Geophysical Research Letters

RESEARCH LETTER
10.1029/2024GL112792

# A Machine Learning-Based Dissolved Organic Carbon Climatology

**Thelma Panaïotis**[1], **Jamie Wilson**[2], **and BB Cael**[1,3]

[1]National Oceanography Centre, Southampton, UK, [2]University of Liverpool, Liverpool, UK, [3]Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois, USA

**Key Points:**

- A machine learning model was fitted to relate DOC observations to other environmental variables (e.g., temperature, dissolved oxygen)
- Inferred relationships between environmental predictors and DOC were used to generate layer-wise climatologies of DOC
- By integrating our predictions to the global ocean, we propose a refined estimate of the total DOC content of 690 Pg C

**Supporting Information:**

Supporting Information may be found in the online version of this article.

**Correspondence to:**

T. Panaïotis,
thelma.panaiotis@noc.ac.uk

**Citation:**

Panaïotis, T., Wilson, J., & Cael, B. (2025). A machine learning-based dissolved organic carbon climatology. *Geophysical Research Letters*, *52*, e2024GL112792. https://doi.org/10.1029/2024GL112792

Received 30 SEP 2024
Accepted 17 FEB 2025

**Author Contributions:**

**Conceptualization:** Thelma Panaïotis, Jamie Wilson, BB Cael
**Data curation:** Thelma Panaïotis
**Formal analysis:** Thelma Panaïotis
**Funding acquisition:** BB Cael
**Software:** Thelma Panaïotis
**Supervision:** BB Cael
**Validation:** Thelma Panaïotis, Jamie Wilson
**Visualization:** Thelma Panaïotis
**Writing – original draft:** Thelma Panaïotis, Jamie Wilson
**Writing – review & editing:** Thelma Panaïotis, Jamie Wilson, BB Cael



**Abstract** Marine dissolved organic carbon (DOC) is a major carbon reservoir influencing climate, but is poorly quantified. The lack of a comprehensive DOC climatology hinders model validation, estimation of the modern DOC inventory, and understanding of DOC's role in the carbon cycle and climate. To address this problem, we used boosted regression trees to relate a compilation of DOC observations to different environmental climatologies, and extrapolated these inferred relationships to the entire ocean to compute annual layer-wise DOC climatologies with uncertainties. Prediction performance was satisfactory, with $R^2$ values within 0.6–0.8 for all layers and prediction error comparable to within-pixel measurement variability. DOC was mainly predicted by dissolved oxygen in the bathypelagic layer, and by nutrients in other layers. We estimate the total oceanic DOC inventory to be around 690 Pg C. Our results exemplify that machine learning is a powerful tool for constructing climatologies from limited observations.

**Plain Language Summary** Marine dissolved organic carbon (DOC) is a large and important component of the Earth's carbon cycle that influences climate. However, we do not have a good understanding of how much DOC is in the oceans. This lack of information makes it difficult to improve climate models and fully understand how DOC affects the global carbon cycle. To address this, we used a machine learning technique (boosted regression trees) to relate available DOC data to environmental factors. We then applied this analysis to the entire ocean to produce annual estimates of DOC concentrations, along with the associated uncertainties. Our model performed well, explaining between 60% and 80% of the variance across different ocean layers. We found that dissolved oxygen seems to the main factor influencing DOC in deep waters, while nutrients were more important in the upper layers. We estimate that the total amount of DOC in the ocean is about 690 billion tonnes of carbon. Our work shows that machine learning can be a useful tool to generate global estimates from limited data.

## 1. Introduction

Ocean dissolved organic carbon (DOC) is a major reservoir of carbon in the ocean-atmosphere-climate system (662 Pg C: Hansell, 2013), comparable in size to the preindustrial atmosphere. The reservoir is also characterized by a bulk radiocarbon age in the deep ocean (>1,000 m) of between 4,000 and 6,000, years requiring DOC to survive multiple cycles of ocean overturning. The size and apparent persistence of DOC has fueled interest in DOC as an additional facet of the marine carbon cycle, but there remains considerable uncertainty about the processes driving these features (Arrieta et al., 2015; Hansell, 2013). DOC production is mainly due to ecological processes (e.g., zooplankton grazing, virus-induced phytoplankton cell lysis) in the euphotic layer and solubilization from marine snow aggregates, while its removal is due to remineralization by bacteria or abiotic processes (e.g., UV oxidation or adsorption to particles) (Carlson & Hansell, 2015). Despite being one of the least constrained parts of the global carbon cycle, DOC is widely expected to be a reactive reservoir that is capable of impacting climate (Sexton et al., 2011; Wagner et al., 2020).

Understanding the cycling of DOC and its associated impact on the Earth system relies fundamentally on robust estimates of the global DOC inventory and resolving the spatial patterns of DOC concentrations. The DOC inventory constrains the climate impact of DOC as changes in marine carbon reservoirs scale predictably to changes in atmospheric $CO_2$ (Goodwin et al., 2008). Spatial gradients in DOC concentrations are key to diagnosing DOC remineralization rates (Hansell & Carlson, 2013; Sulpis et al., 2023) that provide a basis for supporting the varied hypotheses of DOC persistence. Inverse estimates of DOC export production (Roshan & DeVries, 2017) and numerical modeling of past and future changes in DOC cycling (Gilchrist & Matsumoto, 2023; Matsumoto

et al., 2022) also rely on constraints from spatial patterns (e.g., at a 1° horizontal resolution) of DOC concentrations.

A key challenge for global DOC observations as constraints is the relative sparsity of observations compared to standard quantities such as macronutrients (Hansell et al., 2021; Letscher & Moore, 2015). Spatial gradients have previously been interpolated to a higher resolution by assimilating observations in a numerical model (Hansell et al., 2012) and using artificial neural networks as part of a wider study on DOC export (Roshan & DeVries, 2017). Both approaches provide a broad climatological view of the horizontal and vertical distribution of DOC. Temporally resolved products limited to the surface ocean have also been derived from satellite products (Aurin et al., 2018; Bonelli et al., 2022; Siegel et al., 2002) but typically focus on the subset of DOC compounds that are optically measurable.

In this study, we take advantage of machine learning (ML) to relate DOC observations to other variables for which climatologies are already available, and use the inferred relationships to predict DOC values where no observations are available, hence generating a DOC climatology. We build upon the work of Roshan and DeVries (2017) by using a more comprehensive DOC data set as well as a more robust machine learning algorithm (Hastie et al., 2009). Indeed, compared to artificial neural networks (i.e., multilayer perceptron), tree-based ensemble methods (e.g., boosted trees, random forest) are less prone to overfitting and more effective at handling common challenges with predictors, such as missing values, outliers or irrelevant variables. With respect to recent approaches by Laine et al. (2024) and Bonelli et al. (2022), we focus on the annual and seasonal scales and we provide DOC estimates for deeper layers. Furthermore, we also detail the relationships between the output and predictors. Finally, we provide a ML-based estimate of total DOC by integrating predicted DOC content across oceanic layers.

## 2. Materials and Methods

### 2.1. Data Inputs and Processing

#### 2.1.1. DOC

DOC observations come from a global compilation of dissolved organic matter collected between 1994 and 2021 (Hansell et al., 2021). An initial quality check discarded problematic observations (e.g., missing depth or date, flag indicating a problem with the observations).

To account for the depth-dependent processes driving the DOC concentration and to provide DOC climatologies suitable for different analyses (e.g., surface climatology for constraining satellite products), DOC observations were assigned to four layers based on their depth: surface (0–10 m), epipelagic (10–200 m), mesopelagic (200–1,000 m) and bathypelagic (>1,000 m). These depth layers are intended to approximately delineate differences in cycling as expected from semi-labile, semi-refractory and refractory DOC which have a strong depth dependence (Hansell, 2013). In addition, surface layer DOC observations were assigned to four meteorological seasons (northern hemisphere) based on the sampling month to construct a surface seasonal climatology. Finally, DOC observations were averaged on a 1° grid to match the environmental predictors described below, resulting in a global coverage of ∼2000 observations for annual predictions, and 500–700 observations for seasonal predictions, with less coverage during winter at high latitudes (Table S1, Figures S1, and S2 in Supporting Information S1).

To focus on autochthonous DOC, we excluded high DOC values from allochthonous sources in the Kara Sea and Laptev Sea (Dittmar & Kattner, 2003). Specifically, we discarded 96 surface layer pixels (4.2% of the 2,301 originally available) located within the polygons 50–100°E, 70–80°N (Kara Sea) and 100–150°E, 70–80°N (Laptev Sea). These excluded DOC values ranged from 77 to 567 μmol $kg^{-1}$, with a median of 240 μmol $kg^{-1}$.

#### 2.1.2. Environmental Predictors

We used monthly climatologies from the World Ocean Atlas (WOA18) (Garcia et al., 2019) on a 1° grid to create annual climatologies for temperature, salinity, density, oxygen, apparent oxygen utilization (AOU), silicate, phosphate, and nitrate in each depth layer. Seasonal climatologies were also generated for the surface layer. Since WOA nutrients were unavailable below 800 m, we used deep nutrient data from GLODAPv2 (Key et al., 2015; Lauvset et al., 2016) for the bathypelagic layer. Additional predictors such as thermocline, pycnocline, mixed

layer depth (MLD), and nutricline depths were derived from these variables, using the depth of maximum variance in the relevant variable along a sliding window. We also generated annual and seasonal climatologies from satellite data for net primary production, euphotic depth, particle backscattering ($b_{bp}$), microphytoplankton fraction, log of surface chlorophyll a concentration ($\log([chl_a])$), particulate inorganic carbon, slope of phytoplankton size spectra and irradiance (Cael et al., 2023). Since processes at a given depth are influenced by shallower processes (e.g., sinking particles), for each layer, we used predictors from that layer and from the layers above it and satellite data. A full list of predictors is available in Supporting Information S1 (Table S2).

### 2.2. DOC Modeling

#### 2.2.1. Model Definition

We used Boosted Regression Trees (BRTs) to regress layer-wise DOC concentrations against all the biogeochemical predictors described above. BRTs combine regression trees, which relate a response variable to predictors through recursive binary splitting (Breiman et al., 1984), with boosting, which improves prediction performance by combining many small models (Hastie et al., 2009; Schapire, 2003). Thus, BRTs are an ensemble method in which many "base learners" (i.e., small trees) are combined, each working on the residuals of the previous one. Tree-based methods offer advantages like flexibility with predictors (type, outliers, missing values, relevance), fitting complex non-linear relationships, and the ability to handle interactions between predictors (Hastie et al., 2009). Boosting overcomes the limited prediction performance of single trees, making BRTs a very powerful and popular ML algorithm, with higher performance than many other modeling methods (Elith et al., 2008). BRTs can also use bootstrapping of observations and predictors to reduce overfitting (Friedman, 2002; Hastie et al., 2009). They can adapt to different response variable types and distributions with an appropriate loss function (Elith et al., 2008). In our study, a log-transformation of DOC values ensured normally distributed residuals, making root mean square error suitable. Finally, BRTs are accessible through various programming languages (e.g., R, Python), ensuring usability for non-ML specialists.

#### 2.2.2. Model Training and Assessment

During the training phase, the model generates a succession of small trees to relate log-transformed DOC concentrations to environmental predictors. To properly estimate the generalization ability of the model, a subset of the data is held out during training and used as an independent test set to assess performance. However, the choice of observations included in the test set can influence the performance estimate. A common solution is to use cross-validation (CV), where the data is split into $k$ folds. In each of the $k$ iterations, one fold serves as the test set while the remaining folds are used for training. This process results in $k$ performance estimates, providing a more robust and reliable evaluation of the generalization performance of the model.

Furthermore, BRTs offer the possibility of hyperparameter tuning, such as adjusting the number and size of trees. Optimizing these hyperparameters is essential to prevent overfitting and ensure good generalization performance (Elith et al., 2006, 2008). Hyperparameter optimization typically involves training multiple models with different hyperparameters on the training data and evaluating their performance on a separate validation set. To avoid underestimating errors, it is essential for the validation set to be distinct from the test set. However, as with performance evaluation, the selection of observations included in the validation set can influence the selected hyperparameters. Once again, a common solution is to use CV for a robust hyperparameter optimization.

Thus, to address both generalization error estimation and hyperparameter tuning, we used a nested CV (Varma & Simon, 2006). As the name suggests, this procedure consists of two nested CV. In the first (or outer CV), one fold serves as the test set to estimate generalization error, while the remaining folds are used for model training. Within these training folds, a second (or inner) CV optimizes hyperparameters by iteratively using one fold as the validation set and the others as the training data. In our study, hyperparameter tuning was performed using a space-filling parameter grid of size 30, exploring the following parameters and their respective sampling ranges (note that the extrema may not have been sampled): number of trees [1, 2000], maximum depth of trees [1, 15], minimum number of objects in a node to split further [2, 40], and learning rate $[10^{-10}, 10^{-1}]$. Error stabilization around a minimum indicated sufficient grid size to explore the hyperparameters space. Both inner and outer CV used 10 folds, with stratification based on deciles of the response variable (log(DOC)) to ensure similar distribution across resamples. This approach resulted in 10 distinct models, each with potentially different hyperparameters selected through optimization, rather than a single finalized model.

To avoid overestimating model performance due to spatial correlation between outer and inner folds (i.e., test and learning sets), we performed another nested CV based on spatial location, for the surface layer only, as spatial variation in environmental variables decreases with depth (Costello et al., 2018). We used spatial block CV (Roberts et al., 2017), dividing the space into a $10 \times 10$ grid and randomly assigning cells to 10 folds. This procedure was used for both the outer and inner resampling of the nested CV. This method results in greater variation between outer and inner resampling compared to stratified CV, making the prediction task more challenging and likely leading to lower performance estimates. Thus, stratified CV provides an upper estimate, while spatial CV provides a lower estimate of model performance. Model performance was computed as both Root Mean Squared Error (RMSE) and $R^2$ between predicted and true values of the outer resampling (test set) for each iteration ($n = 10$), in a log-transformed space to prevent artificial inflation from very high DOC values, resulting in 10 RMSE and $R^2$ values for each prediction task.

All models were fitted using the tidymodels framework (Kuhn & Wickham, 2020) version 1.1.1 and the LightGBM engine (Ke et al., 2017) in R (R Core Team, 2023) version 4.3.2.

#### 2.2.3. Model Interpretation

To identify important predictors for DOC prediction among all available variables and thus understand which parameters were susceptible to drive DOC concentrations, we conducted a feature importance procedure by computing model performance after shuffling the predictor values one at a time over 10 permutations (Breiman, 2001). A large drop in performance (usually computed as a loss) suggests an important predictor. Predictor importance was averaged across CV folds. To further understand how the response variable changes with a predictor while keeping others constant, we computed univariate partial dependence profiles (Friedman, 2001). Specifically, we generated 100 ceteris-paribus (CP) profiles for the most important predictors. To estimate the shape of the response to a predictor and the consistency of that shape, we then calculated the mean and standard deviation of the centered CP profiles. These values were then averaged across CV folds.

#### 2.2.4. DOC Projection

To generate global maps of DOC concentration, we applied the fitted regression to all pixels where at least 90% of predictors were available (covering 84%–92% of pixels depending on the layer). For annual data, the predictors distribution was similar between DOC-annotated data (i.e., points for which DOC is known) and new data (i.e., pixels for which DOC was to be predicted) (Figure S3 in Supporting Information S1). This similarity in distributions is crucial to avoid data set shift and prevent extrapolation beyond the training data range, thereby ensuring robust and reliable predictions (Quiñonero-Candela et al., 2022). For seasonal predictions, especially in summer, there was more variation between DOC-annotated data and new data (Figure S4 in Supporting Information S1), so these predictions should be treated cautiously. Using 10-fold nested CV resulted in 10 predictions per task. For each pixel, the final DOC prediction and uncertainty were computed as the weighted average and standard deviation of these 10 predictions, weighted by the $R^2$ value of each iteration.

### 2.3. Total DOC Estimate

We estimated the total DOC inventory by integrating our predictions over the global ocean volume. First, we computed the surface area of each 1° pixel using latitude (Equation S1 in Supporting Information S1). This area was multiplied by the layer thickness, derived from NOAA bathymetry data (NOAA National Centers for Environmental Information, 2022), to get pixel volume. This volume was then converted to seawater mass using the average seawater density (1,027.7 kg $m^{-3}$) (Wunsch, 2015). The total DOC content was obtained by summing the DOC of each pixel. We performed this calculation for each CV fold to estimate uncertainty. Since some pixels (8%–16% depending on the layer) could not be predicted due to missing predictors, this provided a minimum DOC content estimate. We also provide a second estimate where missing pixels were assigned the layer-wise average DOC prediction.

### 2.4. Assessing DOC ML Predictions Using a Biogeochemical Model

To assess the quality of our ML predictions in unobserved ocean regions, we conducted a similar experiment on the DOC output of a biogeochemical model (Nowicki et al., 2022). In this model, DOC fields were available only for the epipelagic, mesopelagic, and bathypelagic layers. A comparison of modeled and observed DOC revealed

that bathypelagic data had the highest agreement ($R^2$ = 51.6%, RMSE = 3.83 μmol $kg^{-1}$). We thus focused our experiment on this layer. The bathypelagic DOC field was subsampled using the locations of DOC observations ($n$ = 1,148). Using BRT, we regressed these values against the same set of predictors used to predict DOC observations in the bathypelagic layer (Table S2 in Supporting Information S1), following the procedure described earlier. Inferred relationships between modeled DOC and predictors were then applied to reconstruct the bathypelagic DOC field, which was then compared to the bathypelagic DOC field from Nowicki et al. (2022).

## 3. Results

### 3.1. Surface Climatologies

The model performance for the annual surface climatology was satisfactory (Figure S5 in Supporting Information S1): RMSE = 0.095 ± 0.007 μmol $kg^{-1}$ and $R^2$ = 77.5 ± 3.3% for the stratified CV, RMSE = 0.120 ± 0.030 μmol $kg^{-1}$ and $R^2$ = 62.8 ± 16.4% for the spatial CV (as mentioned in the methods, spatial and stratified CV provide respectively a lower and an upper estimate of model performance). Lower performance and more variation was expected in the spatial CV due to the increased difference between the training and test sets. In both cases, RMSE values are in the same order of magnitude as the standard deviation across log-transformed DOC measurements (Figure S6 in Supporting Information S1), which can be considered a proxy for both measurement error and seasonal variations. Regarding the $R^2$ values, they highlight that there is variance in the data that could not be captured by the ML model.

In terms of global projections, mid-range values were predicted at mid-latitudes, with lower values (<50 μmol $kg^{-1}$) in the Southern Ocean and to some extent in the equatorial Pacific and high northern latitudes (Figure 1a). In addition, very high DOC values (>150 μmol $kg^{-1}$) were predicted in the Kara, Leptov and East Siberian Seas, while relatively high values (>100 μmol $kg^{-1}$) were also predicted in coastal waters, including the North Sea, East China Sea and Gulf of Guinea. The prediction uncertainty (computed as the standard deviation across CV folds) was higher in coastal waters (up to 30 μmol $kg^{-1}$ in the Arctic Ocean) but low elsewhere (<3 μmol $kg^{-1}$) (Figure 1b). Note that the difference in the CV method did not lead to strong changes in the global projection (Figure S7 in Supporting Information S1).

For the surface seasonal climatologies, RMSE values ranged from 0.071 ± 0.009 μmol $kg^{-1}$ to 0.126 ± 0.030 μmol $kg^{-1}$, while $R^2$ values ranged from 62.4 ± 13.1% to 83.0 ± 3.5% (Figure S5 in Supporting Information S1). Once again, RMSE values are above but in the same order of magnitude as uncertainties of DOC measurements (Figure S6 in Supporting Information S1). Overall, the seasonal prediction uncertainty (Figure S9 in Supporting Information S1) was higher for the summer prediction, consistent with a sparser DOC sampling, especially in the southern hemisphere (Figure S2 in Supporting Information S1). This is also the season for which the model had to make predictions outside of its training range (Figure S4c in Supporting Information S1), making summer DOC predictions less reliable than others. High seasonal amplitude was predicted in the Sea of Okhotsk, East China Sea and Bering Sea, as well as along the coast of Peru and in the Gulf of St. Lawrence (Figure 1c). Conversely, seasonal amplitude was low in the Southern Ocean and at low latitudes in the Western Pacific. However, it cannot be excluded that this latitudinal pattern is due to data limitations that affect the quality of seasonal predictions.

### 3.2. Deeper Climatologies

Regarding deeper climatologies, the prediction performance was high in both the epipelagic ($R^2$ = 78.4 ± 2.2%) and the bathypelagic ($R^2$ = 76.8 ± 4.4%), and a little lower in the mesopelagic ($R^2$ = 61.1 ± 7.2%) (Figure S5 in Supporting Information S1). In the epipelagic layer, the predicted DOC was higher in the subtropical gyres and the Arctic Ocean (60–70 μmol $kg^{-1}$), while lower values (<50 μmol $kg^{-1}$) were found in the Southern Ocean (Figure 2). The pattern in the mesopelagic layer was quite similar to that in the epipelagic layer, but with lower values. Finally, in the bathypelagic layer, the highest DOC values were found in the North Atlantic with 45–50 μmol $kg^{-1}$. Additionally, the sharp patterns observed in the southwest Pacific and next to the Antarctic Peninsula are within the range of prediction uncertainty (Figure S10 in Supporting Information S1), and thus likely not significant.

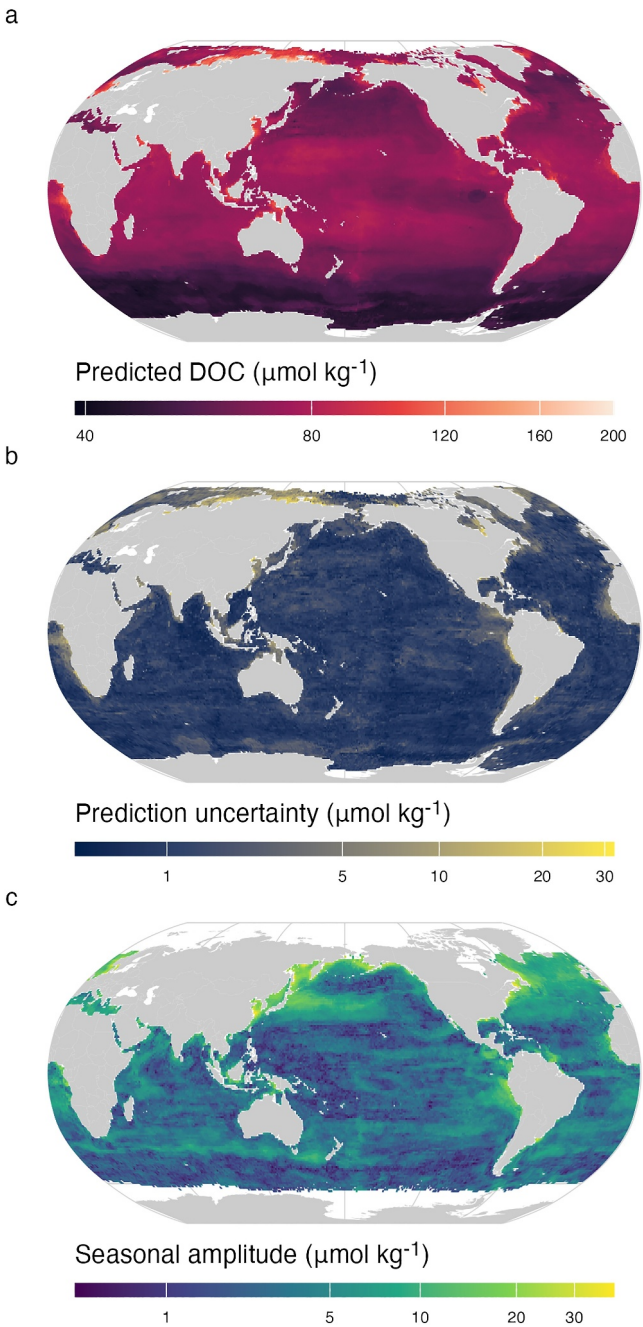


**Figure 1.** The surface (0–10 m) DOC climatology. (a) DOC prediction in the surface layer, computed as the mean across CV folds, weighted by $R^2$ values. (b) Uncertainty of DOC prediction in the surface layer, computed as the standard deviation across CV folds, weighted by $R^2$ values. (c) Seasonal amplitude computed as the difference between maximal seasonal DOC and minimal seasonal DOC, for pixels where seasonal DOC could be predicted for all four seasons. These seasonal projections can be found in Figure S8 in Supporting Information S1, uncertainties can be found in Figure S9 in Supporting Information S1. Note that all color-scales are log-transformed. Areas where no prediction could be made are left blank.

### 3.3. Important Predictors

The most important environment predictors varied across layers (Figure 3). In the surface, epipelagic and mesopelagic layers, macronutrients appeared to drive DOC predictions, with surface nitrate being the best predictor of both surface and epipelagic DOC, while mesopelagic phosphate was the best predictor of mesopelagic DOC. While this pattern is very clear in both the surface and epipelagic layers, the signal in the mesopelagic layer is not as clear, with both surface oxygen and mesopelagic nitrate appearing almost as important as mesopelagic phosphate. In all cases, higher nutrients were associated with lower predicted DOC (Figure S11 in Supporting Information S1), consistent with DOC production resulting from biological activity associated with nutrient consumption (Carlson & Hansell, 2015). Conversely, in the bathypelagic layer, bathypelagic dissolved oxygen was the strongest predictor, with higher dissolved oxygen associated with higher DOC predictions, suggesting that deep ocean DOC concentrations are limited by the remineralization of DOC by heterotrophic bacteria (Carlson & Hansell, 2015).

### 3.4. Total DOC Estimate

By integrating our DOC predictions across the globe, we estimate the total DOC inventory to be a minimum of 679 Pg C. Assigning non-predicted pixels to the average DOC concentration of each layer increases this estimate to 691 Pg C. In both cases, averaging across CV folds resulted in a negligible uncertainty ($2\sigma$ = 0.72 Pg C). In terms of layer-wise content, DOC distribution was around 70% in the bathypelagic layer, 20% in the mesopelagic, 7% in the epipelagic, while the surface layer contained less than 1% (Table 1) in agreement with previous estimates scaled up from mean basin concentrations in the bathypelagic (Hansell et al., 2009).

### 3.5. Verification Against a Biogeochemical Model Output

Our model demonstrated excellent performance in predicting DOC outputs from a biogeochemical model (Nowicki et al., 2022), with $R^2 = 98.9 \pm 0.5\%$ and RMSE = 0.217 ± 0.054 µmol kg$^{-1}$ when evaluated on the outer resamples of each CV iteration. Additionally, the comparison between the ML-reconstructed field and the original field ($n$ = 8,520) showed very strong agreement ($R^2$ = 97.3%, RMSE = 0.290 µmol kg$^{-1}$), with no systematic spatial pattern in prediction error (Figure S12 in Supporting Information S1), apart from a slight underestimation in the Mozambique Channel and a slight overestimation in the Arabian Sea, the Gulf of Mexico, and the Sea of Japan, all of the order of 2 µmol kg$^{-1}$.

## 4. Discussion

### 4.1. Limitations and Potential Improvements

For each DOC prediction task, the prediction error was higher but remained within the same order of magnitude as the DOC measurement error and temporal variation. More specifically, RMSE decreased progressively from the surface to the bathypelagic layer, which is consistent with the lower variability in DOC concentrations in the deeper layers (33 µmol kg$^{-1}$ compared to 220 µmol kg$^{-1}$ in the surface layer). Conversely, $R^2$ decreased from the surface to the bathypelagic layer, except in the mesopelagic layer, where $R^2$ was substantially lower (61.1% compared to 76.8%–78.4% in the other layers). This indicates that the predictors included in our models could not capture all the variance in the response variable (i.e., DOC

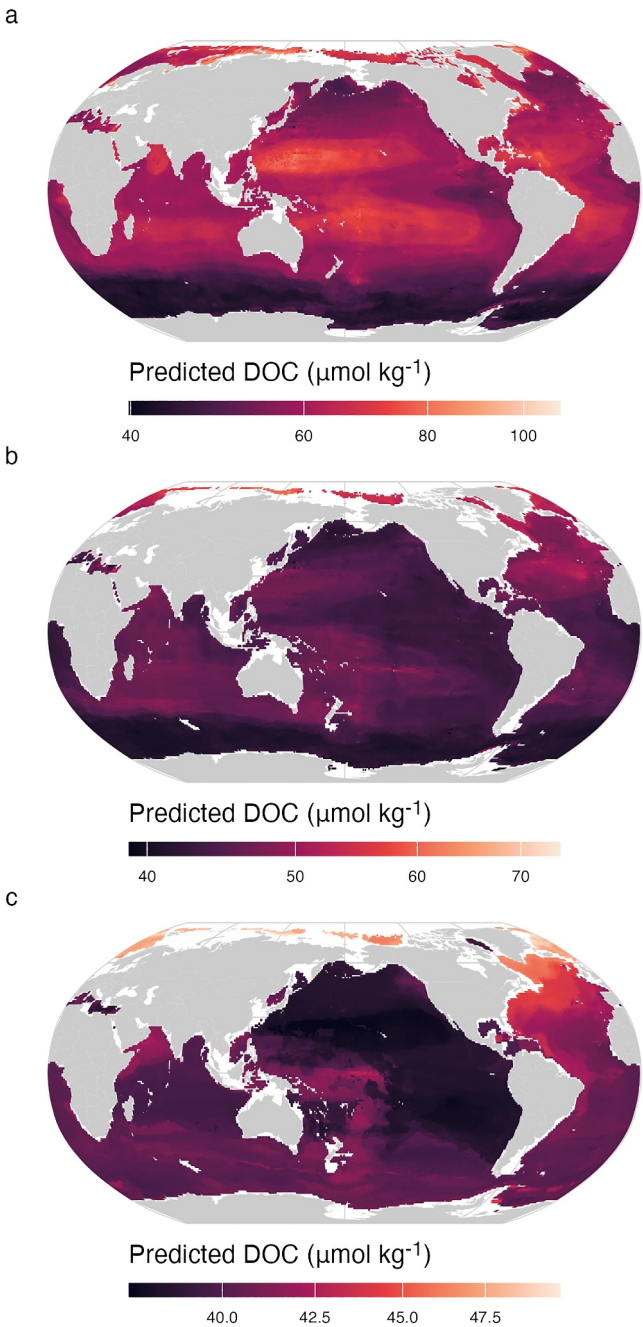


**Figure 2.** Deeper DOC climatologies. DOC prediction in the (a) epipelagic (10–200 m), (b) mesopelagic (200–1,000 m) and (c) bathypelagic (>1,000 m) layers, computed as the mean across CV folds, weighted by $R^2$ values. Note that color scales are log-transformed and vary across plots. Areas where no prediction could be made (missing predictors, shallow waters) are left blank. Prediction uncertainties can be found in Figure S10 in Supporting Information S1.

concentration), either because the set of predictors is not diverse enough, or because some of this variance is attributable to noise. This limitation was particularly true for the mesopelagic layer. One possible explanation is that mesopelagic DOC dynamics may be influenced by processes not captured by the included predictors, such as DOC production by zooplankton and microorganisms through sloppy feeding on sinking particles of organic carbon (Legendre, 2024). The inclusion of predictors encompassing biological processes responsible for DOC production or removal could be a lead for improvement, but these climatologies are still lacking. In terms of seasonal predictions, prediction performance was most of the time lower (lower $R^2$ and higher RMSE) than for annual predictions. This can be explained by the lower number of available observations to train the models (Table S1 in Supporting Information S1). Regarding the spatial and temporal distribution of prediction errors, regions with higher standard deviation across folds in global projections (i.e., higher uncertainty, as shown in Figure 1b, Figures S9 and S10 in Supporting Information S1) are inherently associated with higher prediction errors. This was particularly apparent for the Arctic Ocean, across both layers and seasons.

Furthermore, for a ML model to be able to generalize beyond the training data, it is essential that the new data has a similar distribution to the data seen during training (Quiñonero-Candela et al., 2022). This criterion was met for our annual projections, while there is still room for improvement for seasonal projections. Indeed, the summer surface prediction was particularly limited by the scarcity of observational data. Consequently, we propose that future sampling should focus not only on addressing spatio-temporal gaps (primarily caused by sampling challenges during winter at high latitudes) but also on filling gaps in environmental coverage. More specifically, future DOC estimates could benefit from increased sampling of critical variables for DOC prediction (nutrients in the upper layer, dissolved oxygen in the bathypelagic).

Despite these limitations, the reconstruction experiment of the DOC field from a biogeochemical model demonstrated that our approach was free from evident systematic biases and produced reliable predictions in unobserved ocean regions. Consequently, the DOC climatologies produced in this study significantly advance the understanding of DOC quantification and the spatial patterns of DOC concentrations.

### 4.2. Previous Climatologies and Total DOC Estimates

Our predicted surface DOC concentrations in the open ocean (40–80 μmol $kg^{-1}$) are consistent with findings from previous studies (Bonelli et al., 2022; Fichot et al., 2023; Laine et al., 2024; Yamanaka & Tajika, 1997). Additionally, the higher DOC concentrations predicted for coastal areas compared to open waters align well with the DOC concentration gradient reported by Massicotte et al. (2017), further validating the ability of our model to capture key spatial patterns influenced by terrestrial inputs. Our projections for the epipelagic and mesopelagic layers also show consistency, both in terms of spatial patterns and concentration values, with the maps presented by Roshan and DeVries (2017). In the bathypelagic layer, however, our results differ from those of Yamanaka and Tajika (1997) and Martin and Fitzwater (1992), who, respectively based on a modeling approach and observations, found no significant difference between the North Atlantic and Pacific. In contrast, we estimate higher DOC concentrations in the North Atlantic, with values approximately 10 μmol $kg^{-1}$ higher, in accordance with results reported by Hansell (2013). Finally, our projections of seasonal variability in the surface layer align with those reported by Laine et al. (2024), supporting the robustness of our temporal predictions. Overall, by integrating

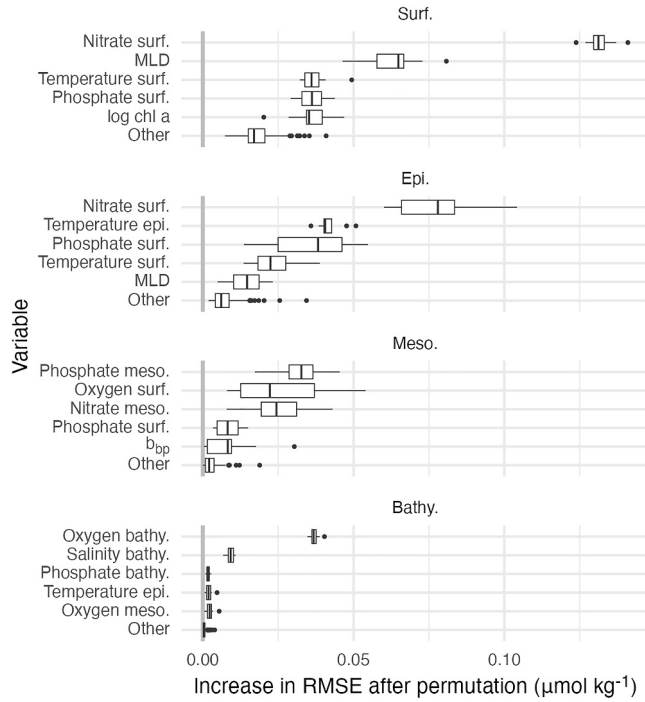


**Figure 3.** Variable importance plots for the five most important predictors for the annual climatology in each layer (importance of all other predictors is averaged and shown as "other"), shown as the increase in root mean square error (RMSE) when a predictor is removed compared to the full model (gray vertical line).

these various aspects, our work not only reinforces previous findings but also significantly advances our understanding of spatio-temporal variability of DOC distribution.

Observation-based inventory estimates (662±32 Pg C) are derived by scaling from mean deep (typically >1,000 m) concentrations at a basin scale with global ocean volume (600 Pg C; 700 Pg C; 685 Pg C: Hansell & Carlson, 1998; Martin & Fitzwater, 1992; Williams & Druffel, 1987). These estimates have been more recently supplemented by the artificial neural network estimate of Roshan and DeVries (2017) and modeling (DeVries & Weber, 2017; Hansell et al., 2012; Nowicki et al., 2022) estimates giving a similar but more constrained inventory of 660 ± 5 Pg C. Our total DOC estimate lies at the higher end of previous estimates Table 1. Using a simple integral model of the preindustrial carbon cycle (Goodwin et al., 2008), the difference in atmospheric $CO_2$ from a complete remineralization of the DOC inventory, given the previous lowest estimate and our higher estimate, is 10 ppm. This suggests that in terms of climate-relevance, these differences in inventory sizes are relatively minimal. The consistency of DOC inventory estimates between simple and complex estimation procedures gives confidence in inventory estimates. Our climatology also agrees closely (<1%) with the proportional distribution

**Table 1**
*Total DOC Content (Pg C) in Each Layer for the a Minima (Non Predicted Pixels Were Left Empty) Filled (Non Predicted Pixels Were Filled With the Average DOC Value for the Layer) Predictions, As Well As Previous Estimates From Hansell et al. (2009)*

| Layer | Prediction a minima (PgC) | Prediction filled (PgC) | Previous estimates (PgC) |
|---|---|---|---|
| Surf. | 2.8 | 3.0 | 47[a] |
| Epi. | 45.5 | 47.4 | |
| Meso. | 140.0 | 145.3 | 138 |
| Bathy. | 490.3 | 495.1 | 477 |
| *Total* | *678.6* | *690.7* | *662* |

[a]This covers both surface and epipelagic layers.

over depth intervals found by Hansell et al. (2009) giving further confidence that our climatology provides robust estimates of DOC concentrations.

### 4.3. Use, Future, and Access

In conclusion, we show that ML is a powerful tool for constructing a global climatology from a limited number of DOC observations. Our climatology is useful not only for empirical quantitative constraint of the present DOC inventory, but also for validation of both prognostic and diagnostic models of DOC. Further observations should allow us to refine this product, especially the seasonal projections, and eventually to predict large DOC shifts in the context of global climate change (Beucler et al., 2024). Generated climatologies are made available online on SEANOE (https://doi.org/10.17882/101170).

## Data Availability Statement

Generated climatology products are made available online on SEANOE via Panaïotis et al. (2024). The code to generate the climatologies is archived on Zenodo via Panaïotis (2025).

**Acknowledgments**
TP and BBC received support through Schmidt Sciences (project: CALIPSO) and by the Horizon Europe (project: 101059915, BIOcean5D). Views and opinions expressed are those of the author (s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

## References

Arrieta, J. M., Mayol, E., Hansman, R. L., Herndl, G. J., Dittmar, T., & Duarte, C. M. (2015). Dilution limits dissolved organic carbon utilization in the deep ocean. *Science*, *348*(6232), 331–333. https://doi.org/10.1126/science.1258955

Aurin, D., Mannino, A., & Lary, D. J. (2018). Remote sensing of cdom, cdom spectral slope, and dissolved organic carbon in the global ocean. *Applied Sciences*, *8*(12), 2687. https://doi.org/10.3390/app8122687

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., et al. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250. https://doi.org/10.1126/sciadv.adj7250

Bonelli, A., Loisel, H., Jorge, D., Mangin, A., d'Andon, O., & Vantrepotte, V. (2022). A new method to estimate the dissolved organic carbon concentration from remote sensing in the global open ocean. *Remote Sensing of Environment*, *281*, 113227. https://doi.org/10.1016/j.rse.2022.113227

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Hall/CRC. https://doi.org/10.1201/9781315139470

Cael, B., Bisson, K., Boss, E., & Erickson, Z. K. (2023). How many independent quantities can be extracted from ocean color? *Limnology and Oceanography Letters*, *8*(4), 603–610. https://doi.org/10.1002/lol2.10319

Carlson, C. A., & Hansell, D. A. (2015). Chapter 3—Dom sources, sinks, reactivity, and budgets. In D. A. Hansell & C. A. Carlson (Eds.), *Biogeochemistry of marine dissolved organic matter* (2nd ed., pp. 65–126). Academic Press. https://doi.org/10.1016/B978-0-12-405940-5.00003-0

Costello, M. J., Basher, Z., Sayre, R., Breyer, S., & Wright, D. J. (2018). Stratifying ocean sampling globally and with depth to account for environmental variability. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-29419-1

DeVries, T., & Weber, T. (2017). The export and fate of organic matter in the ocean: New constraints from combining satellite and oceanographic tracer observations. *Global Biogeochemical Cycles*, *31*(3), 535–555. https://doi.org/10.1002/2016GB005551

Dittmar, T., & Kattner, G. (2003). The biogeochemistry of the river and shelf ecosystem of the Arctic Ocean: A review. *Marine Chemistry*, *83*(3), 103–120. https://doi.org/10.1016/S0304-4203(03)00105-1

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Fichot, C. G., Tzortziou, M., & Mannino, A. (2023). Remote sensing of dissolved organic carbon (DOC) stocks, fluxes and transformations along the land-ocean aquatic continuum: Advances, challenges, and opportunities. *Earth-Science Reviews*, *242*, 104446. https://doi.org/10.1016/j.earscirev.2023.104446

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Garcia, H., Weathers, K., Paver, C., Smolyar, I., Boyer, T., Locarnini, M., et al. (2019). World Ocean Atlas 2018, Volume 3: Dissolved oxygen, apparent oxygen utilization, and dissolved oxygen saturation. In *NOAA Atlas NESDIS* (Vol. 83). Retrieved from https://archimer.ifremer.fr/doc/00651/76337/

Gilchrist, M. D., & Matsumoto, K. (2023). Dynamics of the marine dissolved organic carbon reservoir in glacial climate simulations: The importance of biological production. *Paleoceanography and Paleoclimatology*, *38*(7), e2022PA004522. https://doi.org/10.1029/2022PA004522

Goodwin, P., Follows, M. J., & Williams, R. G. (2008). Analytical relationships between atmospheric carbon dioxide, carbon emissions, and ocean processes. *Global Biogeochemical Cycles*, *22*(3). https://doi.org/10.1029/2008GB003184

Hansell, D. A. (2013). Recalcitrant dissolved organic carbon fractions. *Annual Review of Marine Science*, *5*(1), 421–445. https://doi.org/10.1146/annurev-marine-120710-100757

Hansell, D. A., & Carlson, C. A. (1998). Deep-ocean gradients in the concentration of dissolved organic carbon. *Nature*, *395*(6699), 263–266. https://doi.org/10.1038/26200

Hansell, D. A., & Carlson, C. A. (2013). Localized refractory dissolved organic carbon sinks in the deep ocean. *Global Biogeochemical Cycles*, *27*(3), 705–710. https://doi.org/10.1002/gbc.20067

Hansell, D. A., Carlson, C. A., Amon, R. M., Álvarez-Salgado, X. A., Yamashita, Y., Romera-Castillo, C., & Bif, M. B. (2021). Compilation of dissolved organic matter (DOM) data obtained from global ocean observations from 1994 to 2021. Version 2 (NCEI Accession 0227166). https://doi.org/10.25921/s4f4-ye35

Hansell, D. A., Carlson, C. A., Repeta, D. J., & Schlitzer, R. (2009). Dissolved organic matter in the ocean: A controversy stimulates new insights. *Oceanography*, *22*(4), 202–211. https://doi.org/10.5670/oceanog.2009.109

Hansell, D. A., Carlson, C. A., & Schlitzer, R. (2012). Net removal of major marine dissolved organic carbon fractions in the subsurface ocean. *Global Biogeochemical Cycles*, *26*(1). https://doi.org/10.1029/2011GB004069

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30). Retrieved from https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Key, R. M., Olsen, A., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., et al. (2015). Global ocean data analysis project, Version 2 (GLODAPv2). https://doi.org/10.3334/CDIAC/OTG

Kuhn, M., & Wickham, H. (2020). Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

Laine, M., Kulk, G., Jönsson, B. F., & Sathyendranath, S. (2024). A machine learning model-based satellite data record of dissolved organic carbon concentration in surface waters of the global open ocean. *Frontiers in Marine Science*, *11*. https://doi.org/10.3389/fmars.2024.1305050

Lauvset, S. K., Key, R. M., Olsen, A., van Heuven, S., Velo, A., Lin, X., et al. (2016). A new global interior ocean mapped climatology: The 1° × 1° GLODAP version 2. *Earth System Science Data*, *8*(2), 325–340. https://doi.org/10.5194/essd-8-325-2016

Legendre, L. (2024). Jigsaw puzzle of the interwoven biologically-driven ocean carbon pumps. *Progress in Oceanography*, *229*, 103338. https://doi.org/10.1016/j.pocean.2024.103338

Letscher, R. T., & Moore, J. K. (2015). Preferential remineralization of dissolved organic phosphorus and non-redfield dom dynamics in the global ocean: Impacts on marine productivity, nitrogen fixation, and carbon export. *Global Biogeochemical Cycles*, *29*(3), 325–340. https://doi.org/10.1002/2014GB004904

Martin, J. H., & Fitzwater, S. (1992). Dissolved organic carbon in the Atlantic, southern and Pacific Oceans. *Nature*, *356*(6371), 699–700. https://doi.org/10.1038/356699a0

Massicotte, P., Asmala, E., Stedmon, C., & Markager, S. (2017). Global distribution of dissolved organic matter along the aquatic continuum: Across rivers, lakes and oceans. *The Science of the Total Environment*, *609*, 180–191. https://doi.org/10.1016/j.scitotenv.2017.07.076

Matsumoto, K., Tanioka, T., & Gilchrist, M. (2022). Sensitivity of steady state, deep ocean dissolved organic carbon to surface boundary conditions. *Global Biogeochemical Cycles*, *36*(1), e2021GB007102. https://doi.org/10.1029/2021GB007102

NOAA National Centers for Environmental Information. (2022). ETOPO 2022 15 arc-second global relief model. https://doi.org/10.25921/FD45-GT74

Nowicki, M., DeVries, T., & Siegel, D. A. (2022). Quantifying the carbon export and sequestration pathways of the ocean's biological carbon pump. *Global Biogeochemical Cycles*, *36*(3), e2021GB007083. https://doi.org/10.1029/2021GB007083

Panaïotis, T. (2025). ThelmaPana/DOClimato (Version v2.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.14906514

Panaïotis, T., Wilson, J., & Cael, B. (2024). A machine learning-based dissolved organic carbon climatology. *SEANOE*. https://doi.org/10.17882/101170

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2022). *Dataset shift in machine learning*. Mit Press.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Roshan, S., & DeVries, T. (2017). Efficient dissolved organic carbon production and export in the oligotrophic ocean. *Nature Communications*, *8*(1), 2036. https://doi.org/10.1038/s41467-017-02227-3

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear estimation and classification* (pp. 149–171). Springer. https://doi.org/10.1007/978-0-387-21579-2_9

Sexton, P. F., Norris, R. D., Wilson, P. A., Palike, H., Westerhold, T., Rohl, U., et al. (2011). Eocene global warming events driven by ventilation of oceanic dissolved organic carbon. *Nature*, *471*(7338), 349–352. https://doi.org/10.1038/nature09826

Siegel, D. A., Maritorena, S., Nelson, N. B., Hansell, D. A., & Lorenzi-Kayser, M. (2002). Global distribution and dynamics of colored dissolved and detrital organic materials. *Journal of Geophysical Research*, *107*(C12), 21-1-21–14. https://doi.org/10.1029/2001JC000965

Sulpis, O., Trossman, D. S., Holzer, M., Jeansson, E., Lauvset, S. K., & Middelburg, J. J. (2023). Respiration patterns in the dark ocean. *Global Biogeochemical Cycles*, *37*(8), e2023GB007747. https://doi.org/10.1029/2023GB007747

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91

Wagner, S., Schubotz, F., Kaiser, K., Hallmann, C., Waska, H., Rossel, P. E., et al. (2020). Soothsaying DOM: A current perspective on the future of oceanic dissolved organic carbon. *Frontiers in Marine Science*, *7*. https://doi.org/10.3389/fmars.2020.00341

Williams, P. M., & Druffel, E. R. M. (1987). Radiocarbon in dissolved organic matter in the central north Pacific Ocean. *Nature*, *330*(6145), 246–248. https://doi.org/10.1038/330246a0

Wunsch, C. (2015). *Modern observational physical oceanography: Understanding the global ocean*. Princeton University Press.

Yamanaka, Y., & Tajika, E. (1997). Role of dissolved organic matter in the marine biogeochemical cycle: Studies using an ocean biogeochemical general circulation model. *Global Biogeochemical Cycles*, *11*(4), 599–612. https://doi.org/10.1029/97GB02301