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: "*Furthermore, we introduced a method for quantifying uncertainty using Monte Carlo dropout. The method was applied to our neural network model and underlying features to generate a map of information gain. The map shows the expected increase in model knowledge that can be 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. "*Against this background, 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… and undersampled geological settings, and should do a better job at capturing the non-linear relationships between TOC and other geological features. ”
84	I made note in the Appendix about this section, it is worth nothing these features are very old and many of them are outdated or irrelevant (where there is no variance across a feature for observed locations).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: “*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 available in the updated version.*”
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