NN-TOC v1: GLOBAL PREDICTION OF TOTAL ORGANIC CARBON IN MARINE SEDIMENTS USING DEEP NEURAL NETWORKS

Naveenkumar Parameswaran[1,2], Everardo González[1], Ewa Burwicz-Galerne[3], Malte Braack[2], and Klaus Wallmann[1]

[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[2]Department of Mathematics, Kiel University, Kiel, Germany
[3]MARUM - Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

Correspondence: Naveenkumar Parameswaran (nparameswaran@geomar.de)

Abstract. Spatial predictions of total organic carbon (TOC) concentrations and stocks are crucial for understanding marine sediments' role as a significant carbon sink in the global carbon cycle. In this study, we present a geospatial prediction of global TOC concentrations and stocks on a 5 x 5 arc minute grid, using a novel neural network approach. We also provide and apply a new compilation of over 21,000 global TOC measurements and a new set of predictors including features such as seafloor lithologies, benthic oxygen fluxes, and chlorophyll-a satellite data. Moreover, we compare different machine learning models based on their performance metrics and predictions and assess their strengths and limitations. For the dataset used, we find that the performance metrics of the models are comparable and the neural network approach outperforms on unseen data compared to methods such as k Nearest Neighbors and random forests, which tend to overfit to the training data. We provide estimates of mean TOC concentrations and stocks in both continental shelves and deep-sea settings across various marine regions and oceans. Our model suggests that the upper 10 cm of oceanic sediments harbor approximately 156 Pg of TOC stock and has a mean TOC concentration of 0.61%. Furthermore, we introduce a standardized methodology for quantifying predictive uncertainty using Monte Carlo dropout. The method was applied to our neural network model and underlying features to generate a map of information gain, that measures the expected increase in model knowledge, achieved through additional sampling at specific locations which is pivotal for sampling strategy planning.

1 Introduction

Burial of particulate organic carbon in marine sediments removes carbon dioxide (CO_2) from the atmosphere and generates molecular oxygen (O_2) that accumulates in the atmosphere (Berner, 1982; Hedges and Keil, 1995). It is a key process in the global carbon cycle that largely controls the atmospheric partial pressures of O_2 and CO_2 on geological timescales (Berner, 1982, 2004). The mechanisms controlling concentrations, standing stocks, degradation and accumulation rates of organic carbon at the seabed are, however, complex and remain a topic of active research (Arndt et al., 2013; Burdige, 2007; Hedges and Keil, 1995; LaRowe et al., 2020b; Bradley and Arndt, 2022). Furthermore, present estimates on the spatial distribution of sed-

imentary carbon concentrations and stocks across the global ocean, including shelf regions, are limited due to sparse data and the large spatial variability observed in shelf deposits (Atwood et al., 2020; Diesing et al., 2021; Lee et al., 2019; Legge et al., 2020; Seiter et al., 2004). An improved map of global organic carbon concentrations and stocks in marine surface sediments, including the continental shelf, could, hence, help to better understand processes governing the turnover and accumulation of organic carbon at the seabed.

Sedimentary organic carbon concentrations are typically reported as total organic carbon (TOC in weight percent), which includes particulate organic carbon bound to sediment grains and a minor contribution by organic carbon dissolved in sediment porewater (Hedges and Keil, 1995). TOC varies between different geological environments (Emerson and Hedges, 1988). Fine-grained shelf and delta sediments deposited close to river mouths typically contain 0.5 – 1.0% TOC at 0 – 10 cm sediment depth (Berner, 1982). A major fraction of TOC deposited in these environments (up to 67%) is not formed by marine plankton but produced by land plants (Burdige, 2005). Shelf regions where neritic carbonates are formed by corals and other organisms at the seabed contain about 1% TOC (Berner, 1982). However, large parts of the continent shelf (about 50 - 70%) do not receive sediment inputs and are covered by relict sands (Emery, 1968; Hall, 2002) that contain only minor amounts of TOC (about 0.1%). Typical deep-sea sediments, that are not associated with high productivity regions, contain about 0.2 – 0.4% TOC (Baturin, 2007; Berner, 1982; Lee et al., 2019; Seiter et al., 2004). In oceanic upwelling regions with high productivity, large amounts of TOC are rapidly deposited at the seabed such that sedimentary TOC concentrations are usually larger than 1% and may reach up to 10% (Berner, 1982; Lee et al., 2019; Seiter et al., 2004). Elevated TOC values are also reported for surface sediments deposited in the Arctic Ocean (1.0%) and the deep basins of the Black Sea (2.0%) (Berner, 1982; Lee et al., 2019; Seiter et al., 2004). Considering these observations, the global mean TOC concentration in both shelf and deep-sea sediments seems to be close to 0.5 to 1.0%.

The inventory or standing stock of TOC in surface sediments (in mass of carbon per seafloor area) is calculated by multiplying TOC concentrations with the dry bulk density of sediments and the thickness of the considered surface layer. Different methods have been applied to derive the standing stock of TOC at regional and global scales. An early estimate based on limited data and expert knowledge concluded that the global TOC stock is 146 Pg TOC for a 30 cm thick surface layer (Emerson and Hedges, 1988). The first estimate of the global TOC inventory derived by a machine-learning approach (k-Nearest Neighbors (kNNs)) using an extended database (5,623 data points) yielded a global inventory of 87 $\pm$ 43 Pg TOC in the top 5 cm layer (Lee et al., 2019). In subsequent publications with an extended database (11,574 sediment cores) and a more advanced machine-learning approach (random forest model), the global inventory was estimated as 2322 Pg TOC for the top 1 m of the sediment column (Atwood et al., 2020). This inventory exceeds the global TOC inventory in terrestrial soils and suggests that TOC in marine surface sediments is the largest TOC pool at the surface of the Earth (Atwood et al., 2020). Another estimate of the global TOC inventory was derived by reactive transport modeling of sedimentary processes employing a range of global datasets (LaRowe et al., 2020a). This model yields a global inventory of 170 Pg TOC for the top 10 cm affected by biological mixing processes.

Since about 70% of the Earth's surface is covered by oceans, and sampling sediments at the seafloor is costly, data coverage will always be sparse. Therefore, advanced methods are required to derive spatial information on sediment properties from a

limited number of point measurements. Machine learning approaches, which have rapidly advanced in recent years, are the most promising approach to tackle this challenge. So far, k-nearest neighbors and random forest models have been applied to derive global maps of sediment porosity (Martin et al., 2015), TOC concentration (Lee et al., 2019), TOC inventory (Atwood et al., 2020), sedimentation rate (Restreppo et al., 2021, 2020), and regional estimates of TOC accumulation rates (Diesing et al., 2021). However, machine-learning techniques have their own challenges and limitations. Overfitting issues are often encountered, and a standardized approach for estimating predictive uncertainty has not yet been established (Lee et al., 2019).

Given these challenges, this paper aims to derive more robust maps of TOC concentrations and inventories for the global ocean. These maps, including the continental shelf, are based on a new larger TOC measurement database and an extended collection of predictors to improve the accuracy of predictions for highly heterogeneous and undersampled geological settings. We compiled an enlarged database of TOC concentrations in surface sediments with 21,125 entries and applied a deep neural network (DNN) as a more advanced machine-learning approach that considers the non-linear relationships between TOC and other geological features. The global ocean was divided into two different domains (shelf and deep-sea), and the network was trained separately for each of these domains. Moreover, we introduced a standardized methodology called Monte Carlo Dropout to quantify predictive uncertainties in the DNN model and derive information gain to guide future sampling efforts.

2 Materials

2.1 Features

An extensive repository of features from both the sea surface and the seafloor at a 5 x 5 arc minute grid resolution has been compiled using previously reported feature lists (Lee et al., 2019; Restreppo et al., 2021; Hart-Davis et al., 2021) that include a range of oceanographic, geological, geographic, biological, and biogeochemical parameters. It is worth noting that oceanographic features are updated very often from newer models and measurements, and some of the features used here might be outdated. Features deemed irrelevant to TOC distributions (e.g. crustal and mantle properties; distance to plate boundary, continental ridges, trenches) were excluded. Additional features that may influence TOC distributions were added to improve TOC predictions. These include total oxygen uptake (respiration rates) at the seabed (Jørgensen et al., 2022), sediment lithology (Garlan et al., 2018), tidal velocities (Hart-Davis et al., 2021), and chlorophyll-a concentrations at the sea surface (NASA, 2014).

99 raw feature grids are compiled for a comprehensive representation of the marine environment, providing the necessary input for the neural network analysis in this study to predict TOC concentrations in marine sediments. Most of these features are easily measurable from the sea surface by e.g. satellite observations, making them a reliable dataset compared to the less accessible properties of the seafloor. Some of the seafloor feature grids used in this work were previously generated from raw data using machine-learning methods (e. g. porosity grid provided by Martin et al. (2015)). Others were reprocessed in this work to achieve global coverage at a resolution of 5 x 5 arc minutes (e.g. sediment lithology (Garlan et al., 2018)).

Neighborhood information was incorporated for a subset of the features. Specifically, 40 of the initial 99 features were spatially averaged using a 50 km radius (Lee et al., 2019). Spatial averaging was applied when TOC concentrations are assumed

to be affected not only by the local feature value but also by feature values in the surrounding area. This approach was used for selected physical (e.g., current velocity), chemical (e.g., dissolved compounds), and biological parameters (e.g., biofauna abundance).

Overall, a total of 139 features including 99 original features and 40 additional spatially averaged features are used in the model. The complete feature list is presented in Appendix S1.

2.2 TOC Data

The dataset for TOC concentrations (in weight percent) utilized in this study has been compiled from multiple sources. It includes global data sets (Seiter et al., 2004; Romankevich et al., 2009; van der Voort et al., 2021; Paradis et al., 2023) and regional data sets for the northern Gulf of Mexico (Beazley, 2003) and the North Sea (personal communication, W. Zhang, HEREON). Each label represents a known measurement (TOC concentration) and is paired with the nearest grid point on the 139-feature grids via L2 distance computation, resulting in the association of a feature vector with each label. For those stations where TOC is reported as function of sediment depth, we calculated the mean TOC concentration for the top 10 cm and used this mean as model label. For many stations, values are only reported for the top 1-2 cm (around 19,000 measurements). We included these stations in our model since they contain valuable information but acknowledge that they may be somewhat higher than those integrated over the top 10 cm since TOC concentrations tend to decrease with sediment depth due to ongoing TOC degradation. However, most sediments deposited on the continental shelf and in high-productivity regions of the open ocean are affected by intense biogenic and physical mixing processes (Boudreau, 1997) such that the down-core TOC decrease is usually small within the mixed surface layer (0 – 10 cm sediment depth) The labeled data is preprocessed to enhance the reliability and robustness of the dataset for subsequent model development and validation. We first searched for duplicates in our combined data base that may arise when the same data are reported in multiple data bases. They were removed from the combined data base when longitudes, latitudes and TOC concentrations were identical. Moreover, coastal regions often exhibit clustered measurements, potentially resulting in shared feature vectors, as all the measurements lie in the same feature grid cell. To mitigate this, a variance assessment is conducted. Labels, that share the same feature vectors, exhibiting high variance (the standard deviation of these labels is higher than 20% of the maximum of these labels) are excluded, while those with low variance are averaged, and the shared feature vector is assigned. Also, some data points situated in close proximity to land were not adequately captured by the 5 x 5 arc minute grid. To address this, reasonable values are assigned by interpolating from the nearest points, ensuring the overall quality of the dataset. Our database includes a total of 110,149 data points that have been consolidated as discussed above such that the final TOC database employed in the model is composed of 21,125 entries (Figure 1). Both the datasets for labels and features can be downloaded at https://doi.org/10.5281/zenodo.11186224.

3 Methods

The primary objective of this study is to build a supervised prediction model that uses feature grid maps as inputs to predict TOC concentrations as outputs. Additionally, we aim to quantify prediction uncertainties using Monte Carlo dropout and

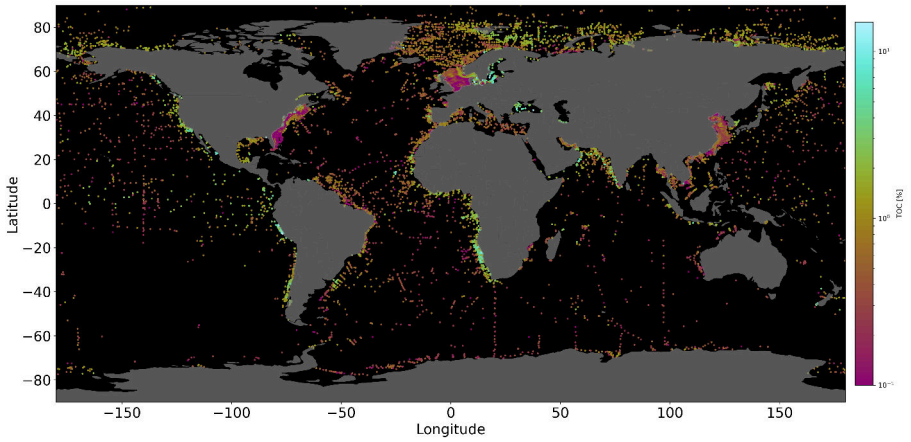

Figure 1. Quantitative TOC measurements (i.e., labels) acquired from various sources (Seiter et al., 2004; Romankevich et al., 2009; van der Voort et al., 2021; Beazley, 2003; Paradis et al., 2023). Notably, data point clusters are observed in close proximity to coastal regions[1].

information theory techniques. The supervised model is trained using the set of labels (TOC data) and their corresponding feature vectors. Due to the non-linearity in the relationships between data and features, we choose deep learning models, which are good at understanding such patterns. Deep Neural Networks (DNNs) transform data non-linearly with non-linear activation functions such as ReLU (Rectified Linear Unit), a piecewise linear function that outputs 0 for negative inputs and the input itself for positive inputs, introducing non-linearity in the DNN. Therefore, even after one layer, multicollinearity in the data is eliminated. In our case of a deep neural network, the final output is controlled by numerous combinations of ReLU functions involving higher order interactions of original features (De Veaux and Ungar, 1994).

3.1 Deep learning model

Deep Neural Networks have achieved state of the art results on a variety of tasks in ocean observation, prediction, and forecasting of ocean phenomena (Song et al., 2023). DNN architectures, that are intrinsically non-parametric and non linear, are less susceptible to the curse of dimensionality. They capture complex relationships between data and features at different levels of abstraction through their hierarchical nature which makes them well-suited to resolve highly complex geoscientific problems (LeCun et al., 2015).

Here, we use a multi-layer perceptron (MLP), feed forward DNN to predict global TOC in sediments and introduce a new approach to map uncertainty in predictions that serves as a quantifiable measure of information gain from sampling. To further improve our predictions, the global ocean was separated into a continental shelf and deep-sea region using the $200m$ water depth horizon as a boundary. Two separate models were trained for these regions (shelf: 0 - 200 m, deep-sea: > $200m$) to consider the different processes that drive sedimentation and control TOC values in the deep-sea and shelf environment. The same set of features is used for both regions, but the interplay of these features differs between the contrasting environments. The weights and the biases in the DNN is initialized using the technique proposed by He et al. (2015). Batch normalization

(which normalizes the inputs of each layer for faster and more stable training) and dropout (which assigns a probability of being deactivated to each node during training and thus prevents overfitting) are applied to each layer for regularization. ReLU is used as the activation function.

The Monte Carlo Dropout method is implemented here to estimate uncertainty in the DNN model, leveraging dropout layers as approximate Bayesian inference (Gal and Ghahramani, 2016). It gives us an ensemble of predictions from different subsets of neurons in the same DNN model. Kullback Leibler (KL) divergence is used to map information gain from the quantified predictive uncertainty. In the field of information theory, KL divergence represents the information gain and is defined as the difference of the cross entropy between the observation and the prediction of an event, and the entropy in the observation of the event (Kullback and Leibler, 1951). In our context, the predicted distribution arises from Monte Carlo dropout prediction ensemble, while the reconstructed observed distribution is modeled with a normal distribution with the predicted value as a mean and the standard deviation of 0.05 TOC%, arising from both technical handling and the precision of the weighing tool (Pape et al., 2020).

Uncertainty and information gain are inherently associated in as far as there cannot be high information gain without high uncertainty, however, information gain also depends on the observation probability distribution and is constrained by it. In other words, information gain measures the expected increase in model knowledge achieved through field sampling at a specific location. This concept provides a strategic guide for determining optimal sampling strategies: taking samples in regions with the highest information gain values is the most efficient way to refine our model's representation of the real world. The mathematical formulation of entropy, cross entropy, and information gain is detailed in Appendix S2.

4 Results and Discussions

Understanding the global distribution of TOC concentrations and stocks is crucial for advancing our knowledge of the carbon cycle and sedimentary environments worldwide. Before delving into the prediction maps from the DNN, we first compare the performance of three methods: DNNs, kNNs, and random forests. Separate models are run for deep-sea and continental shelf regions, and the outcomes are summarized in Table 1. For kNN, 5 neighbors were utilized for continental shelves, and 4 for the deep-sea, based on a sensitivity analysis with respect to model performance. Random forests employed 100 estimators for both marine regions. The DNN consists of 10 layers with 128 nodes each. The choice of hyperparameters in the models are discussed in Appendix S3. This comparison sets the groundwork for a detailed exploration of DNN results. All the methods were run with the same train/test splits of the dataset and the random split is seeded to make the methods reproducible.

The results of this model comparison show that random forest and kNN algorithms exhibit higher correlation coefficients and superior overall performance on the training dataset than the DNN. However, the DNN outperforms the other two algorithms in the test data performance (Table 1), for the dataset used. This discrepancy suggests a potential overfitting issue, where the kNN and random forest models may have become specialized in learning the training data. The emphasis on generalization capabilities is crucial in our context due to data scarcity in many regions, making predictions in unexplored areas a priority. The correlation plot between measured and predicted data shows similar errors for the training and test data sets which confirms that

the DNN-model largely avoids overfitting (Figure 2). The observed underestimation of TOC concentrations at higher values is likely due to the distribution of the ground truth dataset, which is predominantly composed of low TOC concentrations (<1%). Training an NN model on such an imbalanced dataset often results in a model that is biased toward predicting lower values, effectively "erring on the side of caution". Several approaches could be employed to address this issue, such as weighting the gradient descent steps based on concentration values, applying a logarithmic transformation to the TOC scale, or balancing the dataset by withholding low-value labels. However, each of these methods is likely to introduce trade-offs, potentially reducing accuracy in other areas.Ultimately, the most effective way to improve the model's performance in predicting higher TOC concentrations is to obtain additional TOC samples within this higher range.

The prediction map of DNN is presented in figure 3 while maps generated by kNN and random forests are provided in Appendix S3 (Figure S3.1, S3.2). Both the kNN and random forests showed artifacts particularly in the equatorial Pacific and Atlantic oceans, similar to the map published by Lee et al. (2019). As stated by Lee et al. (2019), there is no standard means of quantifying uncertainty in kNN. In random forests, the variance or standard deviation of all the sub-output values to measure the regression uncertainty is considered as an uncertainty quantification method, but it is difficult to provide the uncertainty for an individual base learner (Lucas and Giles, 2016). Estimating the confidence of the predictions should be an important factor in deciding which model to use. On the other hand, uncertainty quantification in DNN is an active field of research and has standardized methods. Nonetheless, kNNs and random forests are useful learning algorithms when computational resources are constrained and requires an out-of-the-box solution.

We also tested a DNN model where the global ocean was not separated into shelf and deep-ocean regions but treated as one entity. The resulting TOC map shows spurious features in the Pacific Ocean (Appendix S5), similar to those that occur in previous predictions. This additional model shows that the separation of the ocean into shelf and deep-sea regions improves the model results.

Method	**Train data**			**Test data (15% of all data)**		
	Pearson CC	**R-squared**	**MSE**	**Pearson CC**	**R-squared**	**MSE**
kNN	0.921	0.45	0.517	0.852	0.719	0.541
Random forests	**0.986**	**0.966**	**0.239**	0.867	0.745	**0.499**
DNN	0.928	0.844	0.492	**0.888**	**0.737**	0.537

Table 1. Comparison of machine learning methods based on performance metrics: Pearson correlation coefficient (Pearson CC), coefficient of determination (R-Squared), and mean squared error for predicted values vs observed labels for the training and testing data. The train:test data ratio is 85:15.

Our DNN-based map of TOC concentrations (Figure 3) shows similarities to maps previously published by Seiter et al. (2004) and Lee et al. (2019), who used geostatistical methods and a kNN model, respectively. All maps show elevated concentrations in the Arctic region and in upwelling areas located along the western continental margins of America and Africa, the equatorial Pacific, and the Arabian Sea. This pattern can be explained by elevated rates of marine primary and export production in upwelling regions delivering large fluxes of TOC to the seabed. The low TOC values in the open oceans are

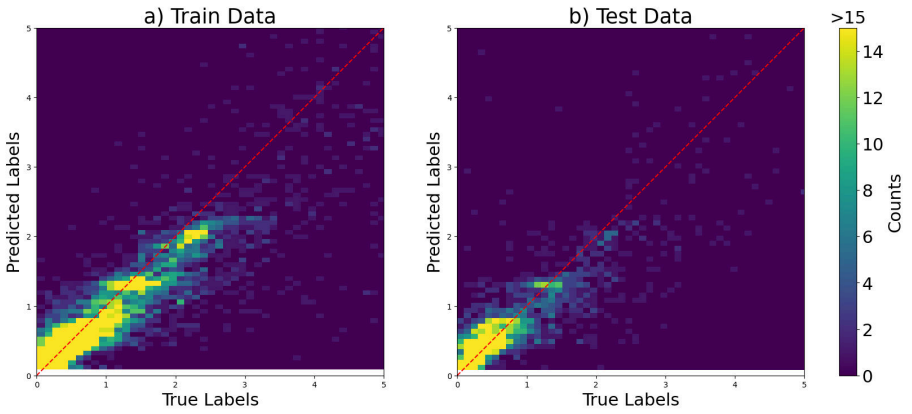

Figure 2. Heat map of the correlation plot between measured (labels) and predicted data (targets) using DNN for a) train data and b) test data, to assess the model performance. The minimal difference observed between train and test errors serves as an indicator of the model's ability to avoid overfitting.

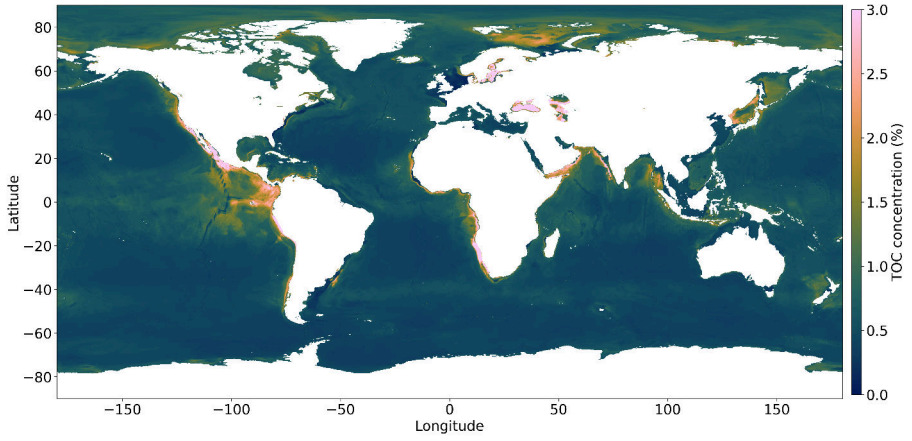

Figure 3. Global prediction map of the TOC concentration using a DNN. Higher TOC concentration is observed in Arctic region and in upwelling areas located along the western continental margins of America and Africa, the equatorial Pacific, and the Arabian Sea.

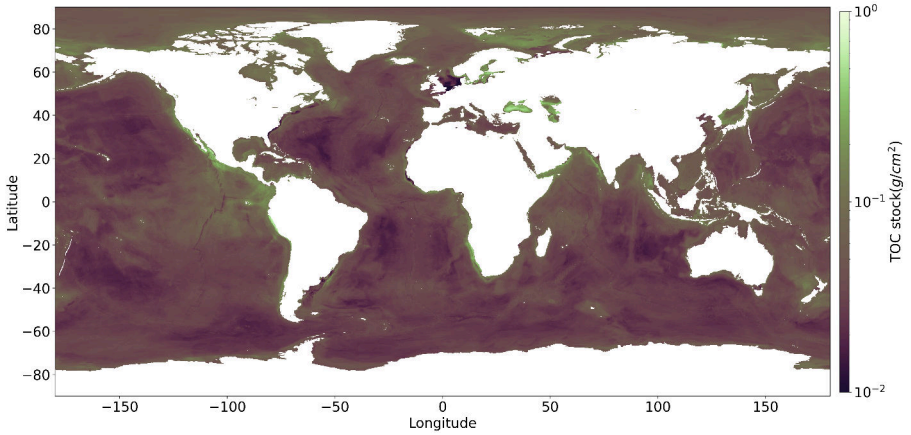

Figure 4. TOC stock map using the global porosity grid provided by Martin et al. (2015). The color map is shown on a logarithmic scale.

related to lower productivity and the large water depths limiting the TOC flux to the deep-sea floor. The predictions in Figure 3 are also consistent with the early work on TOC distributions by Berner (1982) and Emerson and Hedges (1988), showing low TOC values in the open oceans and elevated values for upwelling regions and the Arctic region. The high TOC concentrations predicted for the Black Sea and Baltic Sea (Figure 3) are probably related to the lack of oxygen in bottom waters of these marginal seas that promotes TOC preservation (Hedges and Keil, 1995). The map published by Lee et al. (2019) shows several large areas in the open Pacific that have unusually high TOC concentrations. These patches are probably not realistic since they do not appear in other maps and are not consistent with our understanding of the TOC cycle. They may be artifacts generated by the kNN method and the sparse data coverage in these regions. Our new map avoids these artifacts and presents a pattern that better corresponds to our understanding of TOC accumulation in the seafloor for both the deep-sea and the continental shelf that were never modeled individually in previous maps.

We also produced a map of TOC stocks for the global ocean (Figure 4). The TOC stock was calculated using the global porosity grid provided by Martin et al. (2015) and a density of dry solids (d_s) of $2.6 g/cm^3$. We performed the calculation for the top 10 cm of the sediment column since our TOC data have been measured within this mixed surface layer. Moreover, the top 10 cm are the most vulnerable and dynamic part of the sedimentary TOC pool since they are subject to frequent biological and physical mixing processes (Song et al., 2022) and are affected by human interventions such as bottom trawling (Sala et al., 2021).

$$\text{TOC stock} = (1 - \text{porosity}) \times d_s \times \text{TOC concentration} \times 10 \text{ cm} \quad (1)$$

The TOC stock is computed for global oceans and major seas (Flanders Marine Institute, 2021), considering both continental shelves and deep-sea regions within each ocean and sea (Table 2. Notably, the mean TOC concentration in continental shelves exhibits significant variability across regions. Visualization of the TOC stock in the oceans is provided in Appendix S4.

	Continental shelves			Deep-sea		
Region	**Sum of TOC stock (Pg)**	**Area (million km^2)**	**Mean TOC concentration (%)**	**Sum of TOC stock (Pg)**	**Area (million km^2)**	**Mean TOC concentration (%)**
Arctic Ocean	5.57	5.72	0.94	7.18	9.46	0.88
Indian Ocean	2.63	4.06	0.61	25.86	67.10	0.55
Mediterranean Region	0.61	0.65	0.98	1.95	2.27	1.04
North Atlantic Ocean	2.82	4.26	0.63	14.96	37.46	0.58
North Pacific Ocean	2.74	3.83	0.66	31.16	73.42	0.67
South Atlantic Ocean	1.25	1.86	0.82	12.76	38.67	0.51
South China and Easter Archipelagic Seas	1.56	3.00	0.48	2.58	3.74	0.83
South Pacific Ocean	1.26	1.46	1.02	32.96	83.81	0.58
Southern Ocean	0.21	0.57	0.59	6.39	20.16	0.43
Baltic Sea*	0.77	0.39	3.03			
Caspian Sea*	0.72	0.38	2.27			
Total	20.15	26.20	0.79	135.80	336.08	0.59

Table 2. TOC Stock in the continental shelf and deep-sea regions. *The total sums and the mean concentrations in the continental shelves include the Baltic Sea and the Caspian Sea. Without these regions, the total TOC stock in continental shelves is 18.66 Pg, area of the continental shelves is 25.42 million km^2 and the mean TOC concentration is 0.66%.

According to our model, most the TOC stock can be found in the vast deep-sea basins of the Pacific, Indian and Atlantic oceans which is due to the large area of these basins (Table 2). The shelf region harbors 12.1% of the global stock (Table 2, excluding Baltic Sea and Caspian Sea), similar to the fraction, previously derived by Atwood et al. (2020) who suggested that 11.5% of the global TOC stock is located on the continental shelves. The global TOC stock derived from our model amounts to 155.8 Pg carbon for the 10 cm layer consider in our calculations (Table 2). This value is close to the global stock in the top 10 cm derived by reactive transport modeling (170 Pg, LaRowe et al. (2020a)). The other stock estimates were calculated applying a range of sediment thicknesses. When normalized to 10 cm, the stocks reported by Lee et al. (2019) amounts to 174 Pg while the stock derived by Atwood et al. (2020) results as 232 Pg carbon. The first stock estimate, that was based on expert knowledge and a limited data base, corresponds to only 49 Pg carbon when normalized to 10 cm (Emerson and Hedges, 1988) which is lower than our estimate. Our new global stock assessment, hence, falls into the range of previous estimates.

According to our DNN-model, the mean TOC concentration in continental shelf sediments, excluding the Baltic Sea and the Caspian Sea (0.70%) is close to the concentration in deep-sea sediments (0.59%, Table 2). This is a surprising result since the high marine productivity and low water depths on the shelf induce high TOC fluxes to the seabed that should result in elevated TOC concentrations in surface sediments. Moreover, large amounts of terrestrial particulate organic carbon (POC)

produced by land plants are deposited in shelf sediments (Burdige, 2005) which should further increase TOC concentrations in these deposits. However, TOC concentrations in shelf surface sediments are diminished by a number of factors: i. frequent biological and physical reworking that accelerates TOC degradation processes (Song et al., 2022), ii. dilution of TOC by inorganic material (clay, silt, sand) in delta deposits and other shelf regions with high sedimentation rates (Berner, 1982), iii. strong bottom currents that inhibit sediment deposition such that large shelf areas are covered by relict coarse-grained sediments that were deposited in the geological past and do not contain significant amount of TOC (Emery, 1968), iv: frequent bottom trawling that exposes sedimentary TOC to oxygen and accelerates TOC degradation (Atwood et al., 2020). According to our DNN-model, these factors could potentially decrease TOC concentrations in shelf sediments to such to degree that they attain mean values that are close to those observed in deep-sea sediments (Table 2). It should, however, be noted that most TOC burial occurs on the shelf where sedimentation rates are elevated due to the deposition of riverine particles (Bradley and Arndt, 2022).

A method based on cooperative game theory (SHAP , SHapley Additive exPlanations), is used to further analyze our results and identify features that have a large effect on the predicted distribution of TOC concentrations (Lundberg and Lee, 2017). The higher the SHAP value for a feature, the more important is the feature for the predictions of that particular model According to our model analysis, the total oxygen uptake feature (Jørgensen et al., 2022) has the largest effect (SHAP value) on predicted TOC concentrations in shelf sediments while the global porosity grid (Martin et al., 2015) was the most important feature for deep-sea sediments It should, however, be noted that the feature importance ranking is only valid for our specific model set-up and might not be representative for the real world. Model interpretability and feature importance ranking is further discussed in Appendix S6.

To guide future sampling, a new information gain map is provided (Figure 5). It identifies the regions that should be explored to improve the current model predictions. Some of the main takeaways from the information gain map are: i. Regions with high information gain are found in parts of the equatorial Pacific Ocean, Zealandia and around Papa New Guinea. These regions are less explored geographically and hence the model is not trained with the features in this region. ii. The continental slopes at the western coast of North America, east of Iceland and parts of the eastern coast of Africa have higher information gain, though they have more measurements. This could be due to the steep slopes and rough topography in these regions that may induce a high spatial heterogeneity in TOC values that is not yet resolved by the model. iii. Though the Southern Ocean is not well explored, the higher information gain regions are only found in regions with relatively steep terrain such as areas located close to islands and ocean ridges. These examples show that an abundance of measurements does not necessarily correspond to lower information gain, and vice versa. Information gain depends not only on the geographical proximity of measurements but also on their proximity in the parameter space and the congruence of the measurements made there. Including measurements from a region of higher information gain should lead to higher model knowledge and hence are more valuable compared to regions of low information gain. An experiment showing this is presented in Appendix S2.

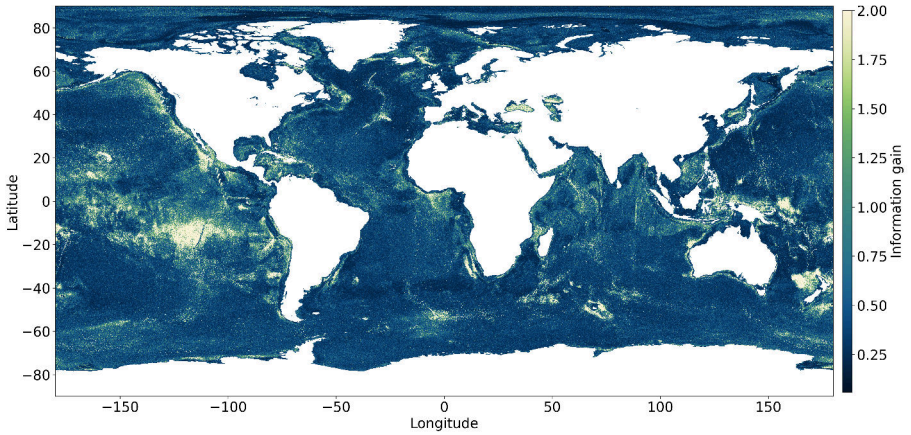

Figure 5. The information gain map serves as a guide for determining optimal sampling locations, i.e. those with high information gain values. The color scheme highlights the high information gain regions with brighter colors. Information gain does not have any units, and is non-negative, $[0, inf)$.

5 Conclusions

The comparison between different modeling approaches, including DNNs, kNNs, and random forests, highlights the effectiveness of each method in predicting TOC concentrations. While kNN and random forest models exhibit higher correlation coefficients and overall performance on the training dataset, the DNN outperforms them on test data performance. This suggests a potential overfitting issue with the kNN and random forest models, where they may have become specialized in learning the training data. Nonetheless, these algorithms remain useful, especially when computational resources are limited.

Our DNN-based map of TOC concentrations shows elevated concentrations in specific regions such as the Arctic and upwelling areas along continental margins. These patterns are consistent with known processes of marine primary and export production. Notably, our model that treats the shelf and deep-sea regions as separate entities captures their individual dynamics with higher accuracy and yields a better global map of TOC concentrations than a model version that simulates the entire ocean as one continuous system. It specifically avoids artifacts like unrealistic high TOC concentrations in open ocean regions with poor data coverage that have also been encountered in previous kNN and forest models.

The computed TOC stock for global oceans and major seas provides valuable insights into the distribution and magnitude of TOC storage. Despite significant variability in mean TOC concentration across continental shelves, our model confirms that the majority of the TOC stock is found in deep-sea basins. Surprisingly, mean TOC concentrations in continental shelves are close to those in deep-sea sediments, suggesting complex processes at play that diminish TOC concentrations in shelf sediments.

The analysis of information gain highlights regions with sparse or contradicting measurements and higher uncertainty, providing guidance for future sampling efforts. It reveals that the abundance of measurements does not necessarily correspond to lower uncertainty, emphasizing the importance of considering both geographical proximity and parameter space proximity in sampling strategies.

In conclusion, our study contributes to a better understanding of global TOC distributions and stocks, shedding light on the complex interplay between biological, physical, and geological processes in marine sedimentary environments. The insights gained from our modeling approach can inform future research and management efforts aimed at preserving and managing marine carbon sinks.

Code availability. The repository of code to run the different models, analyse the outputs is available at: https://doi.org/10.3289/SW_3_2024.

Data availability. Raw features and labels, model outputs are available at: https://doi.org/10.5281/zenodo.11186224.

Appendix S1: Feature list

File names adhere to the naming conventions discussed below. The naming structure is partitioned by underscores and periods in the following order: interface to which the gridded values refer to, quantity of values contained within the grid, units and reference values/units (e.g. meters below sea level), data source, statistic calculated (if applicable), grid pitch, and file extension.

SS – Sea surface – atmosphere interface (may also be average of the entire water column);

SF – Seafloor – water interface (may also be denoted by GL- Ground level);

(r50 km) - Raw feature and feature averaged at a 50km radius used.

Units referenced are as follows:

KGM3 - kilogram per cubic meter;

MS - meters per second;

KM - kilometer;

M_ASL - meters above sea level (i.e. meters referenced to sea level);

MWM2 - milliwatt per square meter;

TGCYR - terragram of carbon per year;

TGYR - terragram per year;

MA - megaannum;

M - meters;

MGCM2 - milligram of carbon per square meter;

DEG - degree;

S - seconds.

Most of the features presented below have been collected by Lee et al. (2020) and Phrampus et al. (2019). The new datasets including the additions from this work are uploaded at https://doi.org/10.5281/zenodo.11186224.

Feature	Explanation	Data Source
GL _COAST _FROM _LAND _IS _1.0 _ETOPO2v2.5m.nc (raw, r50km)	Coastline, with a binary indicator for the presence of coastline. This dataset is derived from ETOPO2v2, a 2-minute gridded global relief data for land topography	National Geophysical Data Center (2006)
GL _COAST _FROM _SEA _IS _1.0 _ETOPO2v2.r50km.men.5m.nc (raw, r50km)	Coastline with a binary indicator for the presence of coastline using ETOPO2v2 relief data for ocean bathymetry	National Geophysical Data Center (2006)
GL _DIST _TO _COAST _KM _ETOPO.r50km.men.5m.grd (raw, r50km)	Distance of ocean grid points to the nearest coast.	National Geophysical Data Center (2006)
GL _ELEVATION _M _ASL _ETOPO2v2.r50km.men.5m.grd (raw, r50km)	Elevation data from ETOPO2v2, representing heights above sea level	National Geophysical Data Center (2006)
GL _RIVERMOUTH _CO2 _TGCYR-1 _ORNL.r50km.men.5m.grd (raw, r50km)	Carbon dioxide flux at river mouths, measured in teragrams of carbon per year (Tg C/yr)	Ludwig et al. (2011)
GL _RIVERMOUTH _DOC _TGCYR-1 _ORNL.r50km.men.5m.grd (raw, r50km)	Dissolved organic carbon flux at river mouths (Tg C/yr)	Ludwig et al. (2011)
GL _RIVERMOUTH _HCO3 _TGCYR-1 _ORNL.r50km.men.5m.grd (raw, r50km)	Bicarbonate HCO_3^- flux at river mouths (Tg C/yr)	Ludwig et al. (2011)
GL _RIVERMOUTH _POC _TGCYR-1 _ORNL.r50km.men.5m.grd (raw, r50km)	Particulate organic carbon flux at river mouths (Tg C/yr)	Ludwig et al. (2011)
GL _RIVERMOUTH _TSS _TGYR-1 _ORNL.r50km.men.5m.grd (raw, r50km)	Total suspended solids flux at river mouths (50 km resolution) (Tg C/yr). All riverine fluxes are binary features with the magnitude of fluxes defined at the coastline and a value of zero for the ocean's interior	Ludwig et al. (2011)
GL _TOT _SED _THICK _M _CRUST1 _NOAA.r50km.men.5m.grd (raw, r50km)	Total sediment thickness in the earth's crust in m	Whittaker et al. (2013)

Feature name	Explanation	Data Source
2N2 _ocean _eot20 _modified.nc;K1 _ocean _eot20 _modified.nc;K2 _load _eot20 _modified.nc;K2 _ocean _eot20 _modified.nc;M2 _load _eot20 _modified.nc;M2 _ocean _eot20 _modified.nc;M4 _load _eot20 _modified.nc;M4 _ocean _eot20 _modified.nc;MF _load _eot20 _modified.nc;MF _ocean _eot20 _modified.nc;MM _load _eot20 _modified.nc;MM _ocean _eot20 _modified.nc;N2 _load _eot20 _modified.nc;N2 _ocean _eot20 _modified.nc;O1 _load _eot20 _modified.nc;O1 _ocean _eot20 _modified.nc;P1 _load _eot20 _modified.nc;P1 _ocean _eot20 _modified.nc;Q1 _load _eot20 _modified.nc;S1 _load _eot20 _modified.nc;S1 _ocean _eot20 _modified.nc;S2 _load _eot20 _modified.nc;S2 _ocean _eot20 _modified.nc;SA _load _eot20 _modified.nc;SA _ocean _eot20 _modified.nc; SSA_load_eot20_modified.nc; SSA_ocean_eot20_modified.nc	Hart-Davis et al. (2021) provides global atlases of both ocean and load tides, containing information about the amplitudes and phases of seventeen tidal constituents (ocean and load) for the global ocean. These constituents include: 2N2, J1, K1, K2, M2, M4, MF, MM, N2, O1, P1, Q1, S1, S2, SA, SSA, and T2, that extends across the entire global ocean ranging from 66°S to 66°N. For higher latitudes, the FES2014b model is used to fill in the gaps. Eleven satellite altimetry missions contribute to this model.	Hart-Davis et al. (2021)
ChlorSummerMean.nc	Average chlorophyll-a concentration during summer (June to November), collected from July 2002 till July 2022	NASA (2014)
ChlorWinterMean.nc	Average chlorophyll-a concentration during winter (Decmeber to May), collected from July 2002 till July 2022	NASA (2014)
DERIVATIVE _GL _ELEVATION _M _ASL _ETOPO2v2.5.nc	Slope from ETOPO2v2.5 data	
GL _HEATFLUX _MWM2 _Becker.5m.nc	Oceanic heat flux data (exchange of heat energy between the ocean surface and the atmosphere) in megawatts per square meter (MW/m^2)	Becker et al. (2014)
GL _LAND _IS _1.0 _ETOPO2v2.5m.nc	Land mask data	National Geophysical Data Center (2006)

Feature name	Explanation	Data Source
POROSITY _global _prediction.grd	Global prediction map for porosity of surface sediments using a random forest method	Martin et al. (2015)
SF _ACTIVE _SEAMOUNTS _KIM.r10km.wct.5m.grd	Active (volcanically) seamounts location data at a 10 km resolution	Kim and Wessel (2011)
SF _AVG _SEA _DENSITY _KGM3 _DECADAL _MEAN _woa13x.5m.grd (raw, r50km)	Sea density in kg/m^3, averaged	Boyer et al. (2013)
SF _COASTLINE _IS _1.0.5m.nc	Coastline data from Global Land One-kilometer Base Elevation (GLOBE)	National Geophysical Data Center (2006)
2 SF _CURRENT _EAST _MS _2012 _12 _HYCOMx.5m.grd;SF _CURRENT _NORTH _MS _2012 _12 _HYCOMx.5m.grd;SF _CURRENT _MAG _MS _2012 _12 _HYCOMx.5m.grd (raw, r50km)	Ocean bottom current data for the east-west, north-south component and total magnitude using the HYCOM model in m/s. The data is provided in 1/12 resolution. The dataset has a time range from August 1, 1995 to December 31, 2012, temporally averaged.	The HYCOM+NCODA Ocean Reanalysis (2014)
SF _GRAINSIZE _D16 _MM _NGDC.5m.nc;SF _GRAINSIZE _D50 _MM _NGDC.5m.nc;SF _GRAINSIZE _D84 _MM _NGDC.5m.nc	Grainsize data with the 16th percentile (D16), median (D50) and the 84th percentile (D84)	National Geophysical Data Center (1976)
SF _SEA _BULKMODULUS _MPA _DECADAL _MEAN _woa13x.5m.nc	Sea bulk modulus in mega pascals (MPa) averaged over six decades, from the year 1955 to 2012. The sea bulk modulus is an important thermodynamic property and is a measure of resistance against the compressibility of a fluid. It is calculated from the International Equation of State of Seawater from Joint Panel on Oceanographic Tables and Standards (1991)	Boyer et al. (2013)
SF _SEA _CONDUCTIVITY _SM _DECADAL _MEAN _woa13v2x.5m.grd (raw, r50km)	Average conductivity of seawater (dissolved ions) at the sea surface over six decades, from the year 1955 to 2012 and the units are in Siemens per meter (S/m)	Boyer et al. (2013)

Feature name	Explanation	Data Source
SF _SEA _OXYGEN _MLL _DECADAL _MEAN _woa13v2x.5m.grd (raw, r50km)	Average dissolved oxygen concentration in seawater in millilitre per litre over a decadal mean	Boyer et al. (2013)
SF _SEA _OXYGEN _PCTSAT _DECADAL _MEAN _woa13v2x.5m.grd (raw, r50km)	Oxygen concentration in seawater percentage saturation averaged over six decades, from the year 1955 to 2012.	Boyer et al. (2013)
SF _SEA _PRESSURE _MPA _DECADAL _MEAN _woa13x.5m.nc	Seawater pressure in mega pascals (MPa) averaged over six decades, from the year 1955 to 2012.	Boyer et al. (2013)
SF _SEA _SALINITY _PSU _DECADAL _MEAN _woa13v2x.5m.nc	Seawater salinity in practical salinity units averaged over six decades, from the year 1955 to 2012.	Boyer et al. (2013)
SF _SEA _SEA _OXYGEN _UTILIZATION _MOLM3 _DECADAL _MEAN _woa13v2x.5m.grd (raw, r50km)	Oxygen concentration in seawater utilization in mol/m³ averaged over six decades, from the year 1955 to 2012.	Boyer et al. (2013)
SF _SEA _TEMPERATURE _C _DECADAL _MEAN _woa13v2x.5m.grd (raw, r50km)	Seawater Temperature in Celcius averaged over six decades, from the year 1955 to 2012.	Boyer et al. (2013)
SL _GEOID _M _ABOVE _WGS84 _NGA _egm2008.5m.grd	Height of the geoid above the WGS84 reference ellipsoid, in meters (m), and referenced to the National Geospatial-Intelligence Agency (NGA)	Pavlis et al. (2008)

Feature name	Explanation	Data Source
SS _BIOMASS _BACTERIA _LOG10 _MGCM2 _WEI2010x.5m.grd (raw, r50km); SS _BIOMASS _FISH _LOG10 _MGCM2 _WEI2010x.5m.grd (raw, r50km); SS _BIOMASS _INVERTEBRATE _LOG10 _MGCM2 _WEI2010x.5m.grd (raw, r50km); SS _BIOMASS _MACROFAUNA _LOG10 _MGCM2 _WEI2010x.5m.grd (raw, r50km); SS _BIOMASS _MEGAFAUNA _LOG10 _MGCM2 _WEI2010x.5m.grd (raw, r50km); SS _BIOMASS _MEIOFAUNA _LOG10 _MGCM2 _WEI2010x.5m.grd (raw, r50km); SS _BIOMASS _TOTAL _LOG10 _MGCM2 _WEI2010x.5m.grd (raw, r50km);	Distribution of mean biomass predictions for (a) bacteria, (b) fishes, (c) invertebrates, (d) macrofauna, (e) megafauna, and (f) meiofauna. The mean biomass was computed using random forest algorithm. The total biomass was combined from predictions of bacteria, meiofauna, macrofauna, and megafauna biomass. Predictions were smoothed by Inverse Distance Weighting interpolation to 0.1 degree resolution and displayed in logarithm scale (base of 10), which is then converted to 5 arc minute grids by Lee et al. (2019)	Wei et al. (2010)
SS _CHLOROPHYLL _LOG _MG _M3 _MODIS _Aqua _MISSION _MEANx.5m.grd (raw, r50km); SS _PIC _LOG _MOL _M3-1 _MODIS _Aqua _MISSION _MEANx.5m.grd (raw, r50km); SS _POC _LOG _MOL _M3-1 _MODIS _Aqua _MISSION _MEANx.5m.grd (raw, r50km)	The Moderate Resolution Imaging Spectroradiometer (MODIS), is a 36-band spectroradiometer measuring visible and infrared radiation and obtaining data that are being used to derive the near-surface concentration of chlorophyll-a (chlor_a) in $mgm^{-}3$. It is calculated using an empirical relationship derived from in situ measurements of chlor_a, concentrations of Particulate Organic Carbon (POC) and Particulate Inorganic Carbon (PIC) (i.e., calcium carbonate or calcite) and blue-to-green band ratios of in situ remote sensing reflectances (Rrs).	NASA (2014)
SS _CORIOLIS.5m.nc	Coriolis data, generated using empirical means	Lee et al. (2020)

Feature name	Explanation	Data Source
SS _DENSITY _KGM-3 _SACD _Aquarius _MISSION _MEANx.5m.grd	The Aquarius/SAC-D satellite mission, launched on 10 June 2011, was a joint venture between NASA and the Argentinean Space Agency (CONAE). The mission featured the sea surface salinity sensor Aquarius and was the first mission with the primary goal of measuring sea surface salinity (SSS) from space. The monthly maps of sea surface density are derived from Aquarius sea surface salinity and ancillary sea surface temperature. The time period used to average is between 25 August 2011 to 07 June 2015.	NASA (2011)
SS _GEOID _ANOMALY _NGA _egm2008.5m.nc (raw, r50km)	The regional Free-air and Bouguer gravity anomaly grids (averaged over 2,5 arc-minute by 2,5 arc-minute) are computed at BGI from the EGM2008 spherical harmonic coefficients	Pavlis et al. (2008)
SS _MIXED _LAYER _DEPTH _MAX _M _Goyetx.5m.grd (raw, r50km); SS _MIXED _LAYER _DEPTH _MIN _M _Goyetx.5m.grd (raw, r50km)	Shows the geographical distribution of the maximum and minimum depth (m) of the mixed layer. The observations are in the time span between March 1995 and February 1996.	Goyet et al. (2000)

Feature name	Explanation	Data Source
SS _PHOTO _AVAIL _RAD _EINSTEIN _M-2 _DAY _SNPP _VIIRS _MISSION _MEANx.5m.grd (raw, r50km); SS _PHYTO _ABSORPTION _443NM _M-1 _SNPP _VIIRS _MISSION _MEANx.5m.grd	Daily average photosynthetically available radiation (PAR) at the ocean surface in $Einstein/m^2/day$ The Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi National Polar-orbiting Partnership (SNPP) have been developed for global ocean color products. PAR is defined as the quantum energy flux from the Sun in the 400-700nm range. For ocean color applications, PAR is a common input used in modeling marine primary productivity. An average of the sensors and the 443 nm wavelength maps are used as features	NASA (2014)
SS _WAVE _DIRECTION _DEG _2012 _12 _WAVEWATCH3x.5m.grd (raw, r50km); SS _WAVE _HEIGHT _M _2012 _12 _WAVEWATCH3x.5m.grd (raw, r50km); SS _WAVE _PERIOD _S _2012 _12 _WAVEWATCH3x.5m.grd (raw, r50km)	Mean Wave direction in , wave height in m and wave period in s. The data is provided in 1/12 resolution. The dataset is averaged over the time range from August 1, 1995 to December 31, 2012. Features are based on the 3rd generation wave model WAVEWATCH III®.	The HYCOM+NCODA Ocean Reanalysis (2014)
SS _WINDSPEED _MS-1 _SACD _Aquarius _MISSION _MEANx.5m.grd (raw, r50km)	Mean wind speed in m/s from the Aquarius/SAC-D satellite mission. The time period used to average is between 25 August 2011 to 07 Jun 2015.	NASA (2011)
TOU _Jorgenson2022.nc	Global map of the total oxygen uptake (TOU) of the seabed.	Jørgensen et al. (2022)

Feature name	**Explanation**	**Data Source**
litho_maps_type1_.nc	Lithology map: Mudflats binary map (Median grain size <0.05 mm)	Garlan et al. (2018)
litho_maps_type2_.nc	Lithology map: Fine sand binary map (Median grain size: 0.05 mm - 0.5 mm)	Garlan et al. (2018)
litho_maps_type3_.nc	Lithology map: Sand binary map (Median grain size: 0.5 mm - 2 mm)	Garlan et al. (2018)
litho_maps_type4_.nc	Lithology map: Clay binary map (Median grain size: <0.01 mm)	Garlan et al. (2018)
litho_maps_type5_.nc	Lithology map: Gravel and stone binary map (Median grain size: >2 mm)	Garlan et al. (2018)
litho_maps_type6_.nc	Lithology map: Bed rock binary map	Garlan et al. (2018)
lithology _grain _size _global _8.nc	Global seabed sediment map with 24 different classes or types of sediments based on a logarithmic progression of median grain size.	Garlan et al. (2018)

Table S1.1: Feature list with description and references, that is used as input to all the models in the paper.

Appendix S2: Information gain

In this paper, KL divergence, also known as information gain or relative entropy, has been used to quantify model uncertainty. As Rényi (1961) points out, in the absence of observational information, the amount of information can be taken numerically equal to the amount of uncertainty concerning the model prediction. The mathematical derivation of KL divergence under the theoretical background of information theory (Shannon, 1948) is presented below. The information entropy of a random variable X, with a probability distribution P is represented as:

$$H(P) = -\sum_i P(x_i) \log P(x_i) \tag{S2.1}$$

Shannon (1948)'s definition of entropy determines the minimum channel capacity required to reliably transmit the information as encoded binary digits. Usually, the true distribution $P(X)$ denotes observed data, measurements, or an exact probability distribution. Here, $P(X)$ is constructed using a normal distribution with a mean value equal to Monte Carlo dropout prediction, and a standard deviation of 0.05 TOC%, which arises from both technical handling and the precision of the weighing tool (Pape et al., 2020). The predicted distribution $Q(X)$ is derived from the Monte Carlo dropout prediction ensemble. The measure $Q(X)$ typically represents a theoretical framework, a model, a description, or an approximation of $P(X)$. The cross entropy between $P(X)$ and $Q(X)$ measures the average number of binary digits to represent an event from $P(X)$, by $Q(X)$. It is represented as:

$$H(P,Q) = -\sum_i P(x_i) \log Q(x_i) \tag{S2.2}$$

The information gain measures the difference between the cross entropy (Equation S2.2) and the entropy (Equation S2.1), is represented as $D_{\mathrm{KL}}(P\|Q)$.

$$D_{\mathrm{KL}}(P\|Q) = H(P,Q) - H(P) = \sum_i P(x_i) \log\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{S2.3}$$

$D_{\mathrm{KL}}(P\|Q)$ is always non negative, remains well-defined for continuous distributions. To obtain the continuous distribution for the predicted distribution $Q(X)$, the prediction ensemble is binned into histograms, to obtain an approximate probability density function (PDF). This PDF is then modeled using curve fitting techniques, typically fitted to a Gaussian distribution (Algorithm 2). $D_{\mathrm{KL}}(P\|Q)$, is calculated globally for each prediction, and plotted in the information gain map.

In supervised machine learning, a model's predictive performance is usually determined by withholding a test dataset during the training phase and comparing the final model outputs to these known values. Such a procedure is not possible when evaluating the performance of information gain: firstly, the concept of a ground-truth for the information gain values does not exist. Secondly, we aim to measure the effect that data point selection guided by information gain has on the model output, and not on the information gain itself. Thus, in order to explore the effect that information gain has on data sampling and model refinement, we devised the following experiment: A DNN model with the same parameters as the original one was trained

while withholding one-third of the original training dataset: $\phi_{wh}(x, W, b)$. Afterwards, this model was used to calculate the information gain for each point in the withheld data. These additional data points were sorted according to their information gain values and divided into two subsets of equal size. Each of these subsets was used along with the initial two-thirds to train two new DNN models: one with added high information gain data points ($\phi_{wh+high\_ig}(x, W, b)$), and one with added low information gain data points ($\phi_{wh+low\_ig}(x, W, b)$). To validate for the entirety of the training data, the process was repeated two more times, withholding a different third of the dataset each time.

In two of the three executions, the (test) performance of $\phi_{wh+high\_ig}(x, W, b)$ was superior to that of $\phi_{wh+low\_ig}(x, W, b)$ (Table S2.1). While the difference in performance from the different data subset might be small in magnitude, the selection of high information gain points also has a positive effect in the structure of the global inference patterns: in figure S2.1 we took the prediction maps for both models of the worst performing data subset, $\phi_{wh+high\_ig}(x, W, b)_2$ and $\phi_{wh+low\_ig}(x, W, b)_2$, and calculated the absolute difference between them and the inference map of the original model $\phi(x, W, b)$ in figure 3. Regardless of the performance metrics, the high information gain model resembles the output of the original model more closely than the low information one.

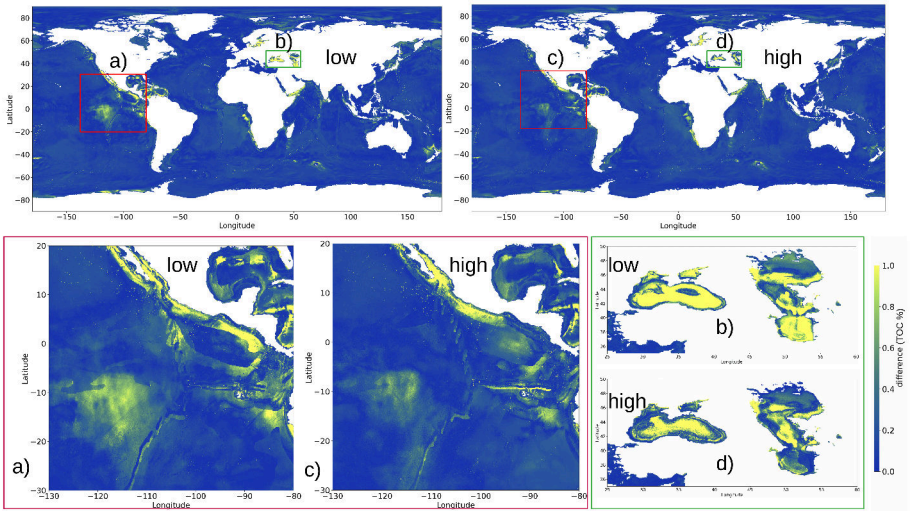

Figure S2.1. Top left: Difference in the prediction of TOC concentration between $\phi(x, W, b)$ and $\phi_{wh+low\_ig}(x, W, b)_2$; Top right: Difference in the prediction of TOC concentration between $\phi(x, W, b)$ and $\phi_{wh+high\_ig}(x, W, b)_2$. Brighter colors (shades of yellow) show higher difference and darker colors (shades of blue) show lesser difference; Bottom left (red box) : Zoomed in version of the equatorial Pacific region in a) and c), Bottom right (green box): Zoomed in version of the Caspian and the black sea in b) and d).

DNN model with different training datasets	Train data			Test data (15% of all data)		
	Pearson CC	**R-squared**	**MSE**	**Pearson CC**	**R-squared**	**MSE**
$\phi_{wh+low\_ig}(x,W,b)_1$	0.918	**0.840**	**0.535**	0.814	0.641	0.692
$\phi_{wh+high\_ig}(x,W,b)_1$	0.918	0.833	0.586	**0.825**	**0.679**	**0.640**
$\phi_{wh+low\_ig}(x,W,b)_2$	**0.927**	**0.856**	**0.405**	**0.887**	**0.784**	**0.525**
$\phi_{wh+high\_ig}(x,W,b)_2$	0.924	0.843	0.470	0.877	0.761	0.593
$\phi_{wh+low\_ig}(x,W,b)_3$	**0.935**	**0.864**	**0.410**	0.819	0.668	0.575
$\phi_{wh+high\_ig}(x,W,b)_3$	0.932	0.853	0.428	**0.855**	**0.728**	**0.475**

Table S2.1. Performance metrics of models trained on different subsets of data based on information gain for different splits of data (seeds): Pearson correlation coefficient (Pearson CC), coefficient of determination (R-Squared), and mean squared error for predicted values vs observed labels for the training and testing data. The train:test data ratio is 85:15.

Appendix S3: Comparison of methods

One of the drawbacks of using DNN is the number of hyperparameters that needs to be tuned. The number of layers and nodes in each layer were decided on a trial and error method starting with the simplest configuration of 3 layers of 8 neurons. The model complexity was increased till the validation and the training performance was comparable, thus avoiding overfitting, but still getting relatively good performance on the test dataset. The initial learning rate was chosen based on the model convergence. The DNN model had 10 layers of 128 nodes each with a learning rate of 0.01. The batch size, decided based on the amount of data, was set as 500, was also chosen based on model convergence. On the other hand, the parameters that were tuned in the random forest algorithm and kNNs were the number of trees in the forest (controlled by number of estimators in sklearn) and number of neighbours respectively. They are tuned using the performance metrics for 1- 50 neighbours for kNN. number of estimators = 10, 20, 30, .. 100, for random forests.

Though it is difficult to tune the DNN model, Table 1 highlights superior performance on the training dataset for kNNs and random forests, while their test performance or generalisation capability lags behind that of DNNs. Figure S3.1 and S3.2, the global predictions from kNNs and random forests respectively, show artifacts particularly in the equatorial Pacific and Atlantic oceans.

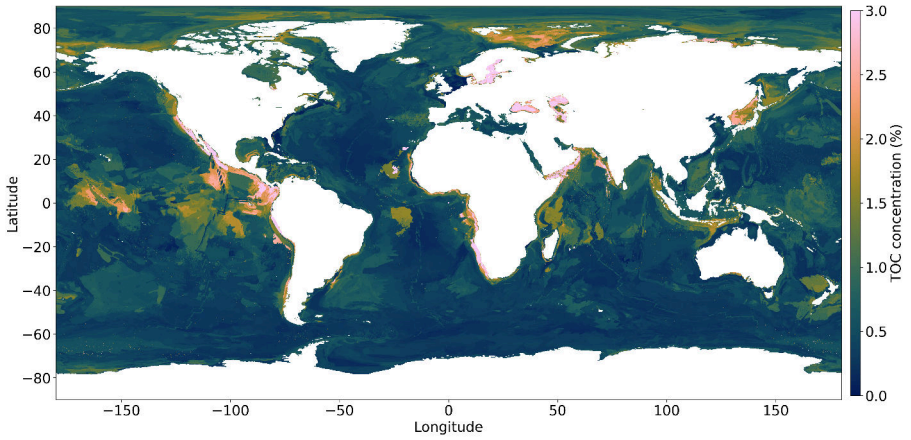

Figure S3.1. Global prediction map of TOC concentrations using a K-Nearest Neigbours algorithm with 5 nearest neighbors in the continental shelves and 4 nearest neighbors in the deep-sea. Spurious patches are observed in the equatorial Pacific ocean and in the Atlantic ocean.

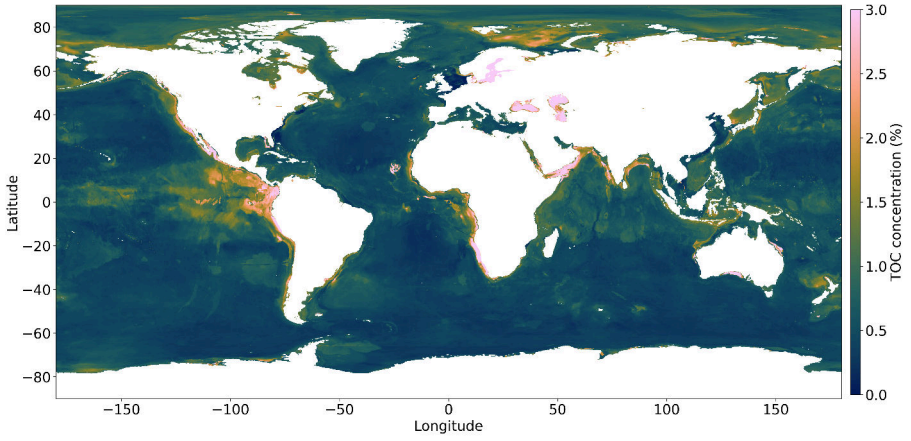

Figure S3.2. Global prediction map of TOC concentrations using a random forest algorithm with 100 estimators. Spurious patches are observed in the Atlantic ocean and the Bay of Bengal

Appendix S4: TOC stock in different marine regions

The table in Table 2 breaks down how much TOC stock is found in different parts of the ocean. Each region is listed, showing how much TOC is there. Here we show a visualization of the different regions in Figure S4.1.

In Figure S4.2, we use a waffle chart to make it easier to see how the TOC is split among these regions. It's like dividing a pie into slices, but here we use squares. With a total of about 156 Pg of TOC worldwide, the South Pacific Ocean gets the biggest share, while the Baltic Sea gets the smallest.

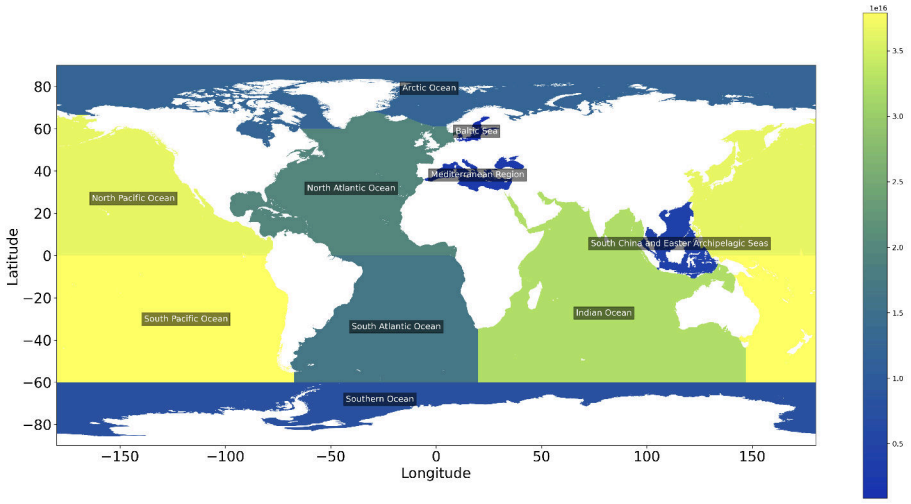

Figure S4.1. TOC stocks in different oceans

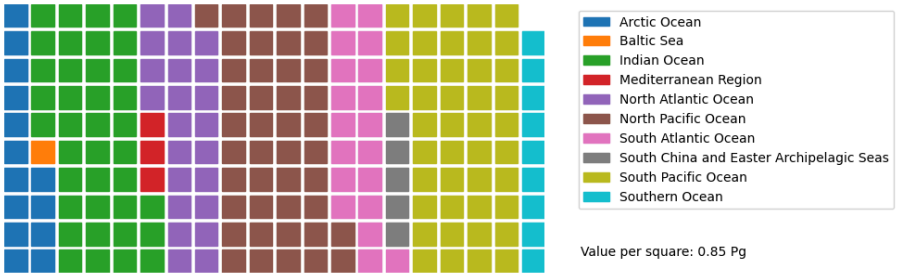

Figure S4.2. TOC stocks in different oceans: Waffle chart

Appendix S5: DNN model run without separation of deep-sea and shelf environments

Here, we test a DNN model where the global ocean was not separated into shelf and deep-ocean regions but treated as one entity. The resulting TOC map shows spurious features in the Pacific Ocean , similar to those that occur in the map published by Lee et al. (2019). These results underscore the importance of separating shelf and deep-ocean regions to achieve more accurate and realistic model outcomes.

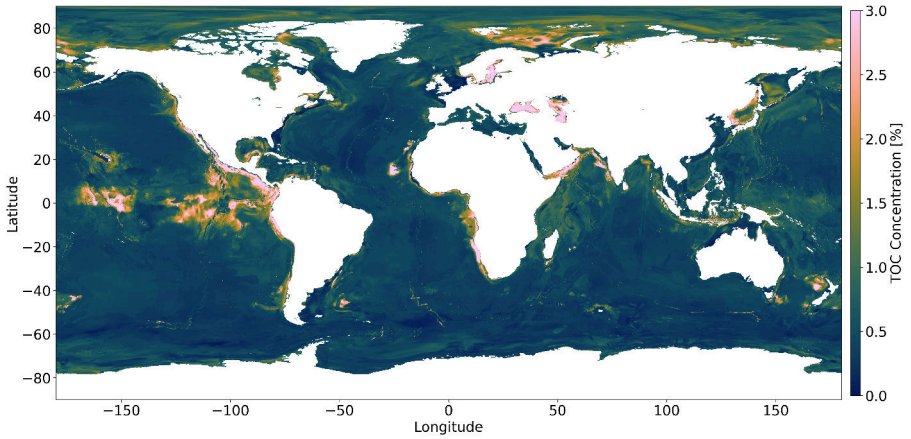

Figure S5.1. TOC concentration map when the DNN model was not separated into shelf and deep-ocean regions. We see unrealistic TOC concentrations especially in the Pacific Ocean.

Appendix S6: Model interpretability using SHAP values

Explaining and understanding why a model makes a certain prediction is as crucial as accuracy and uncertainty in the predictions. This becomes particularly challenging in high-dimensional spaces, where interpreting complex models can be more intricate compared to simpler yet less accurate models. Lundberg and Lee (2017) propose SHAP (SHapley Additive exPlanations) as a unified framework for interpreting predictions. SHAP assigns importance values to each feature for a particular prediction, providing a comprehensive understanding of the model's decision-making process. In our supervised learning model f trained on features $X \in \mathcal{X} \subseteq \mathbb{R}^d$ to predict outcomes $Y \in \mathcal{Y} \subseteq \mathbb{R}$, SHAP, a feature attribution method, considers the model predictions to be decomposed as a sum: $f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi(j, \mathbf{x})$, where ϕ_0 is the baseline expectation (i.e., $\phi_0 = \mathbb{E}[f(\mathbf{x})]$) and $\phi(j, \mathbf{x})$ denotes the Shapley value of feature j at point x.

In our analysis, we aim to simplify the interpretation process by presenting the average importance of features across all predictions, from the deep-sea and the continental shelves. All effects describe the behavior of the model and are not necessarily causal in the real world.

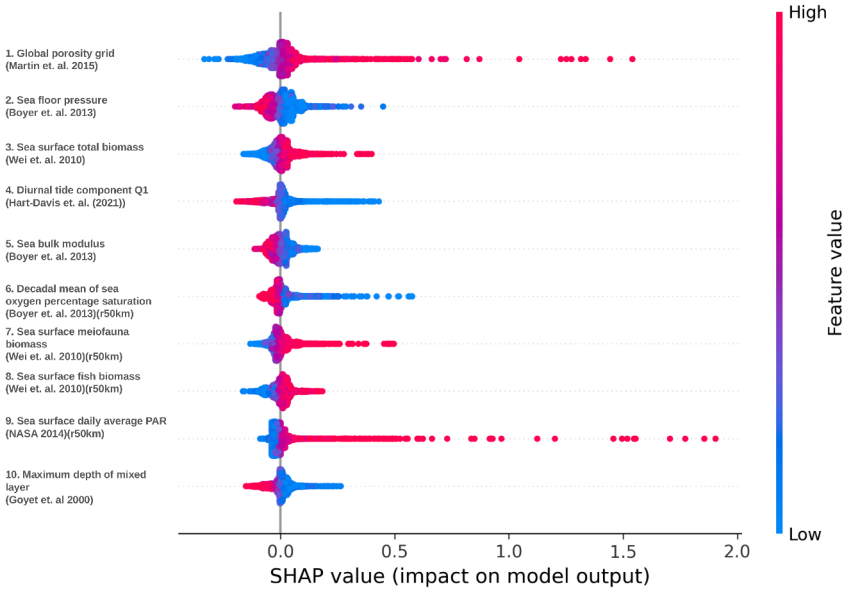

Figure S6.1. Summary plot of Shapley values of the deep-sea DNN model. The global porosity grid (Martin et al., 2015) has the highest feature importance. Regions with high porosity lead to higher TOC concentrations, and vice versa. In the mechanistic model from Bradley and Arndt (2022), porosity is positively correlated to the organic carbon flux through a specific depth. The biological features that includes total biomass, meiofauna and fish biomass in the sea surface (Wei et al., 2010), oxygen concentration in bottom waters (Boyer et al., 2013), daily average PAR (NASA, 2014) show that higher biomass or marine productivity lead to higher TOC concentrations as expected. On the other hand, higher oxygen saturation leads to oxic conditions, resulting in the oxidation of the organic carbon and hence lower TOC concentration. The other features which dominate are the physical oceanographic features, where higher feature values result in lower TOC concentration, such as tidal features (Q1 loading) (Hart-Davis et al., 2021), sea bulk modulus (Boyer et al., 2013) and sea floor pressure (Boyer et al., 2013).

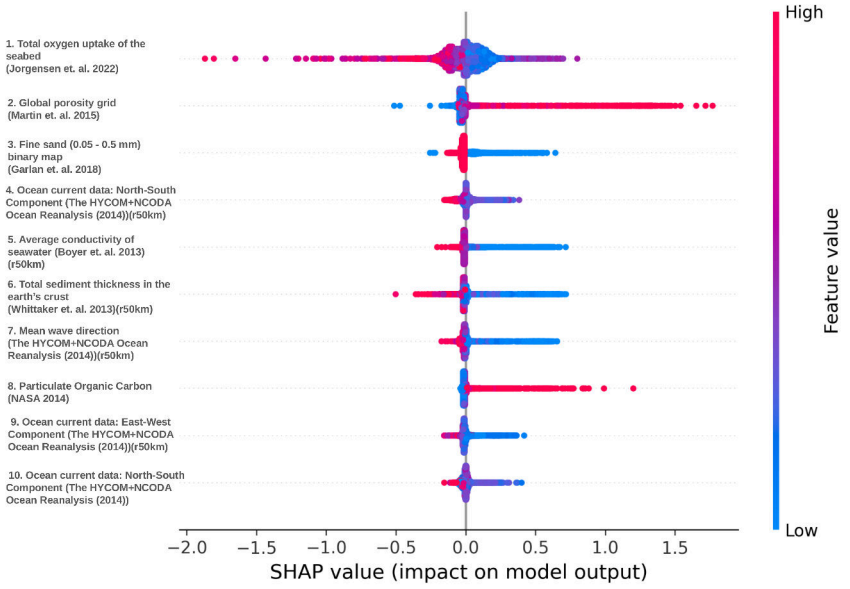

Figure S6.2. Summary plot of Shapley values of the continental shelf DNN model. The total oxygen uptake (Jørgensen et al., 2022) of the seabed has the highest feature importance, with regions of higher oxygen uptake resulting in lower TOC concentrations, denoting oxic conditions. Regions with higher porosity (Martin et al., 2015) result in higher TOC concentrations, while regions with lower porosity result in lower TOC concentration, but with lesser impact. The lithology map is a binary map. Regions with fine sand, with the grain size between 0.05 mm and 0.5 mm (1, being the higher feature value) have low TOC concentration. Higher sediment thickness in the earth's crust lead to lower TOC concentration because of dilution (Berner, 1982). The bottom current components, north-south and east-west, result in reduced TOC concentration, due to higher resuspension of sediments, due to the inhibition of sedimentation, and hence burial of organic carbon. Higher average seawater conductivity results in lower TOC concentration. Higher Particulate Organic Carbon (POC) in the water column has a positive impact on the TOC concentrations, as expected. It can be seen that the feature importance is not as clearly defined as in the deep ocean (as the high (red) and the low (blue) feature values are mixed), because of the complex dynamics on continental shelves.

The summary plot in Figures S6.1 and S6.2 combines the feature importance with feature effects. The summary plot displays Shapley values representing the impact of features on predictions. Each point represents a Shapley value for a feature and an instance. The y-axis position indicates the feature, while the x-axis position corresponds to the Shapley value. Feature values are represented by color, ranging from low (blue) to high (red). To visualize feature importance, points are spread along the y-axis to reveal the distribution of Shapley values per feature. The features are ordered based on their importance, determined by the mean absolute Shapley values across all predictions. The Shapley value is expressed in the same units as the TOC concentration. This indicates the extent to which a specific feature value influences the TOC concentration, whether it drives it towards higher or lower values.

Appendix S7: Algorithms

Algorithm 1. *DNN Training with Batch Normalization and Dropout including Monte Carlo Dropout for inference*

Require: *Labeled dataset* $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$

x_i*: feature vector for the* i*th label*

y_i*: corresponding TOC%*

1: ***Input:*** *feature vector* x_j

2: ***Output:*** *Predicted TOC% predicted% for* x_j

3: ***Method:*** *Construct a neural network with 10 layers and 128 nodes per layer:* $\phi(x, W, b)$

4: Apply batch normalization and dropout to each layer.

5: Initialize optimizer (e.g., Adam) with appropriate learning rate and parameters.

6: Initialize loss function (e.g., Mean Squared Error) for regression.

7: Train the neural network with D *for 1000 epochs:*

8: ***for*** *epoch* $= 1$ *num_epochs* ***do***

9: Randomly shuffle the training dataset..

10: ***for*** (x_i, y_i) *in* D ***do***

11: Forward pass: compute predictions $\hat{y}_i = \phi(x_i, W, b)$.

12: Compute target loss: $loss_{target} = MSE(y_i, \hat{y}_i)$.

13: Back-propagation: update weights and biases using optimizer, with $loss_{target}$ *as the cost function.*

14: ***end for***

15: ***end for***

16: Set dropout to active during inference

17: Perform Monte Carlo dropout for M *forward runs:*

18: $\hat{y}_j^{ensemble} = \phi(x_j, W, b, dropout\_mask_T)$

19: Predicted TOC%, $\hat{y}_j$ *for* $x_j = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_j^{ensemble}$

end

Algorithm 2. *Calculating information gain for the predictions*

Require: *Monte Carlo dropout prediction ensemble,* $\hat{y}^{ensemble}$*, for each grid cell*

1: ***for*** *each grid cell* ***do***

2: Fit a gaussian probability density function $Q_j(x)$ *for* $\hat{y}_j^{ensemble}$ *using histograms and curve fitting algorithm.*

3: Generate original distribution $P_j(x)$ *with mean* $\hat{y}_j$ *and standard deviation 0.05 (sampling error).*

4: Calculate Kullback-Leibler divergence:

$$D_{KL}(P_j \| Q_j) = \sum_i P_j(x_i) \log \left(\frac{P_j(x_i)}{Q_j(x_i)} \right)$$

5: ***end for***

end

Author contributions. Naveenkumar Parameswaran: data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft preparation

Everardo González: conceptualization, data curation, formal analysis, investigation, methodology, software, supervision, visualization, writing – original draft preparation

Ewa Burwicz-Galerne: data curation, investigation, methodology, supervision, validation, writing – original draft preparation

Malte Braack: funding acquisition, project administration, supervision, validation, writing – original draft preparation

Klaus Wallmann: conceptualization, funding acquisition, methodology, project administration, supervision, validation, writing – original draft preparation

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was partially funded by the Cluster of Excellence 'The Ocean Floor – Earth's Uncharted Interface' (EXC 2077) financed by Deutsche Forschungsgemeinschaft (DFG) - Project number 390741603 hosted by the Research Faculty MARUM - Center for Marine Environmental Sciences, University of Bremen, Germany.

The first author wants to thank the Helmholtz School for Marine Data Science (MarDATA), for direct financial support.

The authors would like to thank Dr. Taylor R. Lee for her insightful review of the first submission of this work, which led to further improvement of the manuscript.

AI tools has been used to correct the manuscript, and optimize the code.

References

Arndt, S., Jørgensen, B., LaRowe, D., Middelburg, J., Pancost, R., and Regnier, P.: Quantifying the degradation of organic matter in marine sediments: A review and synthesis, Earth-Science Reviews, 123, 53–86, https://doi.org/https://doi.org/10.1016/j.earscirev.2013.02.008, 2013.

Atwood, T. B., Witt, A., Mayorga, J., Hammill, E., and Sala, E.: Global Patterns in Marine Sediment Carbon Stocks, Frontiers in Marine Science, 7, https://doi.org/10.3389/fmars.2020.00165, 2020.

Baturin, G. N.: Issue of the relationship between primary productivity of organic carbon in ocean and phosphate accumulation (Holocene-Late Jurassic), Lithology and Mineral Resources, 42, 318–348, https://doi.org/10.1134/S0024490207040025, 2007.

Beazley, M. J.: The significance of organic carbon and sediment surface area to the benthic biogeochemistry of the slope and deep water environments of the northern Gulf of Mexico. Master's thesis, Texas A&M University, http://hdl.handle.net/1969.1/534, 2003.

Becker, J. J., Wood, W. T., and Martin, K. M.: Global Crustal Heat Flow Using Random Decision Forest Prediction, in: AGU Fall Meeting Abstracts, vol. 2014, pp. NG31A–3788, 2014.

Berner, R. A.: Burial of organic carbon and pyrite sulfur in the modern ocean: its geochemical and environmental significance, Am. J. Sci., 282, https://doi.org/10.2475/ajs.282.4.451, 1982.

Berner, R. A.: Processes of the Long-Term Carbon Cycle: Organic Matter and Carbonate Burial and Weathering, in: The Phanerozoic Carbon Cycle: CO_2 and O_2, Oxford University Press, https://doi.org/10.1093/oso/9780195173338.003.0005, 2004.

Boudreau, B. P.: Diagenetic models and their implementation, vol. 505, Springer Berlin, 1997.

Boyer, T. P., Antonov, J. I., Baranova, O. K., Coleman, C., Garcia, H. E., and Grodsky, A.: World Ocean Database 2013, NOAA Atlas NESDIS 72, Technical Ed. Silver Spring, MD, https://doi.org/10.7289/V5NZ85MT, last Access: 09/18/2014, 2013.

Bradley, J. A., H. D. L. D. E. and Arndt, S.: Transfer efficiency of organic carbon in marine sediments, Nature Communications, 13, 7297, https://doi.org/10.1038/s41467-022-35112-9, 2022.

Burdige, D. J.: Burial of terrestrial organic matter in marine sediments: A re-assessment, Global Biogeochemical Cycles, 19, https://doi.org/https://doi.org/10.1029/2004GB002368, 2005.

Burdige, D. J.: Preservation of Organic Matter in Marine Sediments - Controls, Mechanisms, and an Imbalance in Sediment Organic Carbon Budgets?, Chemical Reviews, 107, 467–485, https://doi.org/10.1021/cr050347q, pMID: 17249736, 2007.

Crameri, F.: Scientific colour maps, https://doi.org/10.5281/zenodo.8409685, 2023.

De Veaux, R. D. and Ungar, L. H.: Multicollinearity: A tale of two nonparametric regressions, in: Selecting Models from Data, edited by Cheeseman, P. and Oldford, R. W., pp. 393–402, Springer New York, New York, NY, 1994.

Diesing, M., Thorsnes, T., and Bjarnadóttir, L. R.: Organic carbon densities and accumulation rates in surface sediments of the North Sea and Skagerrak, Biogeosciences, 18, 2139–2160, https://doi.org/10.5194/bg-18-2139-2021, 2021.

Emerson, S. and Hedges, J. I.: Processes controlling the organic carbon content of open ocean sediments, Paleoceanography, 3, 621–634, https://doi.org/https://doi.org/10.1029/PA003i005p00621, 1988.

Emery, K. O.: Relict sediments on continental shelves of the world, Am. Assoc. Petr. Geol. B., 52, 445–464, 1968.

Flanders Marine Institute: Global Oceans and Seas, version 1, https://www.marineregions.org/, https://doi.org/10.14284/542, 2021.

Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: Proceedings of The 33rd International Conference on Machine Learning, edited by Balcan, M. F. and Weinberger, K. Q., vol. 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, PMLR, New York, New York, USA, https://proceedings.mlr.press/v48/gal16.html, 2016.

Garlan, T., Gabelotaud, I., Lucas, S., and Marchès, E.: A World Map of Seabed Sediment Based on 50 Years of Knowledge, World Academy of Science, Engineering and Technology. International Journal of Geological and Environmental Engineering, 12, 2018.

Goyet, C., Healy, R., and Ryan, J.: Global Distribution of Total Inorganic Carbon and Total Alkalinity Below the Deepest Winter Mixed Layer Depths, ORNIJCDIAC-127 NDP-076, id: 1970, 2000.

Hall, S. J.: The continental shelf benthic ecosystem: current status, agents for change and future prospects, Environmental Conservation, 29, 350–374, http://www.jstor.org/stable/44520615, 2002.

Hart-Davis, M., Piccioni, G., Dettmering, D., Schwatke, C., Passaro, M., and Seitz, F.: EOT20 - A global Empirical Ocean Tide model from multi-mission satellite altimetry, https://doi.org/10.17882/79489, https://doi.org/10.17882/79489, seanoe, 2021.

He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, 2015.

Hedges, J. I. and Keil, R. G.: Sedimentary organic matter preservation: an assessment and speculative synthesis, Marine Chemistry, 49, 81–115, https://doi.org/https://doi.org/10.1016/0304-4203(95)00008-F, 1995.

Joint Panel on Oceanographic Tables and Standards: Processing of Oceanographic Station Data, UNESCO, Paris, illus., 1991.

Jørgensen, B. B., Wenzhöfer, F., Egger, M., and Glud, R. N.: Sediment oxygen consumption: Role in the global marine carbon cycle, Earth-Science Reviews, 228, 103 987, https://doi.org/https://doi.org/10.1016/j.earscirev.2022.103987, 2022.

Kim, S. S. and Wessel, P.: New global seamount census from the altimetry-derived gravity data, Geophysical Journal International, 186, 615–631, https://doi.org/10.1111/j.1365-246X.2011.05076.x, last access: 09/22/2014, 2011.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, Ann. Math. Stat., 22, 79–86, http://www.jstor.org/stable/2236703, 1951.

LaRowe, D., Arndt, S., Bradley, J., Estes, E., Hoarfrost, A., Lang, S., Lloyd, K., Mahmoudi, N., Orsi, W., Shah Walter, S., Steen, A., and Zhao, R.: The fate of organic carbon in marine sediments - New insights from recent data and analysis, Earth-Science Reviews, 204, 103 146, https://doi.org/https://doi.org/10.1016/j.earscirev.2020.103146, 2020a.

LaRowe, D. E., Arndt, S., Bradley, J. A., Burwicz, E., Dale, A. W., and Amend, J. P.: Organic carbon and microbial activity in marine sediments on a global scale throughout the Quaternary, Geochimica et Cosmochimica Acta, 286, 227–247, https://doi.org/https://doi.org/10.1016/j.gca.2020.07.017, 2020b.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Lee, T. R., Wood, W. T., and Phrampus, B. J.: A Machine Learning (kNN) Approach to Predicting Global Seafloor Total Organic Carbon, Global Biogeochemical Cycles, 33, 37–46, https://doi.org/https://doi.org/10.1029/2018GB005992, 2019.

Lee, T. R., Wood, W. T., Skarke, A., Phrampus, B. J., and Obelcz, J.: Data files associated with the k-nearest neighbor global prediction of isopachs for present to middle Miocene, https://doi.org/10.5281/zenodo.3675364, 2020.

Legge, O., Johnson, M., Hicks, N., Jickells, T., Diesing, M., Aldridge, J., Andrews, J., Artioli, Y., Bakker, D. C. E., Burrows, M. T., Carr, N., Cripps, G., Felgate, S. L., Fernand, L., Greenwood, N., Hartman, S., Kröger, S., Lessin, G., Mahaffey, C., Mayor, D. J., Parker, R., Queirós, A. M., Shutler, J. D., Silva, T., Stahl, H., Tinker, J., Underwood, G. J. C., Van Der Molen, J., Wakelin, S., Weston, K., and Williamson, P.: Carbon on the Northwest European Shelf: Contemporary Budget and Future Influences, Frontiers in Marine Science, 7, https://doi.org/10.3389/fmars.2020.00143, 2020.

Lucas, M. and Giles, H.: Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests, Journal of Machine Learning Research, 17, 1–41, http://jmlr.org/papers/v17/14-168.html, 2016.

Ludwig, W., Amiotte-Suchet, P., and Probst, J. L.: ISLSCP II Global River Fluxes of Carbon and Sediments to the Oceans, https://doi.org/10.3334/ORNLDAAC/1028, last Access: 02/15/2015, 2011.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, p. 4768–4777, Curran Associates Inc., Red Hook, NY, USA, 2017.

Martin, K. M., Wood, W. T., and Becker, J. J.: A global prediction of seafloor sediment porosity using machine learning, Geophysical Research Letters, 42, 10,640–10,646, https://doi.org/https://doi.org/10.1002/2015GL065279, 2015.

NASA: Announcement of Aquarius Level 2 Data Availability, Physical Oceanography Distributed Active Archive Center (PODAAC), https://aquarius.oceansciences.org/cgi/gal_density.htm, 2011.

NASA: MODIS-Aqua Ocean Color Data, Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group, http://dx.doi.org/10.5067/AQUA/MODIS_OC.2014.0, 2014.

National Geophysical Data Center: The NGDC Seafloor Sediment Grain Size Database, first Version. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5G44N6W. Accessed [date], 1976.

National Geophysical Data Center: 2-minute Gridded Global Relief Data (ETOPO2) v2, https://doi.org/10.7289/V5J1012Q, last access: 02/06/2013, 2006.

Pape, T., Bünz, S., Hong, W.-L., Torres, M. E., Riedel, M., Panieri, G., Lepland, A., Hsu, C.-W., Wintersteller, P., Wallmann, K., Schmidt, C., Yao, H., and Bohrmann, G.: Origin and Transformation of Light Hydrocarbons Ascending at an Active Pockmark on Vestnesa Ridge, Arctic Ocean, Journal of Geophysical Research: Solid Earth, 125, e2018JB016 679, https://doi.org/https://doi.org/10.1029/2018JB016679, e2018JB016679 2018JB016679, 2020.

Paradis, S., Nakajima, K., Van der Voort, T. S., Gies, H., Wildberger, A., Blattmann, T. M., Bröder, L., and Eglinton, T. I.: The Modern Ocean Sediment Archive and Inventory of Carbon (MOSAIC): version 2.0, Earth System Science Data, 15, 4105–4125, https://doi.org/10.5194/essd-15-4105-2023, 2023.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K.: The EGM2008 Global Gravitational Model, abstract 2008AGUFM.G22A..01P presented at the 2008 General Assembly of the European Geosciences Union, Vienna, Austria. Last access: 07/10/2014, 2008.

Phrampus, B. J., Lee, T. R., and Wood, W. T.: Predictor Grids for "A Global Probabilistic Prediction of Cold Seeps and Associated Seafloor Fluid Expulsion Anomalies (SEAFLEAs)", https://doi.org/10.5281/zenodo.3459805, 2019.

Rényi, A.: On measures of entropy and information, in: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics, vol. 4, pp. 547–562, University of California Press, 1961.

Restreppo, G. A., Wood, W. T., and Phrampus, B. J.: Oceanic sediment accumulation rates predicted via machine learning algorithm: towards sediment characterization on a global scale, Geo-Marine Letters, 40, 755–763, https://doi.org/10.1007/s00367-020-00669-1, 2020.

Restreppo, G. A., Wood, W. T., and Phrampus, B. J.: A machine-learning derived model of seafloor sediment accumulation, Marine Geology, 440, 106 577, https://doi.org/https://doi.org/10.1016/j.margeo.2021.106577, 2021.

Romankevich, E., Vetrov, A., and Peresypkin, V.: Organic matter of the World Ocean, Russian Geology and Geophysics, 50, 299–307, https://doi.org/10.1016/j.rgg.2009.03.013, 2009.

Sala, E., Mayorga, J., Bradley, D., Cabral, R. B., Atwood, T. B., Auber, A., Cheung, W., Costello, C., Ferretti, F., Friedl, er, A. M., Gaines, S. D., Garilao, C., Goodell, W., Halpern, B. S., Hinson, A., Kaschner, K., Kesner-Reyes, K., Leprieur, F., McGowan, J., Morgan, L. E., Mouillot, D., Palacios-Abrantes, J., Possingham, H. P., Rechberger, K. D., Worm, B., and Lubchenco, J.: Protecting the global ocean for biodiversity, food and climate, Nature, 592, 397–402, https://doi.org/10.1038/s41586-021-03371-z, 2021.

Seiter, K., Hensen, C., Schröter, J., and Zabel, M.: Organic carbon content in surface sediments—defining regional provinces, Deep Sea Research Part I: Oceanographic Research Papers, 51, 2001–2026, https://doi.org/https://doi.org/10.1016/j.dsr.2004.06.014, 2004.

Shannon, C. E.: A mathematical theory of communication, The Bell System Technical Journal, 27, 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x, 1948.

Song, S., Santos, I. R., Yu, H., Wang, F., Burnett, W. C., Bianchi, T. S., Dong, J., Lian, E., Zhao, B., Mayer, L., Yao, Q., Yu, Z., and Xu, B.: A global assessment of the mixed layer in coastal sediments and implications for carbon storage, Nature Communications, 13, 4903, https://doi.org/10.1038/s41467-022-32650-0, 2022.

Song, T., Pang, C., Hou, B., Xu, G., Xue, J., Sun, H., and Meng, F.: A review of artificial intelligence in marine science, Frontiers in Earth Science, 11, https://doi.org/10.3389/feart.2023.1090185, 2023.

The HYCOM+NCODA Ocean Reanalysis: 1/12 deg global HYCOM+NCODA Ocean Reanalysis, url{https://www.hycom.org/data/glbu0pt08/expt-19pt1}, funded by the U.S. Navy and the Modeling and Simulation Coordination Office. Last access: 03/19/2014, 2014.

Thyng, K. M., Greene, C. A., Hetland, R. D., Zimmerle, H. M., and DiMarco, S. F.: True Colors of Oceanography, Oceanography, 29, 10, 2016.

van der Voort, T. S., Blattmann, T. M., Usman, M., Montluçon, D., Loeffler, T., Tavagna, M. L., Gruber, N., and Eglinton, T. I.: MOSAIC (Modern Ocean Sediment Archive and Inventory of Carbon): a (radio)carbon-centric database for seafloor surficial sediments, Earth System Science Data, 13, 2135–2146, https://doi.org/10.5194/essd-13-2135-2021, 2021.

Wei, C.-L., Rowe, G. T., Escobar-Briones, E., Boetius, A., Soltwedel, T., and Caley, M. J.: Global patterns and predictions of seafloor biomass using random forests, PLoS ONE, 5, e15 323, https://doi.org/10.1371/journal.pone.0015323, last access: 06/20/2016, 2010.

Whittaker, J., Goncharov, A., Williams, S., Müller, R. D., and Leitchenkov, G.: Global sediment thickness dataset updated for the Australian-Antarctic Southern Ocean, Geochemistry, Geophysics, Geosystems, https://doi.org/10.1002/ggge.2018, last access: 09/02/2018, 2013.