, We split the corresponding data based on years, minimizing the correlation among these datasets, which are critical in geographical data (i.e., the first law of geography). Years were selected to provide around 10% of the data points for validation, 10% of the data points for testing, and the remaining for training. The training, validation, and test datasets for each compared task are shown in Table 4 and Figure 3-11.

C Model Variants

The encoder of the base model is a 16-layer Transformer with a hidden size of 64 and 4 attention heads. Each discretized bin is embedded as a 64-dimensional vector. The decoder is a 2-layer Transformer with a hidden size of 32 and 2 attention heads. Each feed-forward network contains 256 hidden units. Each model variant uses these values unless otherwise specified in Table 5.

Table 4: Training, validation, and test datasets for pCO_2, surface-ocean O_2 concentration, air-sea O_2 flux, NCP, NPP, GPP, CR, N_2 fixation rate, and nitrification, respectively. z-score is to remove the mean, normalized by the standard deviation.

DATASET	TRAINING	VALIDATION	TEST	TOTAL	TRANSFORM	Z-SCORE
pCO_2	2,607,432	116,225	99,867	2,823,524	NO	YES
O_2	124,851	24,025	22,997	171,873	NO	YES
AIR-SEA O_2 FLUX	124,851	24,025	22,997	171,873	NO	YES
NCP	24,774	6,580	4,600	35,954	NO	YES
NPP	1,076	118	139	1,333	LOG	YES
CR	520	73	116	709	LOG	YES
GPP	281	36	45	362	LOG	YES
N_2 FIXATION	836	169	114	1,119	LOG	YES
NITRIFICATION	219	52	33	304	LOG	YES

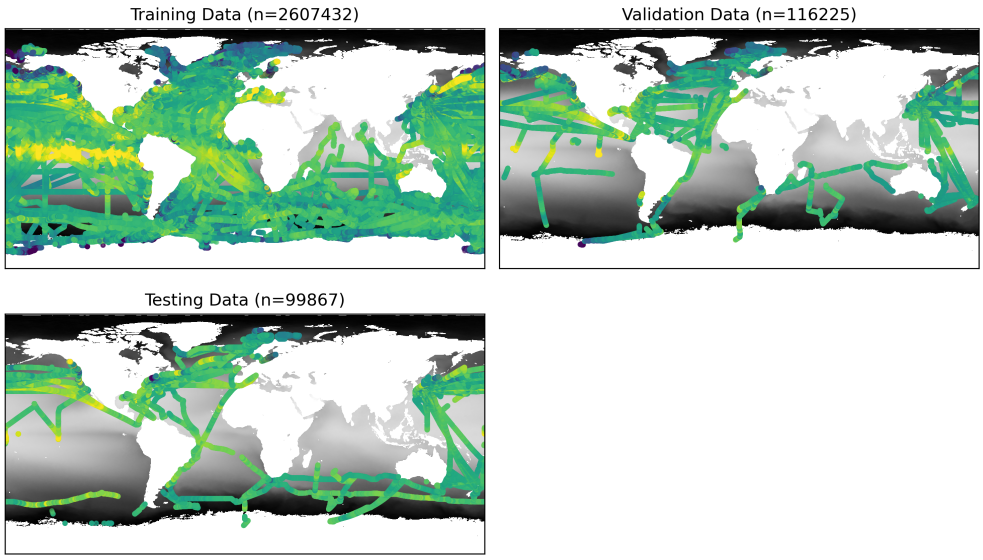

Figure 3: Locations of the pCO_2 measurements used for training (2002-2021), validation (2022), and testing datasets (2023). Data points are color-coded according to the corresponding pCO_2 at the surface ocean. The gray background is monthly sea surface temperature (SST) in August.

D EXPERIMENT

Model optimization uses Adam optimizer with parameters $\beta_1 = 0.95$ and $\beta_2 = 0.99$ Loshchilov & Hutter (2019). The learning rate follows a cosine decay schedule with a base rate of 1×10^{-4} and a 10-epoch warm-up Loshchilov & Hutter (2017). Weight decay is set at 0.05, and the batch size is set at 256, determined by GPU memory constraints. The model is trained for 100 iterations, ensuring that each pretraining data point is statistically seen at least once. In each iteration, data from different years are randomly distributed across GPUs, each of which processes the 46 8-day averages. In total, the model is exposed to over 1.8 billion image cubes during training.

We did a simple experiment to select the optimal base rate (1×10^{-4}) from four candidates $[2.5 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}]$, using the prediction accuracy of air-sea O_2 flux by the corresponding fine-tuning model. The optimal base rate is applied to all the other models, unless specified otherwise.

To alleviate the computational overhead associated with repeatedly loading large satellite imagery files, in each data loading step (i.e., iteration), 100 batches of 256 records are randomly drawn,

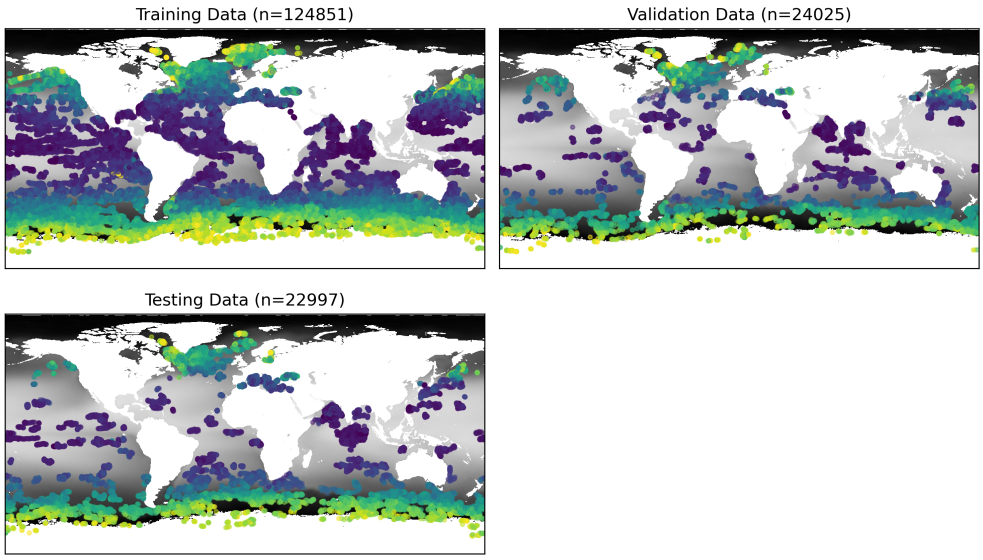

Figure 4: Locations of the O_2 measurements used for training (others), validation (2016-2017), and testing (2018-2019) datasets. Data points are color-coded according to the corresponding surface-ocean O_2 concentrations. The gray background represents monthly sea surface temperature (SST) in August.

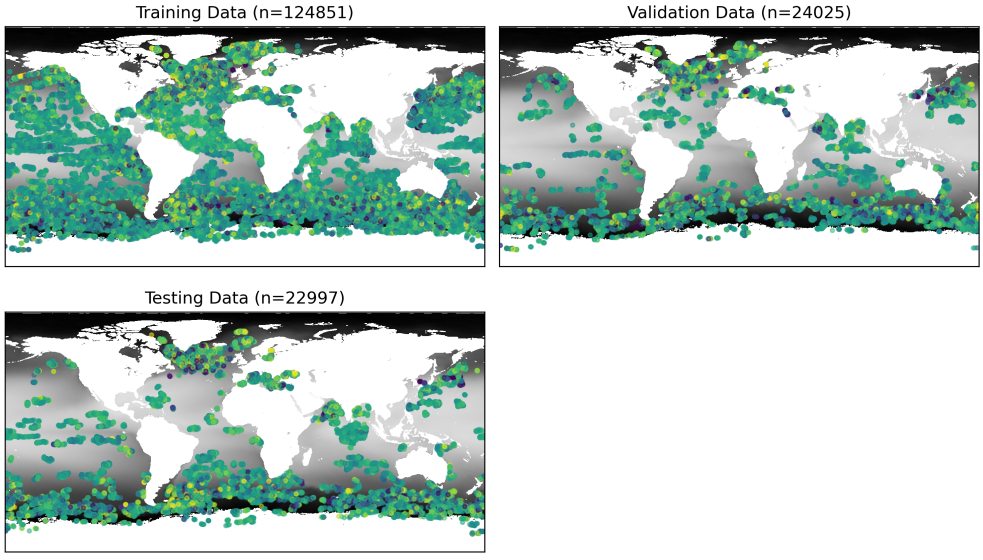

Figure 5: Locations of the air-sea O_2 flux used for training (others), validation (2016-2017), and testing (2018-2019) datasets. Data points are color-coded according to the corresponding air-sea O_2 fluxes. The gray background is monthly sea surface temperature (SST) in August.

ensuring diversity across both spatial and temporal domains. This approximate sampling strategy improves statistical representativeness, stabilizes gradient updates, and substantially reduces I/O latency. Model pretraining was conducted on 16 NVIDIA RTX 4090 GPUs for approximately 2-7 days dependent on model variants.

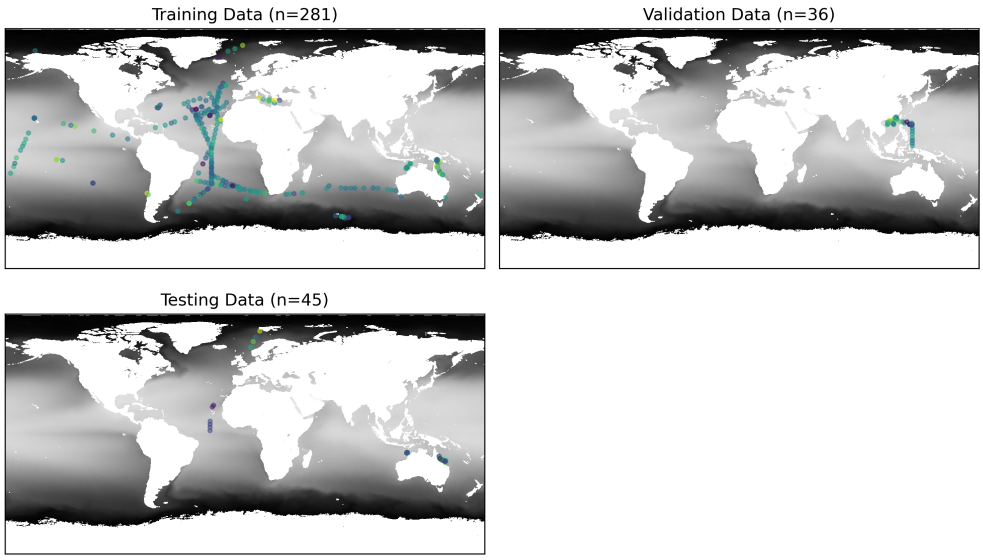

Figure 6: Locations of the gross primary production (GPP) measurements used for training (others), validation (2012-2019), and testing (2009-2010) datasets. Data points are color-coded with the corresponding GPP values. The gray background represents monthly sea surface temperature (SST) in August.

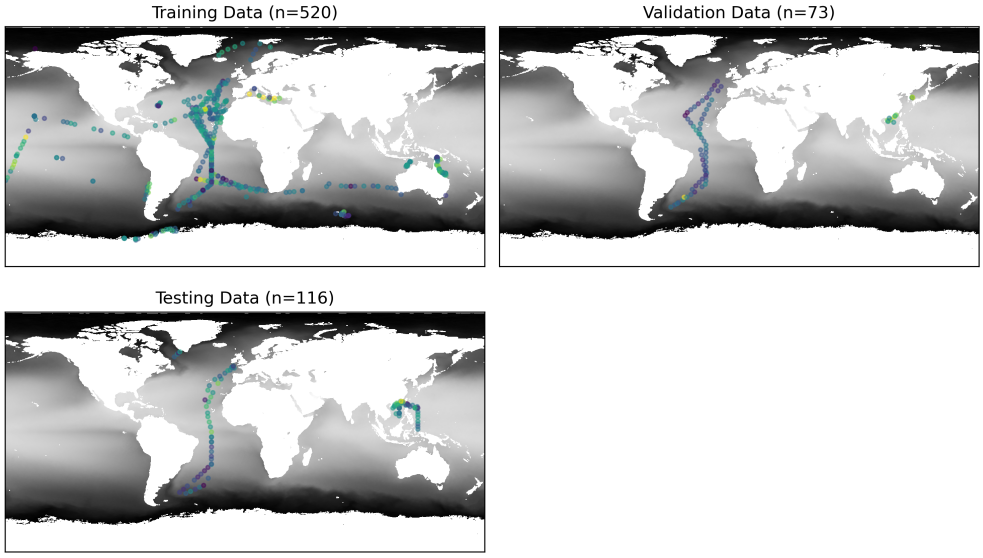

Figure 7: Locations of the community respiration (CR) measurements used for training (others), validation (2012-2013), and testing (2014-2019) datasets. Data points are color-coded according to the corresponding CR values. The gray background is monthly sea surface temperature (SST) in August.

We fine-tune separate models for each task by adding a linear output layer on top of the pretrained encoder. The input to this layer is the concatenation of the [CLS] token (see He et al. (2022); Devlin et al. (2019)) and the center pixels of the SST and Chl sub-cubes. The centers of SST and Chl are selected because of their spatiotemporal linkage to the downstream targets. The embedding vector

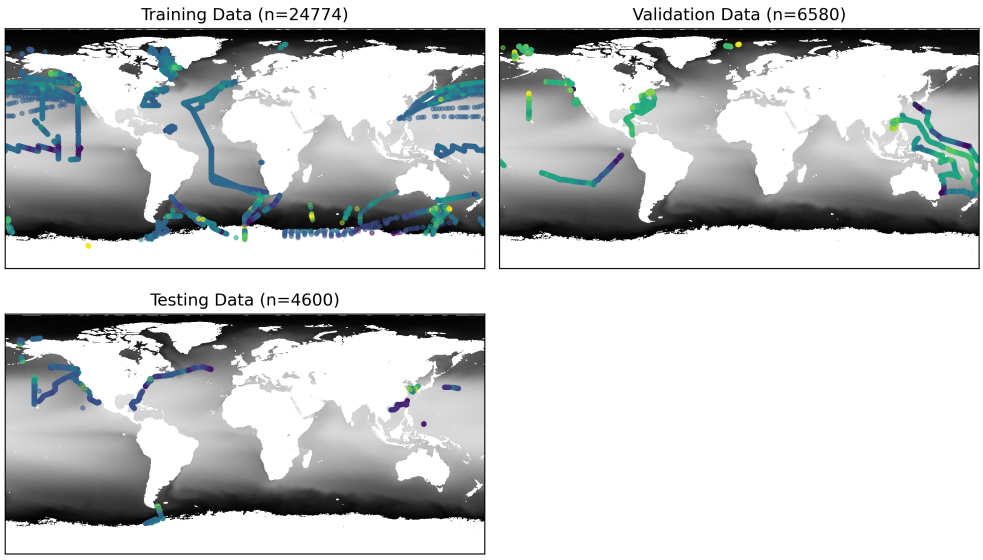

Figure 8: Locations of the net community production (NCP) measurements used for training (others), validation (2016-2017), and testing (2018-2019) datasets. Data points are color-coded according to the corresponding NCP values. The gray background is monthly sea surface temperature (SST) in August.

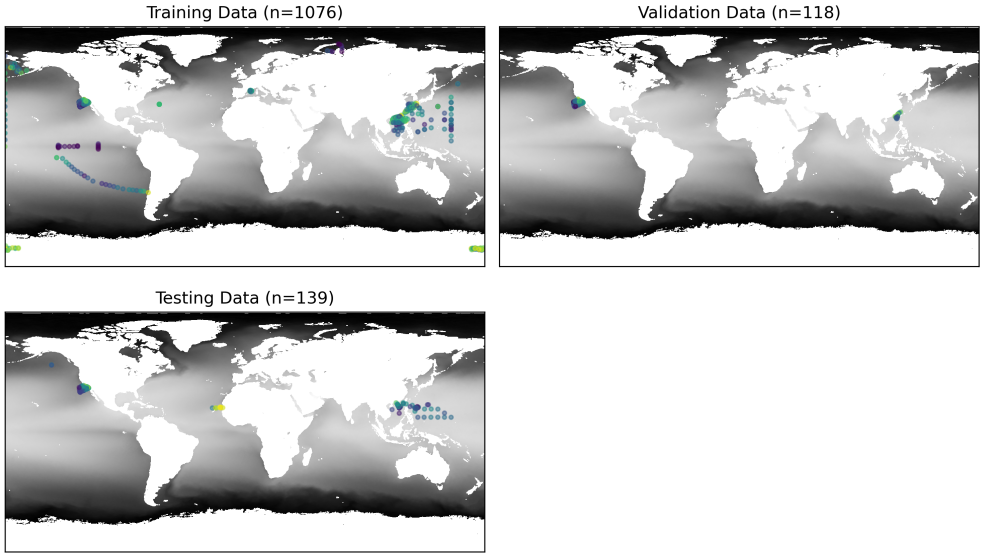

Figure 9: Locations of the net primary production (NPP) measurements used for training (others), validation (2016-2017), and testing (2018-2019) datasets. Data points are color-coded according to the corresponding NPP values. The gray background represents monthly sea surface temperature (SST) in August.

for each discretized bin is fixed (not updated) during the fine-tuning step. Fine-tuning uses a cosine learning rate decay, with a base learning rate of 1×10^{-4} and a 10-epoch warm-up. The batch size is set at 256, and epoch is set at 100 for all the downstream tasks. The weight decay is set to 0.01 for all downstream tasks. Early stopping with a patience of 10 epochs is applied during fine-tuning.

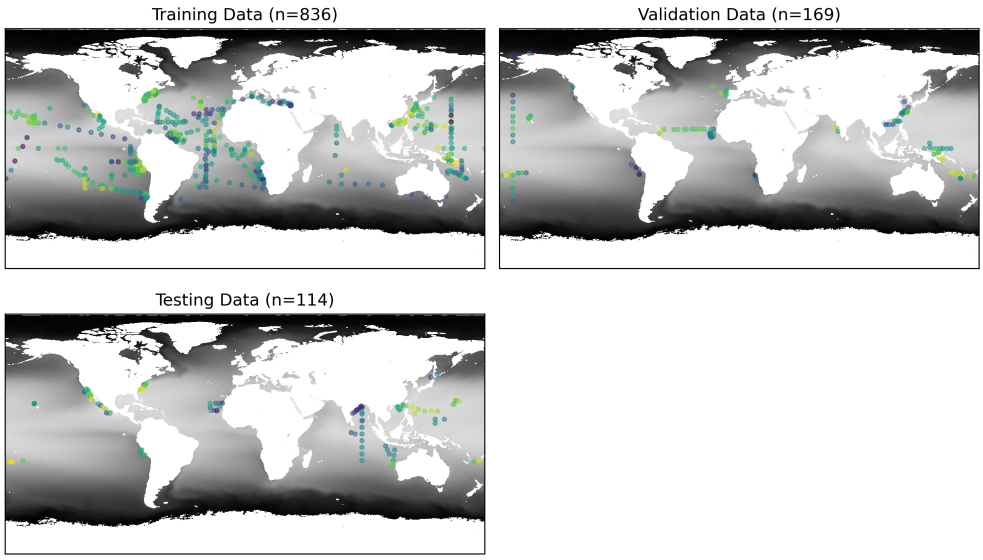

Figure 10: Locations of the N_2 fixation rate measurements used for training (others), validation (2013-2015), and testing (2016-2019) datasets. Data points are color-coded according to the corresponding N_2 fixation rates. The gray background represents monthly sea surface temperature (SST) in August.

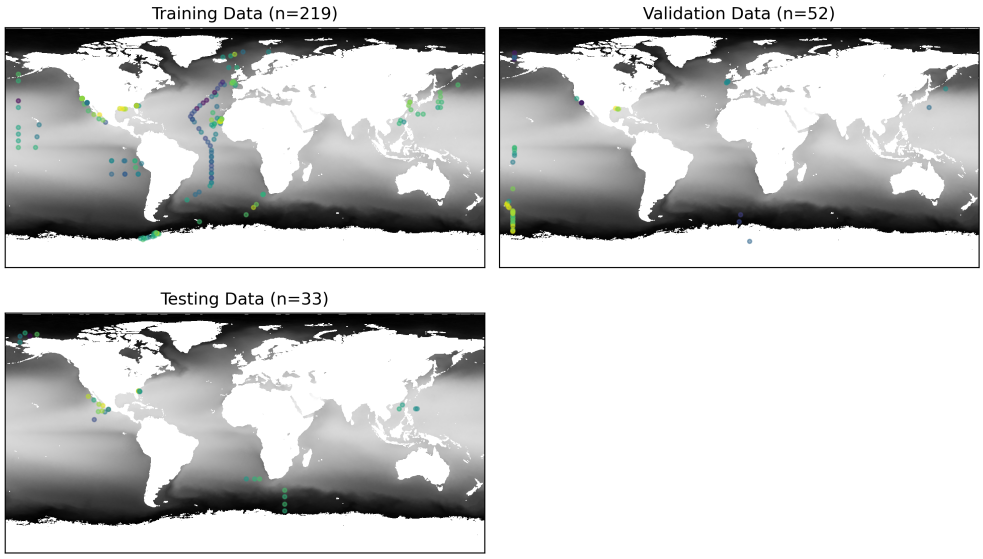

Figure 11: Locations of nitrification measurements used for training (others), validation (2016), and testing (2017-2019) datasets. Data points are color-coded according to the corresponding nitrification values. The gray background represents monthly sea surface temperature (SST) in August.

The uncertainty for prediction is quantified as one standard deviation of the prediction accuracies. For each OcMAE model variant, we perform 11 independent fine-tuning runs. From each run, we extract the prediction accuracy at the epoch with the optimal validation performance, along with the accuracies from the immediately preceding and subsequent epochs to account for local variability. This results in a total of 33 prediction accuracy values for the testing dataset per downstream task for each model. For the baselines without fine-tuning, the prediction accuracy values are obtained by

Table 5: Model variant configurations. *Encoder* denotes the number of encoder layers; *Bins* the number of discretization bins used to partition each input variable; *Dimension* the embedding dimension per bin; *Sampling* the proportion of pixels randomly sampled from each 8-day average image; *Image size* the spatial dimensions of the input (row×column); and *Time length* the number of consecutive 8-day averages for input variable SST, Chl, PAR, Wind speed, Wind U, and Wind V. The number of inputs for MLD and SSS does not change.

MODEL	ENCODER	BINS	DIMENSION	SAMPLING	IMAGE SIZE	TIME LENGTH
BASE	16	1024	64	10%	3×3	3
BIN32	16	32	64	10%	3×3	3
BIN256	16	256	64	10%	3×3	3
BIN512	16	512	64	10%	3×3	3
BIN2048	16	2048	64	10%	3×3	3
L4	4	1024	64	10%	3×3	3
L8	8	1024	64	10%	3×3	3
L32	32	1024	64	10%	3×3	3
D32	16	1024	32	10%	3×3	3
D128	16	1024	128	10%	3×3	3
S5	16	1024	64	10%	5×5	3
S7	16	1024	64	10%	7×7	3
T2	16	1024	64	10%	3×3	2
T5	16	1024	64	10%	3×3	5
PD_FEW	16	1024	64	256 PER 8-DAY	3×3	3
PD_0.3	16	1024	64	30%	3×3	3
PD_REG	16	1024	64	NORTH HEM.	3×3	3

training models on 33 bootstrap resamples of the training dataset. With over 30 samples, the sampling distribution of the mean is approximately Gaussian by the Central Limit Theorem, justifying the use of the sample mean and standard deviation to summarize performance and uncertainty.

E SETUP FOR THE OTHER MODELS

We benchmark our model against LR, RF, XGB, MLP, and TabPFN. Among these models, LR and MLP cannot intrinsically handle missing values. In our experiments, we used the mean of the image cube to fill the missing values of corresponding features for these algorithms. For the other algorithms, we did not fill the missing values. Since LR and tree-based models are insensitive to the scale of the training data, we use the original data as the input. For the MLP and TabPFN, we standardize (z-normalization) the input and output variables. TabPFN cannot consume all the training data for the pCO_2 task (over 2,000,000 data points). We sampled every 20 data points to approximate the distribution of the training data and ran the test for every 10,000 data points, which were concatenated to get the prediction of the testing data.

We used default parameters for LR and TabPFN. For tree-based models and MLP, we searched for the optimal hyperparameters through grid-search technique. The optimal hyperparameters for the tree-based models are shown in Table 6. The optimal hyperparameters for the MLP are shown in Table 7. Due to computational constraints, we could not search for a large hyperparameter space. The prediction accuracy may be improved by searching for better hyperparameters for these algorithms.

F TRACER TRANSPORT EQUATION

We consider the horizontal two-dimensional tracer transport equation, which describes spatiotemporal evolution of surface ocean properties inferred from satellite remote sensing observations. For an incompressible fluid (e.g., the ocean), the discrete form of the equation is written as follows:

$$C_{i,j}^{n+1} = C_{i,j}^{n} - \Delta t\, A_{i,j}^{n} + \Delta t\, D_{i,j}^{n} + \Delta t\, S_{i,j}^{n}, \tag{12}$$

where $C_{i,j}^{n}$ is the tracer concentration, which in this study can represent biomass concentration (biological state) or temperature (physical state). $A_{i,j}^{n}$, $D_{i,j}^{n}$, and $S_{i,j}^{n}$ represent the advection,

Table 6: Hyper-parameters for random forests (RF) and extreme gradient boost (XGB). Each value in the list of number of estimates are used for surface-ocean pCO_2, surface-ocean O_2 concentration, air-sea O_2 flux, oceanic NCP over the mixed layer, oceanic GPP integrated over the euphotic zone, oceanic CR integrated over the euphotic zone, and marine N_2 fixation rate, respectively. The numbers of estimates for NPP and nitrification are 100.

HYPER-PARAMETER	RF	XGB
ESTIMATES	[1000, 200, 200, 100, 100, 100, 50]	[1000, 200, 200, 100, 100, 100, 50]
LEARNING RATE	NONE	0.1
MAXIMUM DEPTH	90	6
L_2-REGULARIZATION	NONE	10
MIN_SAMPLES_SPLIT	MAX(0.1% TRAINING DATA, 2)	NONE
MAX_FEATURES	0.5	ALL

Table 7: Hyper-parameters for multi-layer perceptron (MLP). Hidden layers and dimension for NPP and nitrification are 2 and 16.

HYPER-PARAMETER	pCO_2	O_2	O_2 FLUX	NCP	GPP	CR	N_2 FIX
HIDDEN LAYERS	8	12	8	4	1	2	1
HIDDEN DIMENSION	64	64	64	32	16	16	16
BATCH	512	512	512	512	512	512	512
LEARNING RATE	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
WEIGHT DECAY	0.05	0.05	0.05	0.05	0.05	0.05	0.05

diffusion, and source-sink terms, respectively. The indices i, j, and n corresponds to the x axis (east–west direction; positive eastward), y axis (north-south direction; positive northward), and the discrete time step, respectively, and Δt is the time increment between two adjacent time steps.

The horizontal advection term $A^n_{i,j}$ in Equation (12) can be expressed as:

$$A^n_{i,j} = \left(u\frac{\partial C}{\partial x}\right)^n_{i,j} + \left(v\frac{\partial C}{\partial y}\right)^n_{i,j}, \tag{13}$$

where the two terms on the right-hand side of Equation (13) can be written as follows:

$$\left(u\frac{\partial C}{\partial x}\right)^n_{i,j} = u^n_{i,j}\frac{C^n_{i+1,j} - C^n_{i-1,j}}{2\Delta x}, \tag{14}$$

$$\left(v\frac{\partial C}{\partial y}\right)^n_{i,j} = v^n_{i,j}\frac{C^n_{i,j+1} - C^n_{i,j-1}}{2\Delta y}. \tag{15}$$

Equations (14-15) also can be calculated as the difference between adjacent grids as follows:

$$\left(u\frac{\partial C}{\partial x}\right)^n_{i,j} = \begin{cases} u^n_{i,j}\frac{C^n_{i,j} - C^n_{i-1,j}}{\Delta x}, & u^n_{i,j} > 0, \\ u^n_{i,j}\frac{C^n_{i+1,j} - C^n_{i,j}}{\Delta x}, & u^n_{i,j} < 0, \end{cases} \tag{16}$$

$$\left(v\frac{\partial C}{\partial y}\right)^n_{i,j} = \begin{cases} v^n_{i,j}\frac{C^n_{i,j} - C^n_{i,j-1}}{\Delta x}, & v^n_{i,j} > 0, \\ v^n_{i,j}\frac{C^n_{i,j+1} - C^n_{i,j}}{\Delta x}, & v^n_{i,j} < 0. \end{cases} \tag{17}$$

where horizontal velocities $u^n_{i,j}$ and $v^n_{i,j}$ are positive toward the east and north, respectively.

The diffusion term $D^{n+1}_{i,j}$ in Equation (12) can be written as:

$$D^{n+1}_{i,j} = \left(K_x\frac{C^n_{i+1,j} - 2C^n_{i,j} + C^n_{i-1,j}}{\Delta x^2} + K_y\frac{C^n_{i,j+1} - 2C^n_{i,j} + C^n_{i,j-1}}{\Delta y^2}\right), \tag{18}$$

where K_x and K_y represent the diffusion coefficients in the longitudinal and latitudinal directions, respectively.

The source-sink term $S^n_{i,j}$ in Equation (12) varies with tracer. To derive a general model, we assume that the source-sink term $S^n_{i,j}$ has the following form:

$$S^n_{i,j} = f(Q^{1,n}_{i,j}, Q^{2,n}_{i,j}, ..., Q^{p,n}_{i,j}), \tag{19}$$

where $Q^{:,n}_{i,j}$ represents factors that regulate $S^n_{i,j}$ and p represents the number of controlling factors. Equation (19) can be decomposed with respect to time based on controlling factors as follows:

$$\Delta S^n_{i,j} = \sum_{k=1}^{p} \Delta Q^{k,n}_{i,j} \left. \frac{\partial S^n_{i,j}}{\partial Q^{k,.}_{i,j}} \right|_{Q^{k,:}_{i,j}=Q^{k,n}_{i,j}}, \tag{20}$$

where $\Delta S^n_{i,j}$ and $\Delta Q^{k,n}_{i,j}$ are defined as temporal difference as follows:

$$\Delta S^n_{i,j} = S^{n+1}_{i,j} - S^n_{i,j}, \tag{21}$$

$$\Delta Q^{k,n}_{i,j} = Q^{k,n+1}_{i,j} - Q^{k,n}_{i,j}. \tag{22}$$

Substituting Equations (13-20) into Equation (12) gives the following equation:

$$\begin{aligned} C^{n+1}_{i,j} - C^n_{i,j} &+ \Delta t \left(u^n_{i,j} \frac{C^n_{i+1,j} - C^n_{i-1,j}}{2\Delta x} + v^n_{i,j} \frac{C^n_{i,j+1} - C^n_{i,j-1}}{2\Delta y} \right) \\ &- \Delta t \left(K_x \frac{C^n_{i+1,j} - 2C^n_{i,j} + C^n_{i-1,j}}{\Delta x^2} + K_y \frac{C^n_{i,j+1} - 2C^n_{i,j} + C^n_{i,j-1}}{\Delta y^2} \right) \\ &- \Delta t \left(S^n_{i,j} + \sum_{k=1}^{p} \Delta Q^{k,n}_{i,j} \left. \frac{\partial S^n_{i,j}}{\partial Q^{k,.}_{i,j}} \right|_{Q^{k,:}_{i,j}=Q^{k,n}_{i,j}} \right) = 0. \end{aligned} \tag{23}$$

Equation (23) can be expressed as follows after some mathematical reorganizations:

$$\begin{aligned} C^{n+1}_{i,j} &+ \left(-1 + \frac{2\Delta t K_x}{\Delta x^2} + \frac{2\Delta t K_y}{\Delta y^2} \right) C^n_{i,j} \\ &+ \Delta t \left(\frac{u^n_{i,j}}{2\Delta x} - \frac{K_x}{\Delta x^2} \right) C^n_{i+1,j} \\ &+ \Delta t \left(-\frac{u^n_{i,j}}{2\Delta x} - \frac{K_x}{\Delta x^2} \right) C^n_{i-1,j} \\ &+ \Delta t \left(\frac{v^n_{i,j}}{2\Delta y} - \frac{K_y}{\Delta y^2} \right) C^n_{i,j+1} \\ &+ \Delta t \left(-\frac{v^n_{i,j}}{2\Delta y} - \frac{K_y}{\Delta y^2} \right) C^n_{i,j-1} \\ &- \Delta t \sum_{k=1}^{p} Q^{k,n+1}_{i,j} \left. \frac{\partial S^n_{i,j}}{\partial Q^{k,.}_{i,j}} \right|_{Q^{k,:}_{i,j}=Q^{k,n}_{i,j}} \\ &+ \Delta t \sum_{k=1}^{p} Q^{k,n}_{i,j} \left. \frac{\partial S^n_{i,j}}{\partial Q^{k,.}_{i,j}} \right|_{Q^{k,:}_{i,j}=Q^{k,n}_{i,j}} \\ &- \Delta t S^n_{i,j} = 0. \end{aligned} \tag{24}$$

Equation (24) suggests that the model pretrained on input $\left[C^{n+1}_{i,j}, C^n_{i,j}, C^n_{i+1,j}, C^n_{i-1,j}, C^n_{i,j+1}, C^n_{i,j-1}, Q^{k,n+1}_{i,j}, Q^{k,n}_{i,j}\right]$ learns equation coefficients. The spatially adjacent grids constrain the advection and diffusion processes, while temporal masking and reconstruction of controlling factors enable the model to learn the sensitivity of source–sink term to their temporal variability. Equations (23 and 24) can also be derived using Equations (16 and 17), leading to a slightly different equation that is not shown here.

In marine biogeochemistry, tracers can be classified into physical and biological components. Key biological components include nutrient concentration (e.g., nitrate), phytoplankton biomass concentration, detritus, and zooplankton biomass concentration. Key physical components include temperature.

Among these variables, satellite observations can be used to derive sea surface temperature (SST) and chlorophyll concentration (Chl) which is commonly used as a proxy of phytoplankton biomass concentration. Usually, Chl and SST are sufficient to capture key variability in most of the oceanic biogeochemical processes. Here, we derive the source-sink term of Chl as an example. Similar approach can be used to develop the source-sink term for SST, which will not be shown here.

The source-sink term $S^n_{i,j}$ for Chl is influenced by factors including phytoplankton growth through photosynthesis (i.e., gross primary production, GPP), grazing by zooplankton, phytoplankton modality, and sinking out of the surface ocean due to gravitation. Downstream tasks GPP, NCP, and CR are encapsulated in these processes. Primary productivity at the surface ocean includes GPP, net primary production NPP=GPP-AR where AR represent autotrophic respiration due to phytoplankton, and net community production NCP=GPP-CR where CR denotes community respiration due to both phytoplankton and bacteria. Primary productivity is influenced by factors including Chl, light availability, SST, and nutrient availability. Based on data availability, we assume that the $S^n_{i,j}$ term is a function of Chl, light availability, nutrient concentration, and SST:

$$S^n_{i,j} = f(C^n_{i,j}, I^n_{i,j}, T^n_{i,j}, N^n_{i,j}), \tag{25}$$

where $I^n_{i,j}$, $T^n_{i,j}$, and $N^n_{i,j}$ represent photosynthetically active radiation (PAR), SST, and nutrient concentration, respectively. $S_{i,j}$ can be further decomposed as follows:

$$\Delta S^n_{i,j} = \left.\frac{\partial S^n_{i,j}}{\partial C_{i,j}}\right|_{C_{i,j}=C^n_{i,j}} \Delta C^n_{i,j} + \left.\frac{\partial S^n_{i,j}}{\partial I_{i,j}}\right|_{I_{i,j}=I^n_{i,j}} \Delta I^n_{i,j} + \left.\frac{\partial S^n_{i,j}}{\partial T_{i,j}}\right|_{T_{i,j}=T^n_{i,j}} \Delta T^n_{i,j} + \left.\frac{\partial S^n_{i,j}}{\partial N_{i,j}}\right|_{N_{i,j}=N^n_{i,j}} \Delta N^n_{i,j}, \tag{26}$$

where $\Delta S^n_{i,j}$, $\Delta C^n_{i,j}$, $\Delta I^n_{i,j}$, $\Delta T^n_{i,j}$, and $\Delta N^n_{i,j}$ represent temporal changes in source-sink term, tracer concentration, light, temperature, and nutrient concentration, respectively:

$$\Delta C^n_{i,j} = C^{n+1}_{i,j} - C^n_{i,j}, \tag{27}$$

$$\Delta I^n_{i,j} = I^{n+1}_{i,j} - I^n_{i,j}, \tag{28}$$

$$\Delta T^n_{i,j} = T^{n+1}_{i,j} - T^n_{i,j}, \tag{29}$$

$$\Delta N^n_{i,j} = N^{n+1}_{i,j} - N^n_{i,j}. \tag{30}$$

Substituting Equations (25-30) into Equation (24) gives:

$$\begin{aligned}
&\left(1-\Delta t\frac{\partial S_{i,j}^n}{\partial C_{i,j}}\bigg|_{C_{i,j}=C_{i,j}^n}\right)C_{i,j}^{n+1}+\left(-1+\frac{2\Delta tK_x}{\Delta x^2}+\frac{2\Delta tK_y}{\Delta y^2}+\Delta t\frac{\partial S_{i,j}^n}{\partial C_{i,j}}\bigg|_{C_{i,j}=C_{i,j}^n}\right)C_{i,j}^n\\
&\quad+\Delta t\left(\frac{u_{i,j}^n}{2\Delta x}-\frac{K_x}{\Delta x^2}\right)C_{i+1,j}^n\\
&\quad+\Delta t\left(-\frac{u_{i,j}^n}{2\Delta x}-\frac{K_x}{\Delta x^2}\right)C_{i-1,j}^n\\
&\quad+\Delta t\left(\frac{v_{i,j}^n}{2\Delta y}-\frac{K_y}{\Delta y^2}\right)C_{i,j+1}^n\\
&\quad+\Delta t\left(-\frac{v_{i,j}^n}{2\Delta y}-\frac{K_y}{\Delta y^2}\right)C_{i,j-1}^n\\
&\quad-\Delta t\frac{\partial S_{i,j}^n}{\partial I_{i,j}}\bigg|_{I_{i,j}=I_{i,j}^n}I_{i,j}^{n+1}\\
&\quad+\Delta t\frac{\partial S_{i,j}^n}{\partial I_{i,j}}\bigg|_{I_{i,j}=I_{i,j}^n}I_{i,j}^n\\
&\quad-\Delta t\frac{\partial S_{i,j}^n}{\partial T_{i,j}}\bigg|_{T_{i,j}=T_{i,j}^n}T_{i,j}^{n+1}\\
&\quad+\Delta t\frac{\partial S_{i,j}^n}{\partial T_{i,j}}\bigg|_{T_{i,j}=T_{i,j}^n}T_{i,j}^n\\
&\quad-\Delta t\frac{\partial S_{i,j}^n}{\partial N_{i,j}}\bigg|_{N_{i,j}=N_{i,j}^n}N_{i,j}^{n+1}\\
&\quad+\Delta t\frac{\partial S_{i,j}^n}{\partial N_{i,j}}\bigg|_{N_{i,j}=N_{i,j}^n}N_{i,j}^n\\
&\quad-\Delta tS_{i,j}^n=0.
\end{aligned}\tag{31}$$

Equation (31) can be generalized to a grid or image (e.g., 3×3):

$$\mathbf{A}^{n+1}\mathbf{C}^{n+1}+\mathbf{A}^n\mathbf{C}^n-\mathbf{B}_I^n\mathbf{I}^{n+1}+\mathbf{B}_I^n\mathbf{I}^n-\mathbf{B}_T^n\mathbf{T}^{n+1}+\mathbf{B}_T^n\mathbf{T}^n-\mathbf{B}_N^n\mathbf{N}^{n+1}+\mathbf{B}_N^n\mathbf{N}^n+\mathbf{S}^n=0,\tag{32}$$

where $\mathbf{C}^n$ and $\mathbf{C}^{n+1}$ are column vectors that represent 3×3 grids (images) of the tracer concentration fields at the time steps of n and $n+1$, respectively:

$$\mathbf{C}^n=\left[C_{i-1,j-1}^n,C_{i-1,j}^n,C_{i-1,j+1}^n,C_{i,j-1}^n,C_{i,j}^n,C_{i,j+1}^n,C_{i+1,j-1}^n,C_{i+1,j}^n,C_{i+1,j+1}^n\right]^\top,\tag{33}$$

$$\mathbf{C}^{n+1}=\left[C_{i-1,j-1}^{n+1},\ldots,C_{i+1,j+1}^{n+1}\right]^\top.\tag{34}$$

and $\mathbf{I}^n$, $\mathbf{T}^n$, and $\mathbf{N}^n$ are column vectors that represent 3×3 grids of forcing fields at the time step n:

$$\mathbf{X}^n=\begin{bmatrix}X_{i-1,j-1}^n & X_{i-1,j}^n & X_{i-1,j+1}^n & X_{i,j-1}^n & X_{i,j}^n & X_{i,j+1}^n & X_{i+1,j-1}^n & X_{i+1,j}^n & X_{i+1,j+1}^n\end{bmatrix}^\top,\tag{35}$$

where $\mathbf{X}\in\{\mathbf{I},\mathbf{T},\mathbf{N}\}$. $\mathbf{S}^n$ is a column vectors defined as follows:

$$\mathbf{S}^n=-\Delta t\left[S_{i-1,j-1}^n,S_{i-1,j}^n,S_{i-1,j+1}^n,S_{i,j-1}^n,S_{i,j}^n,S_{i,j+1}^n,S_{i+1,j-1}^n,S_{i+1,j}^n,S_{i+1,j+1}^n\right]^\top,\tag{36}$$

$\mathbf{A}^n$, $\mathbf{A}^{n+1}$, $\mathbf{B}_I^n$, $\mathbf{B}_T^n$, and $\mathbf{B}_N^n$ in Equation (32) are matrices defined below. $\mathbf{A}^n$ represents advection and diffusion processes:

$$\mathbf{A}^n = \begin{bmatrix} b_c^{i-1,j-1} & b_E^{i-1,j-1} & 0 & b_N^{i-1,j-1} & 0 & 0 & 0 & 0 & 0 \\ b_W^{i,j-1} & b_c^{i,j-1} & b_E^{i,j-1} & 0 & b_N^{i,j-1} & 0 & 0 & 0 & 0 \\ 0 & b_W^{i+1,j-1} & b_c^{i+1,j-1} & 0 & 0 & b_N^{i+1,j-1} & 0 & 0 & 0 \\ b_S^{i-1,j} & 0 & 0 & b_c^{i-1,j} & b_E^{i-1,j} & 0 & b_N^{i-1,j} & 0 & 0 \\ 0 & b_S^{i,j} & 0 & b_W^{i,j} & b_c^{i,j} & b_E^{i,j} & 0 & b_N^{i,j} & 0 \\ 0 & 0 & b_S^{i+1,j} & 0 & b_W^{i+1,j} & b_c^{i+1,j} & 0 & 0 & b_N^{i+1,j} \\ 0 & 0 & 0 & b_S^{i-1,j+1} & 0 & 0 & b_c^{i-1,j+1} & b_E^{i-1,j+1} & 0 \\ 0 & 0 & 0 & 0 & b_S^{i,j+1} & 0 & b_W^{i,j+1} & b_c^{i,j+1} & b_E^{i,j+1} \\ 0 & 0 & 0 & 0 & 0 & b_S^{i+1,j+1} & 0 & b_W^{i+1,j+1} & b_c^{i+1,j+1} \end{bmatrix}, \tag{37}$$

where the entries in the matrix are defined as:

$$b_c^{p,q} = -1 + \frac{2\Delta t K_x}{\Delta x^2} + \frac{2\Delta t K_y}{\Delta y^2} + \Delta t \frac{\partial S_{p,q}^n}{\partial C_{p,q}}\Big|_{C_{p,q}=C_{p,q}^n}, \tag{38}$$

$$b_E^{p,q} = \Delta t \left(\frac{u_{p,q}^n}{2\Delta x} - \frac{K_x}{\Delta x^2} \right), \tag{39}$$

$$b_W^{p,q} = \Delta t \left(-\frac{u_{p,q}^n}{2\Delta x} - \frac{K_x}{\Delta x^2} \right), \tag{40}$$

$$b_N^{p,q} = \Delta t \left(\frac{v_{p,q}^n}{2\Delta y} - \frac{K_y}{\Delta y^2} \right), \tag{41}$$

$$b_S^{p,q} = \Delta t \left(-\frac{v_{p,q}^n}{2\Delta y} - \frac{K_y}{\Delta y^2} \right). \tag{42}$$

$\mathbf{A}^{n+1}$ represents the sensitivity of the source-sink term to the the tracer:

$$\mathbf{A}^{n+1} = \mathbf{I} - \Delta t \begin{bmatrix} a_{i-1,j-1}^n & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_{i-1,j}^n & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{i-1,j+1}^n & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{i,j-1}^n & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{i,j}^n & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{i,j+1}^n & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{i+1,j-1}^n & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{i+1,j}^n & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{i+1,j+1}^n \end{bmatrix}, \tag{43}$$

where $a_{p,q}^n = \frac{\partial S_{p,q}^n}{\partial C_{p,q}}\Big|_{C_{p,q}=C_{p,q}^n}$.

$\mathbf{B}_I^n$, $\mathbf{B}_T^n$, and $\mathbf{B}_N^n$, represent the sensitivity of the source-sink term to the forcing factors defined as follows:

$$\mathbf{B}_F^n = \Delta t \begin{bmatrix} d_{i-1,j-1}^{F,n} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_{i-1,j}^{F,n} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_{i-1,j+1}^{F,n} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_{i,j-1}^{F,n} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{i,j}^{F,n} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{i,j+1}^{F,n} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & d_{i+1,j-1}^{F,n} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & d_{i+1,j}^{F,n} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & d_{i+1,j+1}^{F,n} \end{bmatrix}, \tag{44}$$

where $d_{p,q}^{F,n} = \frac{\partial S_{p,q}^n}{\partial F_{p,q}}\Big|_{F_{p,q}=F_{p,q}^n}$ and F can be light (I), temperature (T), and nutrient (N).

Given input data $\mathbf{C^n}$, $\mathbf{I^n}$, $\mathbf{T^n}$, and $\mathbf{N^n}$, reconstructing a randomly masked subset of values in the spatiotemporal field of Chl and the associated controlling factors from their visible values enables the model to learn the following processes:

- **Physical processes:** reconstructing masked Chl from spatial neighboring visible values allows the model to learn the contribution of advection and diffusion processes to the local change of Chl (i.e., matrix $\mathbf{A^n}$), as a result capturing the spatial pattern of the Chl field. Importantly, $\mathbf{A^n}$ also shows the boundary effect for pixels at the boundaries.
- **Source-sink processes:** learning the pattern of source-sink processes from its temporal change with respect to the associated controlling factors, such as tracer concentration and other environmental variables (e.g., nutrient, light, and temperature). The model acquires these knowledge through learning the matrices of $\mathbf{A}^{n+1}$, $\mathbf{B}_I^n$, $\mathbf{B}_T^n$, and $\mathbf{B}_N^n$.

For the tracer of SST, advection and diffusion processes, i.e. Equations (12-18), remain the same. Its source-sink term is primarily driven by air-sea heat flux and mixed layer dynamics, which is closely linked to the physical controls on the O_2 at the surface ocean, one of our downstream tasks. Taken together, algorithm design is guided by the following considerations:

- **Operator learning**. The tracer transport equation suggests that the model learns an operator representing the underlying data-generating processes. The mathematical foundation of this approach is supported by the universal approximation (Lu et al., 2021). Additionally, transformers do not introduce inductive bias into adjacent grids, as a result allowing Δx to vary with latitudes. This variation can be captured by geographical location, which is represented as a single value for each image cube and added to individual pixels. Other region-dependent factors, which are not captured by the other input variables, may also be encoded by the geographical location and seasonality (the day of year).
- **Collinearity**. Equations (24 and 31) indicate that spatiotemporal correlations in the input variables can hinder the model from learning efficient data representations due to collinearity. A high masking rate, which acts similarly to feature selection, can reduce such correlation and facilitate model training. This observation is consistent with previous studies showing that higher masking rates can lead to more robust representations in image and video datasets (He et al., 2022; Feichtenhofer et al., 2022).
- **Spatial and temporal scales**. The spatial and temporal scales of the input variables are closely linked to their degree of collinearity. In this study, satellite remote sensing datasets have been resampled to a spatial resolution of $0.083^\circ \times 0.083^\circ$ and a temporal resolution of 8-day averages. In some regions, particularly subtropical area, the input variables are highly correlated at these spatiotemporal scales. Therefore, a high masking rate helps the model learn robust spatiotemporal representations of the underlying data generating processes.
- **Downstream tasks**. Model performance on downstream tasks depends on factors such as nonlinearity and their relevance to the pretraining datasets. **First**, dissolved O_2 at the surface ocean is largely controlled by solubility, which is well described by temperature. Because temperature is included in the pretraining data, this type of downstream tasks is expected to be well predicted by the model. **Second**, for downstream tasks associated with components of the source-sink term, prediction accuracy is expected to improve as pretraining implicitly guilds the model to learn the physical processes from the input variables, acting as prior from a Bayesian perspective. At the same time, fine-tuning dataset is critical for isolating the relevant processes within the source-sink term. **Third**, for downstream tasks whose dependent variables (or some) are not included in the input variables, the system is analogous to a dynamical system with missing states, making predictions more depend on initial states (Brajard et al., 2020; Cheng et al., 2023). This is consistent with the downstream tasks of GPP and N_2 fixation rate, which depend on additional input variables such as iron concentration. Consequently, the model performs poorly on these tasks when the number of labeled samples is limited.

Finally, the influence of nutrient concentration on the oceanic biogeochemical process (e.g., GPP and NPP) varies across the world's oceans. For example, the ecosystem in the subtropical regions is typically limited by the availability of nitrate, while the ecosystem in the Southern Ocean is commonly limited by iron. In addition, there are sparse nutrient concentration measurements. Although there are

Table 8: The prediction accuracies (R^2) of GPP, NPP, nitrification, and N_2 fixation rate (N_2FIX) derived by different model variants. Model configuration is provided in Table 5. The uncertainties are quantified as one standard deviation (see Appendix D).

MODEL	GPP	N_2FIX	NPP	NITRIFICATION
BASE	-0.047 ± 0.056	-0.050 ± 0.020	0.008 ± 0.024	0.046 ± 0.045
BIN32	0.018 ± 0.056	-0.053 ± 0.021	0.023 ± 0.017	0.049 ± 0.045
BIN256	-0.044 ± 0.053	-0.053 ± 0.021	0.012 ± 0.023	0.049 ± 0.044
BIN512	0.042 ± 0.008	-0.052 ± 0.021	0.010 ± 0.024	0.047 ± 0.045
BIN2048	-0.032 ± 0.043	-0.054 ± 0.022	0.007 ± 0.025	0.048 ± 0.044
L4	-0.055 ± 0.022	-0.014 ± 0.044	0.007 ± 0.025	0.050 ± 0.045
L8	0.158 ± 0.123	0.031 ± 0.010	0.006 ± 0.014	0.049 ± 0.044
L32	-0.051 ± 0.050	0.014 ± 0.020	0.009 ± 0.024	0.046 ± 0.045
D32	-0.051 ± 0.020	-0.593 ± 0.074	0.008 ± 0.023	0.048 ± 0.045
D128	-0.057 ± 0.023	0.064 ± 0.012	0.006 ± 0.015	0.048 ± 0.045
S5	-0.052 ± 0.020	-0.125 ± 0.042	0.008 ± 0.024	0.046 ± 0.045
S7	-0.052 ± 0.021	0.055 ± 0.031	0.005 ± 0.015	0.044 ± 0.046
T2	-0.062 ± 0.042	-0.053 ± 0.021	0.014 ± 0.022	0.048 ± 0.045
T5	0.179 ± 0.037	0.229 ± 0.033	-0.047 ± 0.007	0.113 ± 0.034
PD_FEW	0.073 ± 0.037	0.031 ± 0.047	0.017 ± 0.023	0.050 ± 0.044
PD_0.3	-0.053 ± 0.021	0.020 ± 0.006	0.011 ± 0.023	0.047 ± 0.045
PD_REG	-0.053 ± 0.021	-0.598 ± 0.078	0.019 ± 0.023	0.050 ± 0.045
LR	-0.755 ± 11.30	-0.292 ± 0.470	-0.023 ± 0.055	-0.305 ± 0.153
RF	-0.083 ± 0.151	-0.063 ± 0.074	0.313 ± 0.041	-0.058 ± 0.119
XGB	-0.183 ± 0.178	-0.004 ± 0.104	0.264 ± 0.064	-0.088 ± 0.182
MLP	-0.431 ± 0.357	-0.063 ± 0.487	-0.609 ± 1.123	-0.179 ± 0.072
TABPFN	-0.108 ± 0.134	-0.267 ± 0.093	0.376 ± 0.049	0.021 ± 0.078

many different nutrient data simulated by numerical models, we prefer to use wind speed and mixed layer depth data as a proxy of nutrient availability, which is based on the assumption that nutrient is supplied from beneath through mixed layer depth variation driven by wind speed.

G MODEL PERFORMANCE ON GPP, NPP, NITRIFICATION, AND N_2 FIXATION TASKS

Model variants and baseline models performance on the GPP and N_2 fixation tasks are shown in Table 8. Negative values in some predictions mean that model predictions are worse than average.

H PREDICTIONS OF pCO$_2$, O$_2$, AIR-SEA O$_2$ FLUX, AND CR BY OCMAE

The predicted distributions of global pCO$_2$, O$_2$, air-sea O$_2$ flux, and CR estimates are shown in Figures 12-15).

I PREDICTIONS OF BENCHMARK MODELS

Global distributions of the first 8-day pCO$_2$, surface-ocean O$_2$, air-sea O$_2$ flux, NCP, and CR in 2022 by XGB, RF, and TabPFN are shown in Figures 16-18.

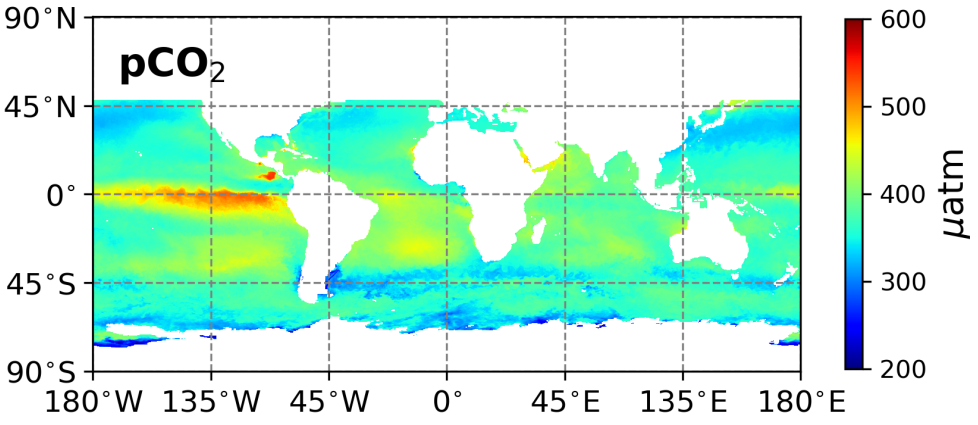

Figure 12: Global distribution of the first 8-day pCO_2 in 2022 estimated by OcMAE.

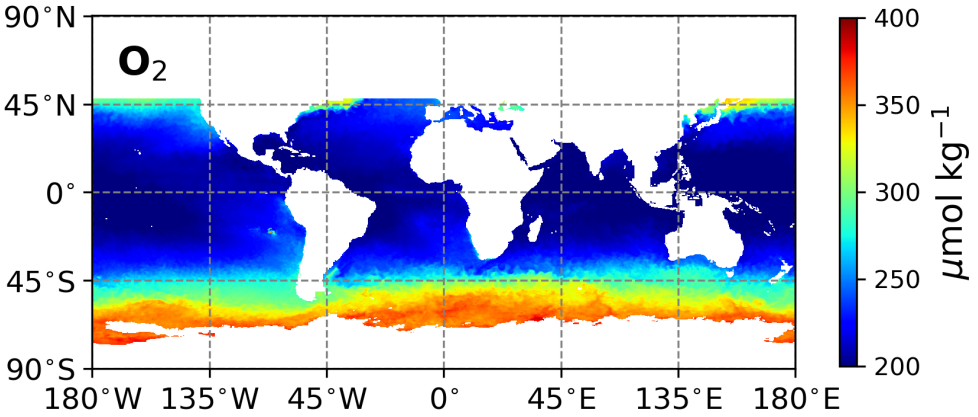

Figure 13: Global distribution of the first 8-day O_2 in 2022 estimated by OcMAE.

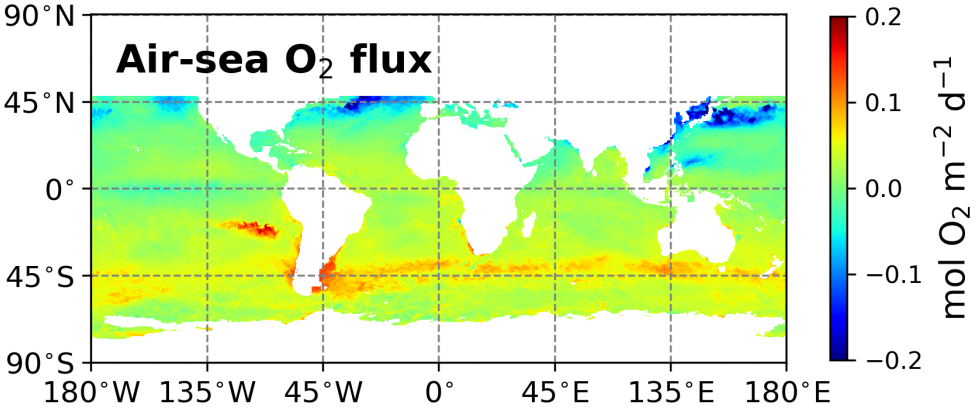

Figure 14: Global distribution of the first 8-day air-sea O_2 flux in 2022 estimated by OcMAE.

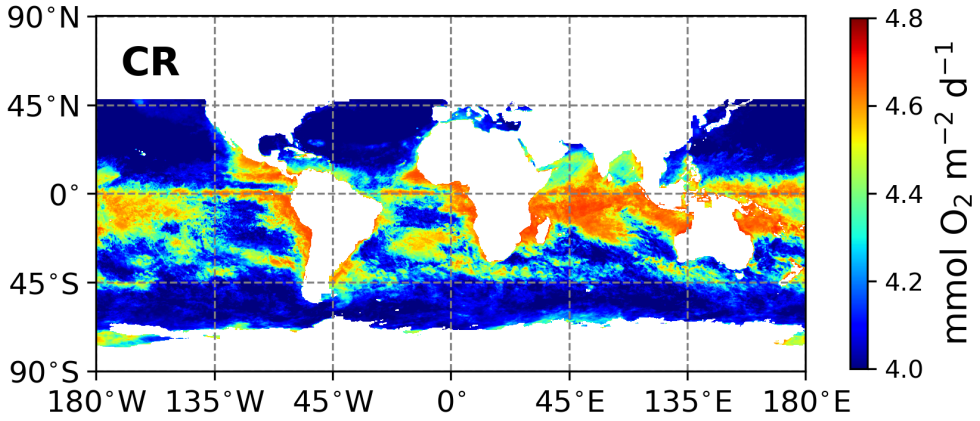

Figure 15: Global distribution of the first 8-day community respiration (CR) on the logarithmic space in 2022 estimated by OcMAE.

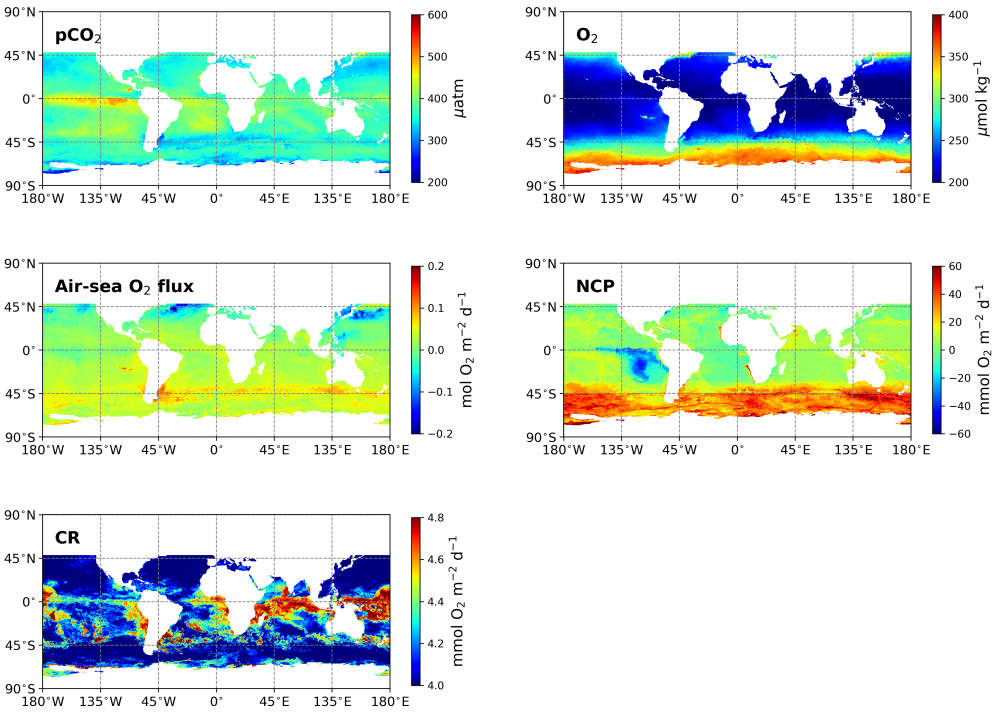

Figure 16: Global distributions of the first 8-day pCO_2, dissolved oxygen (O_2), air-sea oxygen flux, net community production (NCP), and community respiration (CR) on the logarithmic space in 2022 estimated by extreme gradient boosting.

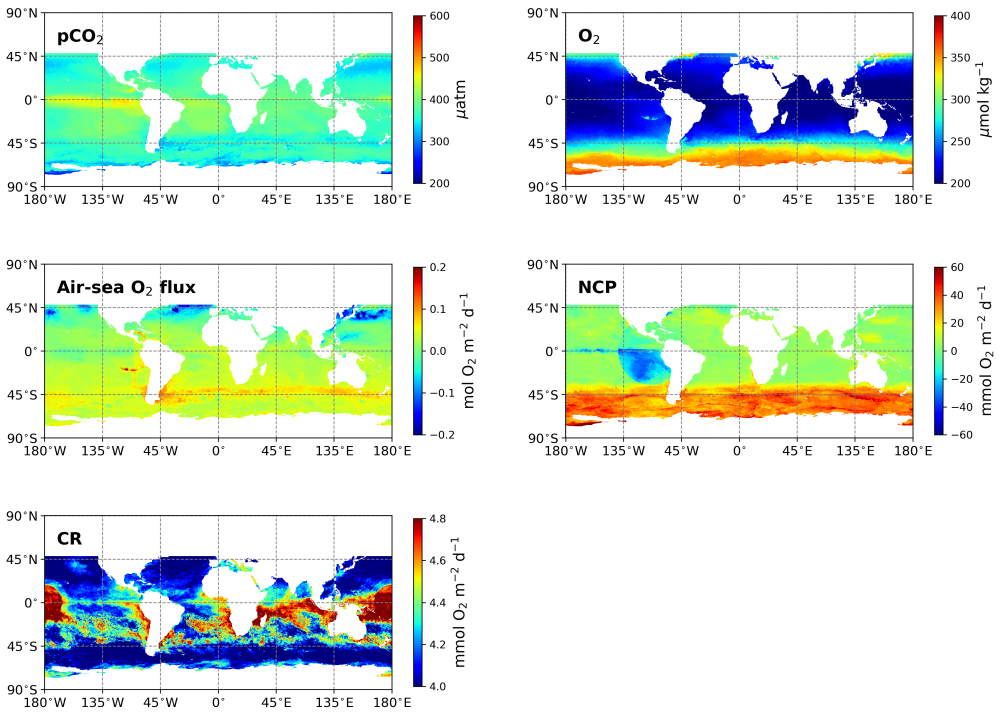

Figure 17: Global distributions of the first 8-day pCO_2, dissolved oxygen (O_2), air-sea oxygen flux, net community production (NCP), and community respiration (CR) on the logarithmic space in 2022 estimated by random forests. The notable artifacts in NCP estimates in the eastern equatorial Pacific arise from input variable latitude and longitude, which are given high weights by the model.

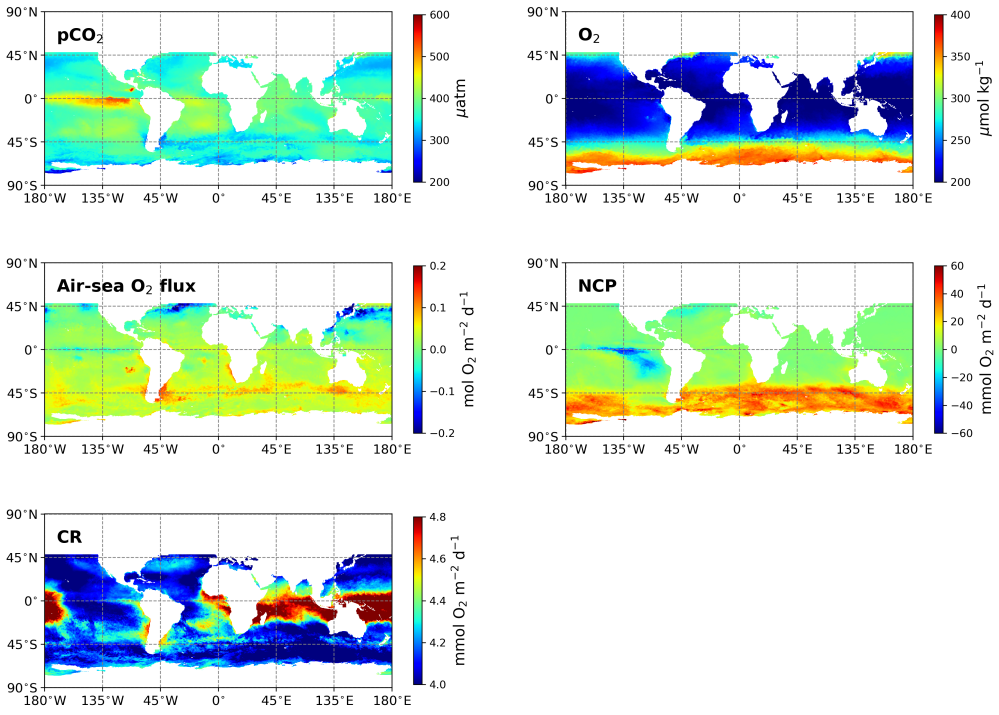

Figure 18: Global distributions of the first 8-day pCO_2, dissolved oxygen (O_2), air-sea oxygen flux, net community production (NCP), and community respiration (CR) on the logarithmic space in 2022 estimated by tabular prior-data fitted networks.