Dear Dr. Lee,

Thank you for your insightful comments and suggestions which significantly improved our manuscript. We have incorporated changes in the text in reply to the suggestions provided. As suggested, we have also uploaded .csv files, in addition to the previously added .npy files and changed the file namings to make it more simpler.

Sincerely,
Naveen Kumar Parameswaran et. al.

Line	Comment by the reviewer	Reply to the comment
13	Based on only the features defined for any given prediction?	Yes, the expected increase in the model knowledge is based only on the features given to the model or to any given prediction. The abstract has been updated as follows now: (lines 11-14, in Abstract) "*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.*"
66	This model and your model produce the same amount of TOC in upper 10 cm… I realize this may just be by chance but to confirm neither of these are constrained by the other, correct	Neither the models of LaRowe nor ours are constrained by one another. Moreover, both methodologies are different.
76	These are not any better resolved, they are still 5.min predictions over the same area. Perhaps the average cell values are more representative of the true acreage across a spatial area but the resolution is not any better.	Thank you very much for pointing this out at multiple locations in the manuscript. We agree with your comment. We have modified the text to better reflect the message we want to convey, i.e.: the maps produced by the DNN model do a better job at capturing complex relationships and non-linearities in the global TOC distribution. (lines 77-79, in section

		“Introduction”) “*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.* ”
84	I made note in the Appendix about this section, it is worth nothing these features are very old and many of them are outdated or irrelevant (where there is no variance across a feature for observed locations).	At that point, we tried using features available to us. However we took special care in making our methods and code accessible and reproducible to facilitate the future use of the method with newer feature datasets.. We also added this comment in the manuscript: (lines 89-91, Section 2.1) “*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.*”
92	Make sure you get this in there; looks like it may already be right in the Appendix	We oversaw this in the manuscript submission. It is corrected in the updated manuscript.
95	You say 99 here and then 139 at line 100, what are the differences? I know you mention averages over 50 km radius but would that not be for every raw grid	Yes, not all the features are averaged over space. We have assumed that features regarding the sediment characteristics, such as the lithology map, porosity map, and tidal features (base) are rather constant over time and spatially averaging them might not actually help in getting a better prediction about the total organic carbon percentage in a certain location. Features such as physical fields or current velocity need to be spatially averaged due to establishing the average current velocity in a larger region. Similarly, we have assumed that neighborhood information for chemical parameters such as dissolved compounds and biological parameters, such as

		biofauna abundance is necessary to estimate the total organic carbon in a certain location. The text has been changed as follows: "*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 total organic carbon content. Most of the depicted 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. Feature grids that lack global coverage or are only available at inappropriate resolutions have been resampled, cell centered, and interpolated as needed using various techniques, including machine learning. Features regarding the sediment characteristics, such as the lithology map, porosity map, and tidal features (base) are rather constant over time and analysing the neighborhood information by spatially averaging the features might not provide more information about the total organic carbon percentage in a certain location.* *Neighborhood information for features such as physical fields, for ex: current velocity, chemical parameters such as dissolved compounds and biological parameters, such as biofauna abundance is necessary to estimate the total organic carbon in a certain location. We adopted the spatial mean calculation as the averaging method, with a spatial average over a 50 km radius to incorporate neighborhood information alongside raw features. Overall, a total of 139 features are used in the model, including the spatial averages, that are listed in the supplementary information.*"
100	How are you accounting for collinearity between features? Are you doing any feature selection?	The 139 features were chosen in accordance with domain expertise. However, no further selection was undertaken. Line 87 mentions some features that seemed irrelevant to the TOC distributions and were excluded.

		For neural networks, each layer transforms data nonlinearly with activation functions such as ReLU (in our case). Therefore even after one layer, multicollinearity in the data is gone. In our deep neural network, the final output is a function of a lot of combinations of ReLU functions involving higher order interactions of original features. The paper "*Multicollinearity: A tale of two nonparametric regressions*" by De Veaux et al., 1994 states that neural networks generally do not suffer from multicollinearity because they tend to be over parameterized. Neural networks can understand more than the linear relationship of the features. In K Nearest Neighbours, multicollinearity can bring points very close together, hence finding the suitable neighbors could be difficult leading to uncertain results. Random forest model predictions are also robust to issues with collinearity. Additionally, the feature selection process adopted in this study limits the likelihood of collinearity as it would be expected that the addition of a highly correlated predictor variable would not provide significant improvement of prediction accuracy.
110	I cannot see why it is useful to exclude these if they are valid observations. Or are you saying they are outliers? I think potentially there could be that level of variability across one grid cell(~10 x 10km)	These are not outliers, but clearly valid measurements. However the resolution of the input features is too low to handle the resolution of some measurement points in the observation data. When many different or diametrically opposite labels are associated with the same feature grid, this produces contradicting training steps, which in turn decreases training efficiency and increases the aleatory uncertainty in the model.. One solution could be to interpolate the features to the measurement locations for all the measurements. However, this approach is only sensible under the assumption that the features vary linearly (or based on the polynomial approximation used). Unless we have a high resolution of

		features, that could give us a feature value for every measurement, we think that this would only add noise to the model.
115	Why are you including duplicates? Duplicates as in the same measurement just recorded in different databases, or duplicates in different unique observations.	Our merged database initially contained duplicates since some measurements were included in more than one of the underlying data sets that we used for our merged database. We excluded these duplicates and also excluded clusters of points in the same predictor grid cell that had a high variance. Entries with a low variance located within the same grid cell of the predictor mesh were averaged and the averages were included in the database. After these refinements, the total number of entries was reduced from initially 110,149 to the smaller number of 22,192 entries that we used in the model. We realize that the old formulation was prone to cause confusion and have rewritten the sentence as follows:: "*…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 22,192 entries (this excludes duplicates from overlap of different databases and labels of high variance with same feature vectors) …*"
122	Such as?	We meant complex non-linear patterns in the features and their interactions with each other. Most of the features or parameters in earth science are highly non-linear from physical oceanographic features, to geological features. Since the term "complex" could be a very general term for the readers, we changed the term to "non-linear" and hence the sentence is updated as follows: "*...Due to the non-linear patterns in the data, such as oceanographic, biological, and geological features, and in the relationships between each other, we choose deep learning models, which are good at understanding such patterns...*"

131	How is uncertainty different from information gain? IN other prediction frameworks the two are inherently different e.g., In Lee et. al., 2019 you can have high uncertainty and low parametric isolation(similar to information gain). That is, the locations with high uncertainty do not inherently mean the most information gain as these locations have low information gain because they are parametrically similar to the other observed data points. And vice versa.	We derive the uncertainty measure from the variance in the results of single Monte Carlo DropOut inference steps. By fitting a probability distribution *Q* to this variance, we can express the uncertainty using our Information Theory framework: it takes the form of the entropy of Q, H(Q). The information gain, on the other hand, is closely related to the cross-entropy H(P|Q) between the predicted distribution Q and the observed distribution P (our theoretical sampling procedure). It acts as a measure of similarity between the two distributions. In other words, uncertainty as a measure pertains to the model alone, while information gain takes into account the information constraints of a physical observation: a point with high prediction uncertainty will not express any information gain if no further information can be gained from the actual sampling of it.
134	Did you ultimately feed it the same set of predictors though? How were they selected? Did you try to do one model on the entire world? How did the results differ?	We feed the same set of features for both the deep ocean and continental shelves. We think that the model has to be trained on continental shelves and deep ocean separately, because the interaction of the input features are not the same in both regions, because of the different dynamics. Please see line 26 - 39. Yes, we also set up a model where the entire global ocean including shelf and deep-sea was simulated in one model run. We noticed a higher number of artifacts in the deep ocean when this global model was applied. Our observation may confirm that features interact differently in the shelf and deep-sea domain due to the different mechanisms controlling TOC concentrations in these contrasting environments.

		We added a sentence to make this clearer: "*...It is to be noted that the same set of features is used for both the regions. But the interplay of these features could be different in contrasting environments…*"
148	What information do you have to support this?	From expert knowledge, it was considered that when TOC % is estimated and the samples are weighted, it is equally probable to underweight and overweight samples. Pape et. al 2020 provides us the standard deviation of 0.05 % as the standard deviation of the TOC measurement. As a standard approach in science, it is always safe to assume a normal distribution, when we do not have more information or a better mathematical representation of the process.
158 a	Using what metric?	We ran a loop over 1-50 neighbors for both the continental shelf and deep ocean models and estimated these numbers. These numbers have the least combined error for both the train and test dataset.
158 b	How did the results differ if you trained on one model or the two separate models(shelf, and deep)	We saw more artifacts, especially in the Pacific ocean, similar to Lee et al., 2019. The patches did reduce a lot when two separate models were used.
162	How are you generating the test/train splits? Randomly? If so, how often is a rest/train value close in proximity to an observed value? For example if you have 5 grid cells close to one another, and one of them was pulled for testing while the other four neighbors are used for training. Then it may be easy to predict that point given the spatial dependencies inherence to the features that define that point. I also assume you are controlling your train/test splits so each model receives exactly the same splits of data	We are generating the train/test splits using the sci-kit library function: train_test_split. The random split inherently considers that the data points are independent of each other. We could implement specialized splitting, such as cluster-based split. However, we think this is not really necessary since data points here are inherently independent of each other and randomizing the splits to the model creates enough room for the model to have some data points out of distribution during the test. Also, with cross-validation, as in Lee et al., 2019, we would still not know if the data points are chosen the way that the model does not overfit to the training data for each fold. The train/test splits are the same for all the

		methods(random forests, KNNs and DNN). We added this line to make this clear: "*...All the methods were run with the same train/test splits of the dataset and the split is seeded to make the methods reproducible…*"
164	I think some of that information should be moved to the primary manuscript, especially since this seems to be a major point of the paper. The author should be sure to define disadvantages of DNNs. There is a lot of exploration against other sides but there are downsides to all methods depending on what your end goal of a work is. The author should define these.	Thank you for this comment. It is definitely true that all the methods have downsides. DNNs are difficult to use because of their methodological complexity. This results in a hyperparameter space that is much larger than in KNN or RF as well as the implementation effort (i.e. lines of code) which can be one or more orders of magnitude higher. Hyperparameter tuning is very important in the case of neural networks. It also requires higher parametric exploration, and technical knowledge. It is highly data intensive. We chose DNNs for its ability to do well with non-linear datasets, and a strong Bayesian theory with uncertainty quantification. As suggested, parts of the appendix are moved to the main manuscript, as it is one of the main achievements of the paper. We moved parts about the overfitting issue and the artifacts in the prediction maps from the random forests and KNNs to the main manuscript. We still kept the prediction maps from the different methods in the appendix, since it might cause confusion with three different prediction maps, and we would like to give spot light to the prediction map from the DNN.
165	What kind of errors (e.g., 10-fold CV, mean, median, ?) ? Be more explicit in dicsussing this.	We used Mean Square Errors(MSE) for training the model. After randomly splitting the data, a much higher testing error as in the case of random forests and KNNs is a sign of overfitting. This is also updated in Table 1.
Table 1	I would say these are incredibly close... Are they really that different?	Pearson CC is a subpar performance measure and was only included here for

		consistency with prior works. One of its weaknesses is a very slow climb to the theoretical maximal value of 1.0. Regardless of the scale, we see the values of Pearson CC merely as a confirmation that the model performs better and not worse than previous works.
Table 1	Are these CC and R2 generated from 10-fold CV for the training dataset? Or this is just raw final predicted value vs observed values? Are you comparing the labeled (gridded?) datasets or the raw observed values here?	This is the final predicted value vs observed values. We do not use a 10 fold cross validation. We compare the raw observed values here.
Table 1	You should also put the errors here since you reference them in the above section.	Agreed. We changed the text accordingly.
Table 1	Same sets of data for each algorithm? Randomly selected? or over specific regions? In many ways from a geoscience perspective, it would make more sense to withhold a "research cruise dataset" to actually test this not just randomly withhold (related to comment on line 161)	Same sets of data were selected for each algorithm and they were randomly selected. From a data science perspective, the data from the same cruise could be similar to data points at a different location, because of its proximity in a higher dimensional space, than in a geographical space. FOr this we could exper
Figure 2	Might convey more information as a heat map and two subplots in one figure, it is difficult to interpret density of points in a standard plot like this	Agree, two new heatmaps have been generated to substitute the scatter plot. This greatly increases the information that can be read in the first third of the diagonal.
179	This is not true, you are predicting at the same resolution and coverage (globally). Perhaps use a different statement to describe	Addressed in L. 76. The text here now reads: "*Our new map avoids these artifacts and presents a pattern that better corresponds to our understanding of TOC accumulation in the seafloor. This is also true for the shelf regions that were never regarded individually in previous maps.*"
Figure 3	Scale bar makes it difficult to see if there were artifacts. I suspect that there will be some (likely less than the original but more apparent via different scale bar) because some of the same sets of predictors are being used.	Agree that the upper end of the scale is difficult to resolve. The colormap is changed and the upper limit lowered to better display values >3%

197	This value is different than line 66?	The value in line 66 is correct. Also the text here is changed to clarify the reactivity scenario: *"This value is close to the global stock in the top 10 cm derived by reactive transport modeling in the low reactivity scenario (170 Pg, LaRowe et al. (2020a))."*
Figure 5	Indicate what lighter and warmer colors mean, specifically 1.0 is more information gained and 0.0 is less information gained on further sampling? This is based on the features used? The reader should explicitly understand the assumptions (features define this, distribution assumptions, etc) that go into making this map.	Changed the figure text to better convey the meaning of color and values to: "… . *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 these regions with brighter colors."* Also added the following additional clarification to the methods section in line 152: "… . *This concept provides a strategic guide for determining optimal sampling strategies: monitoring regions with the highest information gain values is the most efficient way to refine our model's representation of the real world.*"
216	Is uncertainty and information gain always inherently associated? If they are why should information gain not just be uncertainty. Are there any cases that you see high information gain and low uncertainty. Discuss this, these are inherently different in other analyses; see previous comments.	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 thus, two points with the same uncertainty values can have different amounts of information gain. This is now explained more in detail at the end of the methods section.
219	Why? Is this in any way related to the label scheme you are using (e.g., line 110)	Tangentially related, but the labeling scheme is not the cause of this. If multiple data points with similar sets of features get assigned labels with diametrically different values during training, this will most likely result in a model with high uncertainty and information gain for this feature space

		region.
224	If this is a major statement of the paper then it should be more throughly discussed in the manuscript	Agreed. Moved appendix A to its own section in results.
225	How does the information gain work on the test/training datasets? I.e., if you do a prediction and produce an information gain and some of those observations were involved in the test set how did the predictions change? This would show that your information gain really does work	We made an experiment with information gain where we ran the model with 2/3rds of the data and then calculated the information gain on the 1/3rd of the data. We then split the 1/3rd of the data into two halves based on high and low information gain. We then ran two models, one with 2/3rd data and the low information gain points, and the another one with 2/3rd of the data and high information gain points. We observed that the model that included the high information gain points had predictions closer to the original model, that was trained with the entire training data set. We have added the results of this experiment in the appendix B.
232	Use different words, this is not true.	Agreed. The sentence now reads: "… . *Notably, our two-model approach for shelf and deep sea regions captures their individual dynamics with higher accuracy. Compared to previous maps, this helps avoiding artifacts like unrealistic high TOC concentrations seen in some regions.*"
253	This scale bar does not highlight artifacts	Addressed as in Figure 3
254	This sentence is repeated/not needed	Agreed. The sentence is now removed
257	Why does this matter?	Agreed. The global TOC stocks acts as a validation parameter of the model results from a geoscientific point. But it does not matter if it is actually lower than the results of the DNN, as long as it is around the baseline scenario of 168 Pg of global TOC stock as reported by LaRowe et. al., 2020.
260	So there is no difference between the two? Why show both?	Agree that the sentence was misleading. Added the following missing detail for clarification:

		"*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.*"
Figure A2	Where is this referenced in the text or supplemental?	Reference was indeed missing. It is now referenced together with Figure A1 in line 253: … .Examining predictions from kNN and random forests in this section, Figure A1 and Figure A2 show artifacts, particularly in the equatorial Pacific and Atlantic oceans…
267	Why normal? The obs probably is not normally distributed	See our answer to the comment on line 148
281	A lot of these grids are very outdated...	See our answer to the comment on line 84
282	No feature selection?	See our answer to the comment on line 100
Table C1	Why are you using a grid that is all the same value everywhere there is an obs?	Good observation! The feature might play a role in other network architectures (e.g. convolutional neural networks) that with which we experimented in the context of this publication, but it is of no importance to our current model.
319	Why use this kind of chart?	We use the waffle chart to visualize part-to-whole relationships. While pie charts are a more established plot for this task they come with a number of downsides, the most relevant here being their ineffectiveness at resolving small percentages and small differences among multiple classes (Skau and Kosara 2016). Waffle charts perform better in this aspect because they encode information into length instead of angle; human perception is more accurate at interpreting the former than the later (cleveland, mcgill 1984). As an added benefit, waffle charts allow for actual quantification of values (this by counting squares and multiplying them by indicated "value per square").

References:

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

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

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

De Veaux, R. D., and Ungar, L. H.: Multicollinearity: A tale of two nonparametric regressions, 1994

Skau, D., and Kosara, R.: Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts, Computer Graphics Forum (Proceedings EuroVis), 35, 3, 121–130, https://doi.org/10.1111/cgf.12888, 2016

Cleveland, W. S., & McGill, R.: Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, *79*(387), 531–554. https://doi.org/10.2307/2288400, 1984

In this manuscript, the authors provide a new method to predict the OC content in marine sediments using sparse data, in order to then calculate total OC stocks, providing a step forward to properly quantifying the OC budget in marine sediments. However, there are several studies over the last decade that have modelled the OC content in marine sediments, so the authors have to highlight better the novelty of their work and how it improves previous models and their estimates. They do so by comparing their model performance to random forest and k-nearest neighbors, the two machine learning models used in recent studies, but the comparison is insufficient. I highlighted a few sections where I think this can be improved. In addition, the authors discuss their outcomes very superficially, and do not provide greater insight of the complex mechanisms that affect OC content in marine sediments. This study has a lot of potential and the authors should emphasize their outcomes to enhance the impact of their work within the scientific community. I hope my comments below will help improve this manuscript.

We would like to thank the reviewer for the constructive and helpful comments, which helped us to improve our manuscript. We have addressed all the comments that were raised. During the review process, the model outcomes have changed, and the difference between the current version and the previous version is shown in Figure R2.1,

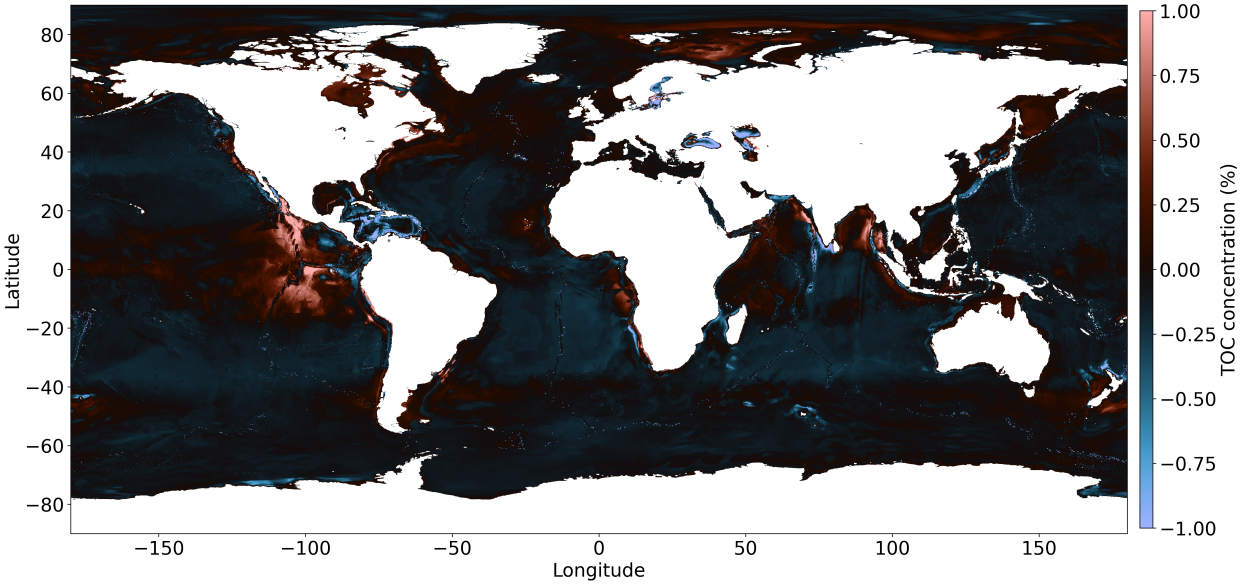

Figure R 2.1: Difference in TOC concentrations between the previously submitted version and the current version, with warmer colors(red) denoting higher TOC concentrations in the current version and the colder colors(blue) denoting lower TOC concentrations in the current version.

The model remains the same as before with the same initialization and hyperparameters. The change in the output is from the pre-processing of the TOC data, and hence the input to the model. The following changes were made:

1. We had some measurements where we had highly variable TOC measurements along the core depth for ex: 0-1 cm x_1% TOC, 1-2 cm x_2% TOC, ... With our pre-processing algorithm when the measurement came from the same location, we averaged them if the standard deviation in the points did not exceed 20% of maximum of the labels and removed all

the measurements if they exceeded this threshold value. This 20 % cut removed around 360 locations, where the depth variance (for 0-10 cm) was greater than 20 % of the maximum value. This was an erroneous approach since we removed values that should have been used to calculate the mean TOC values for the upper 10 cm of the sediment core. We corrected this error and now use all values coming from the top 10 cm to calculate the mean value even if the deviation was larger than 20 %. This included more points in the top 10 cm where the difference is more significant, such as in oligotrophic open ocean regions where bioturbation rates are low. This addition of input in the model changed the model output in the deep-sea, with higher TOC concentrations in regions such as equatorial Pacific, Bay of Bengal, and lower TOC concentrations in general in the deep-sea.

2. For the dataset from Martens et. al 2021, we included data points from greater than 10cm depth. In the current version these data points from higher depths are removed.
   We, hence calculate higher TOC values in the Arctic region, and regions similar to the Arctic, which is shown in Figure R2.1

These reduced the global TOC stock to 156 Pg with a mean TOC concentration of 0.61 %. Including more data points where the difference in the core depth is significant and excluding points from higher depths in the Arctic region makes this model more robust than the initial version.

General comments

1. The authors cleverly divide the global ocean into two regions (shelf and deep-sea) to ensure that the model performs well in these unique environments and avoid the model to simply predict lower TOC concentrations in the deep-sea and higher TOC concentrations on the shelf, and instead ensure that the model captures nuances in each of these settings (which can be highly heterogenous as the authors explain in lines 40-49). The authors use as a cut-off of 200 m. Is this the optimal cut-off? Did they test different water depths? How would the model have performed if the shelf and deep-sea would have not been separated? Is it an improvement to train and validate 2 models? You would need more data to train 2 models than to train one model.

   Have the authors considered doing a third model for the Arctic Ocean? Arctic sediments are affected by OC permafrost and is hence a unique setting. In fact, (Wang et al., 2024) performed 2 models (Arctic and non-Arctic) to predict the radiocarbon signature of surficial sediment on continental margins.

   The edge of the continental shelves rises steeply from depth. They end more or less abruptly at about 200 m below present sea level. We would like to point out that the regions between the 200m depth contour and the coast is typically accepted as the continental shelves. Hence, the continental shelves are defined as submerged parts of the continents (Haas et. al. 2002). The width of the continental shelf varies considerably

though and therefore, the exact water depth used for classification might differ in other approaches. Since this is not a mechanistic or a physics-based model, we did not want to include more complex classifications than required. The model will further classify the feature space on its own by supervised learning.

Due to the accepted norm for bathymetry of 200 m, we did not test other water depths. The model without the split in data between continental shelves and deep-sea resulted in artifacts especially in the Pacific Ocean and other areas, similarly to Lee et al., 2019. This result is shown in Appendix S5 in the corrected manuscript.

The following text has been added in the manuscript (section 4, line 199: *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 the map published by (Taylor R. Lee, Wood, and Phrampus 2019). This additional model shows that the separation of the ocean into shelf and deep-sea regions is required to obtain realistic model results.*

The following figure has been added in the Appendix:

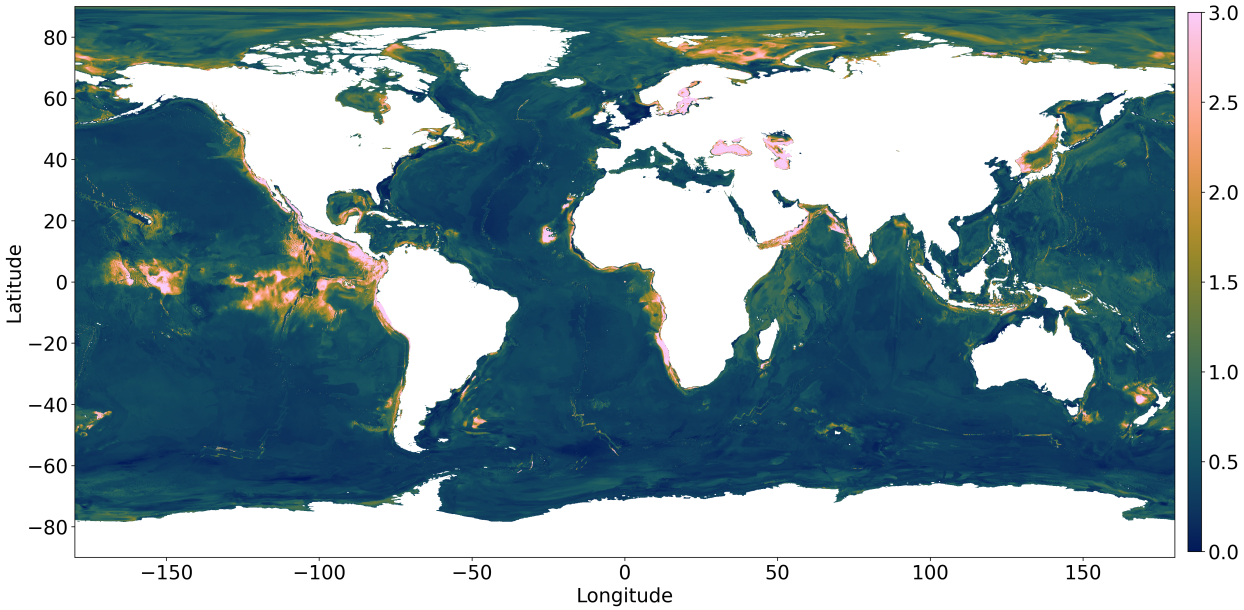

Figure R2.2 *TOC concentration map when the DNN model was not separated into shelf and deep-ocean regions. We see unrealistic TOC concentrations especially in the Pacific Ocean.*

Though we have less data for two models compared to more data for one model, incongruent data points are just noise to the models and increases the aleatoric uncertainty. We have also observed that machine learning models struggle to differentiate the different marine regions when the global ocean is treated as one entity, as seen in Lee et. al 2019 with KNNs and in this work, with DNNs.

We have not considered doing a third model for the Arctic Ocean since this would be beyond the scope of our present work. We do, even, agree that such a model could be a valuable contribution and may address this topic in our future work.

2. A model is just as good as the data that is inputted in it, which is why I went over the 139 features used to train the models, which are listed in Annex C, and the TOC data used in this model.

   In Annex C, the authors provide a table describing each of the features and its sources. In addition to these columns, it would be useful if the authors added an additional column where they argue the importance of including this feature in their models. For instance, bathymetry could affect the OC concentration given the longer transit time through the water column before its deposition on the seafloor. I suggest to do this because there are a lot of features that I don't understand why they are included (e.g., Coriolis).

   We used the features from Lee et. al. 2019 and Restreppo et. al 2021. We excluded features that, based on the expert knowledge, do not affect organic carbon concentrations, for example, the crustal and the mantle properties (see lines 91-92). We included some additional features such as tidal characteristics and chlorophyll concentrations and benthic oxygen fluxes which helped us to improve the predictions. In case of the features with unsure relation to TOC, we decided to keep them. We agree that the model is as good as the input data, however, DNNs are good at extracting only features that are important. From the SHAP values, it was found that the Coriolis effect has a higher importance in the deep ocean than on the continental shelves. We believe that the Coriolis effect indirectly affects sedimentation rates in the deep ocean by influencing the patterns of ocean currents. Some features do not need to have a direct impact on the model output, but could be correlated to a feature that is not included here and could have an impact on the model output. Since it is not a mechanistic model, but a data driven model, we do not a priori distinguish the features that might or might not have an impact on the total organic carbon concentrations. For example, porosity has one of the highest importance in both the continental shelves and deep ocean, and this is a surprising result according to expert knowledge. Possibly, porosity values might represent an underlying geological process that is not directly included in the features, that is more important for the total organic carbon concentration. Hence having an additional column in the feature table about how the TOC concentration is affected by each feature is currently not feasible since the exact processes are still not known for many of the features. As also stated in lines 399-400, 'all effects describe the behavior of the model and are not necessarily causal in the real world". We did, however, included a brief discussion on how the features may affect TOC in Appendix 6 where we address the features that have the highest SHAP values, i.e. the largest effect on our model results.

I also have several questions about some of these features which I include below:

How can the coastline be a feature? A feature should cover the whole region where the model is to be applied to, hence, the whole ocean. Likewise, how can the characteristics of river mouths (carbon dioxide, DOC, bicarbonate, POC, TSS flux at river mouths) be used as a feature in this model?

Coastline is a binary feature where it has a value of one where the bathymetry is zero and a value of zero elsewhere. For the measurements very close to the coast or in the river mouths, coastline is important because some measurements might land on the coastline, because of the grid size. This adds valuable information, if the points are close to the coast or not.

Characteristics of the river mouths are global features similar to the other feature maps. They represent the concentrations and fluxes of different river outputs in the global ocean. The values are non-zero in the coasts or in the vicinity of the river mouths, but the values are set to zero in the deep ocean.

What is the difference between GL_COAST_FROM_LAND_IS_1.0_ETOPO2v2.5m.nc and GL_COAST_FROM_LAND_IS_1.0_ETOPO2v2.r50km.men.5m.nc?

The difference between GL_COAST_FROM_LAND_IS_1.0_ETOPO2v2.5m.nc and GL_COAST_FROM_LAND_IS_1.0_ETOPO2v2.r50km.men.5m.nc is that the first feature is a raw feature and the second feature is averaged over a radius of 50 km. The spatially averaged coastline is a smoothed coastline that is needed for those features that were averaged over a 50 km radius..

Correct grammar for: “Distance OF ocean grid points to the nearest coast”

We corrected the grammar in the manuscript.

Instead of “elevation data”, the authors should be using “bathymetric data”, or am I not understanding well this feature?

The elevation data just includes the elevation on the land as well. In the oceans, elevations and bathymetry are the same. Though we use only the elevation data or the bathymetry in the oceans, the actual feature is elevation, hence the name was kept.

The explanation of Hart-Davis et al.’s (2021) features has a typo (remove “are provided”)
We corrected in the manuscript.

Specify the time range of chlorophyll-a concentrations during summer

and winter.
Chlorophyll-a concentrations in summer is from June to November and in winter is from December to May. The data has been collected from July 2002 till July 2022. This is now detailed in the feature list description (Appendix S1).

When you say “Gradient of elevation”, do you mean “slope”? If that’s the case, please modify.
Yes, we mean the slope. We modified the text in the manuscript accordingly.

Why would you need to have as a feature “Land mask data”. The features should be preprocessed so that they are exclusively of the area of interest. The model shouldn’t learn that if the land is masked, it shouldn’t provide an output.

We used the land mask data as a binary feature that has a value of one on land where the elevation is greater than zero and a value of zero elsewhere. A look at SHAP values from the model runs help us to find if a feature had any impact on the model output. It showed that for the continental shelves model, the land mask had an impact on the model output. The coastline comes from the ETOPO bathymetry data. In the model we used, GL_COAST_FROM_LAND_IS_1.0_ETOPO2v2.5m.nc (the points where the **land topography** is at sea level), and GL_COAST_FROM_SEA_IS_1.0_ETOPO2v2.5m.nc (the points where the **ocean bathymetry** is 0), Though we expected it to be the same, there is an offset in the ETOPO data, between the two. To avoid confusion, we included both files. The coastline data from the ocean bathymetry did not have any impact on the model outputs, while the coastline data from the land had an impact on the model outputs.

In the pre-processing step, we excluded data points, that had a bathymetry of greater than 0 m. But this still included data points from river mouths and very close to the coast, that were inside the coastline boundaries. Hence the non-zero SHAP value (or an effect on the model output) from these features.

Specify what decade is used for the mean sea density, sea bulk modulus (what is that?), average conductivity of seawater, averaged dissolved sea oxygen, sea oxygen percentage saturation, pressure, salinity, oxygen utilization, temperature (etc.), instead of simply saying “over a decade”.
The decadal means are an average over 6 decades from the year 1955 to 2012. This has been updated in the manuscript accordingly in the feature description. More information about the dataset can be found at: https://www.ncei.noaa.gov/data/oceans/woa/WOA13/DOC/woa13documentation.pdf and the features are downloaded from here https://www.nodc.noaa.gov/cgi-bin/OC5/woa13/woa13.pl .

The dataset is obtained from the last depth value in each grid point and hence represent the value at the seafloor.

How does the coastline data “SF_COASTLINE_IS_1.05m.nc” differ from the

previous coastline data provided? Again, how can this be a feature if it doesn't cover the whole area of interest (the ocean)?
The SF_COASTLINE_IS_1.05m.nc, GL_COAST_FROM_LAND_IS_1.0_ETOPO2v2.5m.nc, GL_COAST_FROM_SEA_IS_1.0_ETOPO2v2.5m.nc are similar datasets, but not the same. The SF_COASTLINE_IS_1.05m.nc is derived from NGDC NOAA GLOBE data, while the other two are from ETOPO data (derived from land topography and ocean bathymetry). We found that though all the features are similar, they are not exactly the same and they have an offset. To avoid confusion and knowing the importance of data close to the coast, we used all these features. The training data points can actually have a non-zero value for the feature, if it is very close to the coast. We acknowledge that there are three similar datasets here, that could have been consolidated. However, using all three of these datasets helped us to find points close to the coast, and did not introduce a model error. The sources and the derivation of the datasets are added in the feature list in Appendix S1.

The authors use bottom current data of December 2012, but their TOC dataset encompasses several years. Hence, this feature is not representative of the environmental processes occurring then. The authors should be using average data that encompasses their whole dataset.
We corrected this mistake. The data set is not from December 2012., The dataset has actually a time range from August 1, 1995 to December 31, 2012, temporally averaged. This is now explained in the feature list. More details can be found here: https://www.hycom.org/data/glbu0pt08/expt-19pt1

The feature description has been updated which includes the above information.

This time span, hence, would approximately be representative of the environmental processes occurring.

Sea surface density is extracted from sea surface salinity from the Aquarius project. Why not simply use sea surface salinity instead of sea surface density? Also, please specify the time period averaged to get this feature.
Sea surface density is a function of temperature, salinity and pressure (at the sea surface, this should not be a variable). The time period used to average is between 25 August 2011 to 07 Jun 2015. This is now included in the feature description in Appendix S1. We used surface density rather than salinity in the model since densities reflect vertical mixing processes in the ocean that have a direct effect on biological productivity.

What is the Free-air and Bouguer gravity anomaly?
Free air anomaly is the anomaly arising if the stations are at different distances to the center of the earth and they are often used in marine studies. Bouguer gravity anomaly also includes the rock that is under the instrument and hence includes the anomaly due to it. Hence the gravity

effect of the real topography and the free air is included in the gravity anomaly feature.

Please specify the time period used to extract the maximum and minimum depth of the mixed layer, mean PAR, mean wave direction/height/period, wind speed.
Minimum and maximum depth of the mixed layer is found using two sources. One source is based on observations while the other is based on a numerical ocean model. The fields are from March 1995 to February 1996.

Mean PAR is observed between March 1995 to February 1996.

Mean wave direction/height/period are from the HYCOM data that were measured and averaged over the time range from August 1, 1995 to December 31, 2012.

Wind speed is from the NASA Aquarius mission that was measured and averaged over the time range between 25 August 2011 to 07 Jun 2015.

These specification are now given in the updated feature list (Appendix S1).

Finally, regarding the TOC data:

Why not use the updated MOSAIC database? (Paradis et al., 2023)
We included the MOSAIC updated database and included the data till 2023, as mentioned in line 104. This database and publication is now referred to in the updated manuscript What section depths are included in this dataset? Based on the text, I suppose that TOC concentrations from the upper 10 cm were used, although this is not stated in the Materials section. If so, the same location may have several TOC measurements from the same sediment core, but different section depths (e.g., 0-1 cm, 1-2 cm, 2-3 cm, etc.). TOC concentrations tend to show an asymptotic decrease with depth in a core, being highest at the surface. Hence, TOC measured in different section depths in the same core need to be somehow integrated. How was this done? I imagine the authors did not integrate it using the variance analysis explained in lines 108-112.

Thank you for pointing this out. Ideally, we planned to include all the measurements in the top 10 cm. When we checked through the pre-processing algorithm, we realized that the measurements from the same location that had a depth variance greater than 20 percent were removed. Moreover, we included some data points from the CASCADE dataset from depths higher than 10 cm.

We applied the following corrections: we average the measurements of total organic carbon in the depth of 0-10 cm, if it is from the same location. Afterwards, we perform the variance analysis, in order to reduce the noise due to the lower resolution of the features compared to the measurements. Hence the predictions are the average of the total organic carbon in the top 10 cm of the surface sediments.

This procedure is now explained in the revised manuscript in the section 2.2 (TOC data).

What if a location only had TOC concentrations in the upper 1 cm and not deeper (down to 10 cm), this location would have substantially higher TOC concentration than another core that had measured TOC concentrations down to 10 cm depth.

It is true, that locations where we have the measurements only for the upper 1 cm may yield higher TOC concentration than those where we have sufficient data to calculate the mean TOC over the top 10 cm. The difference is probably small in those regions where bioturbation rates are high (e. g. shelf regions and open ocean regions with high productivity). We also noticed that there were only 360 locations out of 21,000 locations where the variance was greater than 20 percent of the maximum of the measurements. The difference is more significant in oligotrophic open ocean regions were bioturbation rates are low. This issue is now discussed in the manuscript in the TOC data section in lines 100-107. deep-sea.
"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 database includes a total of >110'000 datapoints, including duplicates. How were the duplicates accounted for? If they were accounted through the variance analysis explained in lines 108-112, then the duplicate datapoints would be creating a bias in this variance analysis! Note that these duplicate datapoints (the same sample is reported in more than one dataset) are not duplicate measurements (the same location analyzed in different cruises, or the same sample analyzed in different laboratories).

The duplicates are those measurements that have the same latitude, longitude and TOC concentration. They originate from the overlap of data bases. and were removed from the combined TOC data base. If the same location is analyzed in different cruises or the sample in different laboratories and we have different values, then they are not taken as duplicate measurements, but just considered two different

measurements. But if the standard deviation of the measurements exceeds 20 percent of the maximum value of the measurements in that location, then we eliminate the data, since this would introduce uncertainty to the model. This is now explained in the section on TOC data of the revised manuscript as below in lines:.

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.

3. In this study, the authors compare the performance of their DNN model along with the most often used models in geosciences, k-nearest neighbors and random forest, to prove that their approach is better. However, more detail should be given to how these machine learning models were built. For instance, how were their hyperparameters tuned? What cross-validation approach was used to train the model? This could be explained in Annex A to keep the text simpler.
Hyperparameters: The number of layers and nodes were chosen starting from the simplest model with 3 layers of 8 nodes each. Subsequently, the number of layers and the nodes were increased. In the end, we had 10 layers with 128 nodes in each layer. Even more complex networks with higher number of layers and nodes gave similar results. We chose the simplest network with similar performance on the train and test datasets. We used no structured hyperparameter tuning. We followed a manual approach of choosing the number of layers and nodes based on the train test performance. The weights in the model were initialized with He initialization, which has a zero mean and a specific variance. To introduce non-linearity, ReLU layers are considered a standard in neural networks, apart from specific use cases. This is actually a drawback of the DNNs, and hence not generally preferred because it is not an out-of-the-box method. We updated the text in sections 3 and 3.1 to better explain our approach. Moreover, we added information about the model hyperparameter tuning in Appendix S3 as below in lines 359-367.

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.

Regarding all three models:

How was the train and test dataset generated? Was it random, or did it account for the spatial distribution of the data (ensure that the test dataset comes from all geographical regions), or the feature space (ensure that the test dataset covers a broad range of feature space), or the distribution of the label (ensure that train and test dataset had the same distribution of TOC values as the whole dataset so that the model is trained with all possible TOC values).

We are generating the train/test splits using the sci-kit library function: train_test_split.

The random split inherently considers that the data points are independent of each other. We could implement specialized splitting, such as cluster-based split. However, we think this is not really necessary since data points here are inherently independent of each other and randomizing the splits to the model creates enough room for the model to have some data points out of distribution during the test. Also, with cross-validation, as in Lee et al., 2019, we would still not know if the data points are chosen the way that the model does not overfit to the training data for each fold. This is now better explained in section Results and Discussions in lines 168-169. This issue was also raised by the first reviewer.

“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. “

Were the three models only evaluated once (Table 1)? To properly assess the performance of all three models, it would be better to perform several evaluations with different test datasets, in case the DNN approach happened to perform better using the test dataset presented in this paper.

Yes, the models were evaluated only once. It is possible that the DNN performs worse than random forest and kNN on other splits of the dataset. K- fold cross validation could be a good approach to compare different models and select the most appropriate one for a specific

problem. We do not aim to find the best model for the case, since each of them have their own advantages and disadvantages. The aim was to show that the performance of the DNN was comparable to the KNN and random forests with respect to the performance metrics. For example: Pearson CC is a subpar performance measure and was only included here for consistency with prior works. One of its weaknesses is a very slow climb to the theoretical maximal value of 1.0. Regardless of the scale, we see the values of Pearson CC merely as a confirmation that the model performs better and not worse than previous works. We modified the abstract, hence as follows:

"*For the dataset used, we find that the performance metrics of the models 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 introduce DNN because of its versatility to handle different types of features and a more theoretical approach to uncertainty quantification. This work could be easily extended to unsupervised learning approaches where the parameter space is not only learned on the training data points but the entirety of the oceans. We see this as a potential future work.

Regarding DNN's model performance:

According to Figure 2, the model underestimates TOC concentrations at high label values. What could be the cause of this? How could the model be improved? Since this is an EGU Geoscientific Model Development manuscript, this should be discussed in more detail.

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. This has now been included in the "Results and Discussions" in lines 176-184.

It would also be more informative to plot the spatial distribution of the residuals, to assess whether the residuals present any spatial correlation,

as done in the Supplementary Figures S4-S8 of (Paradis et al., 2024).

We think that the spatial distribution of the residuals would give less information than information gain, since space, which is latitude and longitude, are just two of the features. And we think it is more important to find locations where to sample next to reduce these errors, than finding where the errors are.

The authors identify that, despite the heterogeneous settings in continental margins, the mean TOC concentrations in continental margins (0.69 %) is similar to the mean TOC concentrations in deep-sea sediments (0.66 %). They provide several reasons that could lead to lower TOC concentrations on continental margins such as the effect of sediment reworking, dilution by lithogenics, strong bottom currents, and the effects of bottom trawling. They then state that “According to our DNN-model, these factors decrease TOC concentrations in shelf sediments” (line213). How does the model show that these are the factors responsible to decreasing TOC concentrations? If bottom trawling is a significant factor, then the authors should include it in their model as a feature, and see if this feature is relevant in controlling the distribution of TOC concentrations.
The model does not identify the factors that drive the TOC concentrations predicted. We only provide the possible reasons or causes from the results obtained from a data-driven model. We hence changed the line 242-244 to

“According to our DNN-model, these factors could potentially decrease TOC concentrations in shelf sediments to such a degree that they attain mean values that are close to those observed in deep-sea sediments.”

Including bottom trawling as a feature is a very good idea. We would like to include global trawling data as a feature in our future work.

One of the novelties of this manuscript is the determination of “information gain”, the identification of key regions where more TOC measurements would improve the model’s performance and reduce the uncertainties in these areas, which could be used to guide future research campaigns and fill these gaps. However, this is barely discussed in the text, and in a very confusing way:

Line 215: The choice of words is a bit confusing. It sounds as if the high TOC concentrations in the Norwegian Trench should make this area have a high uncertainty and a high information gain. Please revise.
Line 217: The authors mention that there is a scarcity of data from the Gulf of Mexico (in addition to other areas), but in the Materials section, they mention that they had large regional datasets from the northern Gulf of Mexico (line 105). Please correct this contradiction.
Lines 219-220: The choice of words here is also confusing. In this paragraph, we have the impression that data clusters present lower information gain whereas areas with scarce datapoints have higher information gain. However, this sentence then states that “our analysis also reveals that an abundance of measurements does not necessarily

correspond to lower information gain, and vice versa". Please revise this section so that the reader is not confused by the contradicting sentences. Lines 220-222: The authors explain here that the information gain is a balance of the amount of datapoints and their proximity to parameter space and congruency of the measurements made there. This is very relevant and the authors should emphasize this better. What parameter space is seldomly sampled? Which regions show a low information gain despite the scarcity of datapoints? Why is this the case? What features make these regions have a low information gain? Similarly, which regions show a high information gain due to variability in the measurements? What would be the reason behind this variability in the measurements? Is it seasonality? This would be very insightful and give more relevance to this manuscript.

We have revised the section on information gain (section 3.1). Because of the change in the dataset, we also see a change in the information gain map. We see higher information gain in the equatorial Pacific region, Zealandia and around Papa New Guinea, which are relatively less explored regions. We also found that the information gain was higher in locations with steep slopes deep-sea around ocean islands and ridges Hence, we think that the geology of the region has a greater effect than the seasonality. There is a possibility to find which feature provides more uncertainty, by introducing InfoSHAP by Watson et. al 2023. This is a potential future work, and we would like to focus only on the information gain, considering all the features. We conducted an experiment, where we show that the higher information gain points provide more model knowledge compared to the low information gain points.

The revised section can be found in Results and Discussions in lines 255-267 is as follows:

"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. the regions with the 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 of West 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."

According to the model output, deep-sea basins have large TOC stocks, as was observed in previous modelling approaches (Atwood et al., 2020), and the authors mention that "this underscores the importance of deep-sea environments in the global carbon cycle". However, the large TOC stock is essentially due to the vastness of deep-sea basins, and doesn't necessarily mean that they are more important than continental shelves in the global carbon cycle. Moreover, when accounting for OC burial in marine sediments, the large TOC stock of deep-sea sediments would be reduced due to the low sedimentation rate in these regions in comparison to continental shelves. Hence, the TOC stocks in continental shelves and deep-sea basins are actually not really comparable (at least not to conclude that deep-sea basins are more important in the global carbon cycle). To make this clearer, I suggest the authors discuss the influence of sedimentation rate and the influence of the sediment age at the depth employed (10 cm). See for instance the recent study by (Bradley et al., 2022).
We agree that it does not necessarily mean that deep-sea basins are more important than continental shelves. Hence, we removed this line from the manuscript to avoid any confusion. It is interesting to note that most TOC burial happens in the shelf sediments, where the sedimentation rates are higher, due to the deposition of riverine particles, as stated by Bradley et al,. 2022. We added this information in the manuscript in the Results and Discussion section in lines 244-246.

"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 et al. (2022)"

The authors conclude that "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." (lines 243-246). However, the manuscript doesn't discuss the complex interplay between biological, physical and geological processes in marine sedimentary environments, which would be very insightful for the scientific community. In addition, while the manuscript does an excellent job at identifying future research efforts (albeit it could be improved with the suggestions provided in comment # 6), the authors don't identify regions where management efforts should be made to preserve and manage marine carbon sinks (i.e., vulnerable areas where

high TOC contents could be affected by anthropogenic activities if unprotected). The authors should either modify this concluding sentence or modify the manuscript to discuss the OC mechanisms and vulnerable areas that require preservation.

Since we do not aim to develop a mechanistic model, or a numerical model based on physics and processes, we do not discuss the complex interplay of biological, physical and geological processes. We believe that the processes are already described in detail in papers such as La Rowe et al., 2020. Instead, as a data driven model, we use features that might be relevant for the prediction TOC concentration. Similarly, the main focus of the paper is not management efforts, but rather focuses on the model developed and the advantages of the model.

The authors present additional information in their supplementary information that is not discussed in the text that would be very relevant for this study. For instance, in Appendix E, the authors discuss the model's interpretability and the influence of different features. Although the authors state that "All effects describe the behavior of the model and are not necessarily causal in the real world" (lines 306-307), this analysis could be used to better understand the spatial distribution of TOC (see earlier comment). For instance, why is sediment porosity the most important feature in the DNN model? What are the implications of this? The authors should include a section in the manuscript addressing the model's interpretation.

In Appendix F, the authors visualize the TOC stocks using different visualization techniques, but don't reference it in the text. In my opinion, this does not add additional scientific insight to the manuscript, and since it is not even discussed in the manuscript, I would remove it.

We believe that including model interpretability would confuse the readers and wanted to be careful in highlighting these results, because they might not be representative of the real world, but just the particular model. If any of the hyperparameters of the models are changed, then the feature importance list would be different, with different SHAP values. We included lines 247-254 in the "Results and Discussion" to guide the reader to the Appendix S6 for further reading.

"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 TOC distribution (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."deep-sea

Finally, I suggest the authors remove “NN-TOC v1” from their title, as this naming convention is not used throughout the manuscript.
We agree that this has not been used elsewhere in the manuscript. The journal required a title for the model, as this is a model-description paper. This is the name used in the code repository, and hence has been used in the title here.

Specific comments

Lines 37-39: This sentence is very important in terms of the objectives of the study. However, it cites a paper that already quantifies OC stocks, so what’s the novelty of this study?
The main novelty of the study is the introduction of deep neural networks in geosciences and presenting its potential with respect to quantifying uncertainty and generalization. kNNs and random forests have been widely used before and it has its own shortcomings, especially in the case of higher dimensional spaces. We also introduce the concept of information gain, to guide future sampling. In terms of quantifying OC carbon stocks, the global TOC stock is a good validation check if our model gives acceptable results. As far as we know, this is the first time there is a quantification of TOC stock in different marine regions (shelf and deep-sea). Moreover, we find that the mean TOC concentration in both the continental shelves and deep-sea is not as different as is normally expected.

Lines 40-49: The authors very nicely explain the heterogeneity of continental shelves and deep-sea sediments and how the OC content varies in these different settings. To the best of my knowledge, this study discusses, for the first time, the spatial distribution of relict sands, but this is not a feature that is included in the model. In addition, what about the OC content in other unique geomorphological features, such as fjords and canyons, that should also be taken into account considering their global extension? Finally, the authors finish this paragraph with an estimation of the global mean TOC concentration, which is not only not insightful for this paragraph, but also it is not clear how this was calculated and if the authors have considered the spatial extension of the different regions they have described (deltas, upwelling margins, relict sands, etc.). Instead, I suggest the authors highlight that marine sediments are highly heterogeneous, which complicates a proper quantification of TOC concentrations in marine sediments and its spatial distribution, which is the purpose of this study. This is especially important and novel since they highlight that they will “improve the accuracy of highly heterogenous and undersampled geological settings” (line 78).
The lower concentration of TOC in continental shelves from the model predictions is explained by relic sediments, but we do not take this as a feature. Relic sediments are included to only explain the model predictions. Since the model is not a mechanistic model, we did not include all the features that might be the reason for the model predictions. But these features could be a reason why the model predicted these values. The same reasoning applies for not including other geomorphological features, such as fjords and canyons.

The calculation of the mean TOC concentration does not consider the regions such as deltas, upwelling margins, relic sediments etc. The mean TOC concentration is calculated with the TOC concentration from each cell and the

area of the cell in each of the marine regions (the different oceans and seas). It is then summed up from each of these marine regions, to provide the mean TOC concentration globally. We include the excel sheet used for the calculation of the total mean TOC concentration globally in the supplementary information for better clarification. The code used to calculate the mean TOC concentration in each region from different cells is in /notebooks/TOC/Visualisation.ipynb .

Please be a little bit more descriptive with the Figure captions. Figure 4's caption is: "TOC stock map". You could specify in the figure caption the section depths included in the calculation (for Figure 3 as well), how the TOC stocks were calculated (using porosity map provided by Martin et al., 2015 and sediment density of 2.6 g/cm3), and also note that the colormap is in logarithmic scale. With respect to Figure 5, does Information gain have a unit? I imagine that more sampling should be done in areas that have an information gain of 1 rather than 0. Please include all this relevant information in the figure caption.
We modified the figure captions to be more descriptive. In figure 5, information gain does not have any units. Yes, this is true more sampling should be done in areas that have a higher information gain. Information gain is not bounded by 0 and 1, but is non-negative, and hence ranges between 0 and infinity.

Appendix A: A description of Figure A1 is given, but not of Figure A2. Please be consistent and describe the output of both figures.
We overlooked this and modified it in the manuscript.

How does the spatial resolution of this output compare with previous work by Lee et al. (2019) and Atwood et al. (2020)?
The spatial resolution of 5 arc minute grid is the same compared to the previous works by Lee et al. (2019) and Atwood et al. (2020)
Technical corrections

Line 23: Against what background? What are the authors referring to here?
We referred to the state of the art and the current challenges. We modified the line and can be found in Introduction in lines 24-26.
"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."""

Line 31: Sala et al. (2021) focus their work on the effect of bottom trawling on OC, and not on marine sediment resuspension and erosion. I would suggest citing (Oberle et al., 2016).
Since the line also talks about the effect of bottom trawling on OC, we would like to use both the citations and hence added it in the Introduction section of the manuscript in lines 32-34 as below:
Shelf sediments are also affected by human activities such as bottom-trawling and dredging that erode and disperse large sediment volumes (Sala and Lubchenco, 2021; Oberle et al., 2016)

Line 32. Remove additional "of" in: "It is composed ofboth"
We corrected the manuscript.

Line 72: Refer to “TOC stocks” instead of “TOC inventory” for Atwood et al. (2020) example, to be consistent with the use of this terminology
We changed it into “TOC stocks”.

Line 74: Avoid using “this background”. Makes the user have to interpret what you mean.
Given these challenges, this paper aims to derive more robust maps of TOC concentrations and inventories for the global ocean.

Lines 100-101: Instead of saying that the feature list is in the Supplementary Information, state where we can find it (Appendix C). The same should be done for Line 149 regarding the mathematical formulation of the entropy (Appendix B). Similarly, state that further description of the results of the different models is given in Appendix A in lines 164-165. Finally, restructure the Supplement in order of appearance in the text. Right now, Appendix C is referenced before Appendix B, Appendix B is referenced before Appendix A, and Appendix E and F are not even referenced.
We modified the manuscript by including the Appendix references and restructured the supplement. We also referenced the Appendices that were not included before.

Line 190: I imagine you are referring to TOC stocks, and not TOC concentrations.
We refer to the difference in the TOC concentrations between different marine regions in the continental shelves and the deep-sea in Table 2.

Line 229: Remove this sentence from the conclusions.
This sentence has been removed in the manuscript.

Line 274-275: This sentence is not grammatically correct. I wouldn’t know what change to suggest since I’m not sure I understand it.
The sentence has been modified as:
“The information gain measures the difference between the cross entropy (Equation B2) and the entropy (Equation B1).”

Line 277: DKL is always “positive and” remains well-defined […]
We mean non-negative, because information gain can be zero. Positive would mean that information gain cannot be zero, which is false.

Line 294: Provide Zenodo link
We included the link in this line.

Table 2: Note the typo in the units of TOC stock (Pg).
Typo has been removed in the manuscript.

References in this review:

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

Bradley, J. A., Hülse, D., LaRowe, D. E. and Arndt, S.: Transfer efficiency of organic carbon in marine sediments, Nat. Commun., 13(1), 7297, doi:10.1038/s41467-022-35112-9, 2022.

Oberle, F. K. J., Storlazzi, C. D. and Hanebuth, T. J. J.: What a drag: Quantifying the global impact of chronic bottom trawling on continental shelf sediment, J. Mar. Syst., 159, 109–119, doi:10.1016/j.jmarsys.2015.12.007, 2016.

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 Syst. Sci. Data, 15(9), 4105–4125, doi:10.5194/essd-15-4105-2023, 2023.

Paradis, S., Diesing, M., Gies, H., Haghipour, N., Narman, L., Magill, C., Wagner, T., Galy, V. V., Hou, P., Zhao, M., Kim, J.-H., Shin, K.-H., Lin, B., Liu, Z., Wiesner, M. G., Stattegger, K., Chen, J., Zhang, J. and Eglinton, T. I.: Unraveling Environmental Forces Shaping Surface Sediment Geochemical “ Isodrapes ” in the East Asian Marginal Seas, Global Biogeochem. Cycles, 38(4), doi:10.1029/2023GB007839, 2024.

Wang, C., Qiu, Y., Hao, Z., Wang, J., Zhang, C., Middelburg, J. J., Wang, Y. and Zou, X.: Global patterns of organic carbon transfer and accumulation across the land–ocean continuum constrained by radiocarbon data, Nat. Geosci., doi:10.1038/s41561-024-01476-4, 2024.

Citation: https://doi.org/10.5194/egusphere-2024-1360-RC3

References for the answers:

Martens, J., Romankevich, E., Semiletov, I., Wild, B., Dongen, B., Vonk, J., Tesi, T., Shakhova, N., Dudarev, O., Kosmach, D., Vetrov, A., Lobkovsky, L., Belyaev, N., Macdonald, R., Pie\’nkowski, A., Eglinton, T., Haghipour, N., Dahle, S., Carroll, M., Åstrom, E., Grebmeier, J., Cooper, L., Possnert, G., & Gustafsson, \. (2021). CASCADE – The Circum-Arctic Sediment CArbon DatabasE*. Earth System Science Data, 13(6), 2561–2572.*

Henk de Haas, Tjeerd C.E van Weering, Henko de Stigter: Organic carbon in shelf seas: sinks or sources, processes and products, Continental Shelf Research, Volume 22, Issue 5, 2002, Pages 691-717, ISSN 0278-4343, https://doi.org/10.1016/S0278-4343(01)00093-0. (https://www.sciencedirect.com/science/article/pii/S0278434301000930)

David S. Watson, Joshua O'Hara, Niek Tax, Richard Mudd, & Ido Guy. (2023). Explaining Predictive Uncertainty with Information Theoretic Shapley Values.