I truly enjoyed reading this manuscript and believe it to be in very good shape overall. The study was well-designed, and the manuscript well-written. I recommend publication generally, but would like to raise a few points that the authors and editor should consider before proceeding. I am not an expert on marine carbon, and I am not a statistician, but I have several questions about those components, and can additionally provide perspective on the sediment and geospatial modelling components. I hope these comments are valuable.

General comments

The Introduction provides a good amount of background on the motivations for this work. It appears to sufficiently cover the current state of knowledge regarding broad scale marine carbon modelling. I enjoyed reading it, and have nothing major to suggest.

The level of detail describing the predictor variables is quite good. I believe that the reader has enough information here to go ahead and extract similar data if they wanted. That is great.

The effort undertaken to extract useful data from the literature is very impressive. This is one of the best parts of the manuscript to me. It is highly laudable to attempt an exhaustive compilation of existing data from the published literature, rather than ignoring large amounts of previous work in favour of expedient downloads from one or two large repositories. I would like to see more of this for such regional mapping projects.

The Methods section is a bit long, but this may be acceptable given the scope of the project. If the authors are able to improve the conciseness of this section it would be nice for the reader, but maybe that is not possible. Food for thought.

I am not sure if the method employed to calculate confidence intervals for various location specific carbon parameter predictions, and also for the estimates of carbon stocks across the entire Canadian margin, is appropriate. The same logic is applied in many places to derive confidence intervals for a parameter that is based on confidence intervals from a prior model prediction, which I discuss in a bit more detail below. This is my one major concern with the manuscript. Happy to be corrected if I am mistaken here, but please see my comments on the subject below.

Specific comments

1) **Page 5.** The colour ramp for the bathymetry is unorthodox. This is not necessarily a problem, bit it looks a bit odd. Is there any reason why dark blue is used for shallow and light yellow for deep?

2) **Page 10-11.** Suggest making it clear here that “SPM” refers to suspended particulate matter. It is not formally defined.

3) **Page 11, L 245.** Consider using the word “disparity” rather than “dissimilarity” here.

4) **Page 19**. It is not currently clear to me why the MAR netCDF had to be sampled and modelled when you already had full-extent spatial predictions of MAR. Why not just resample/interpolate these to the appropriate resolution and projection rather than modelling them? They have already

been modelled once in the first place – it seems like a second round of modelling may just compound the error associated with these predictions.

5) **Page 20, L 543-560.** I found this section a bit difficult to follow. It would help to be very consistent with the terminology here. For example, terms like "analysis data" and "assessment data" are used. Are these the same as "training" and "test" data, which are used later on? I am pretty sure these are the same thing, but consistency would be good, especially because the validation procedure was fairly involved.

   Also in this paragraph, you write that "*…This function creates density plots of nearest neighbour distances in multivariate predictor space between all response data…*" By definition, aren't nearest neighbour distances the distance between a data point and its nearest neighbour (in multivariate space), not between all data points?

   Relatedly, could you make it clear what these multivariate distances are? Were the predictors normalized before calculating (Euclidean) distances? That would have a very large impact on the distance calculations. This is also not clear from the appendix figures, which are just labelled "dist".

6) **Page 21-22, L 586-605**. Can you clarify how the a priori feature selection was accomplished? You write that variable importance was calculated using a "*basic random forest model with all training data and predictor variables…*" Does this imply that you used the out-of-bag samples to estimate variable importance, or was it estimated on the training data?

7) **Page 22**. I found this section also hard to follow, and again, believe it would be helpful if the terminology was consistent. I'm assuming the "training" and "test" sets here are the "analysis" and "assessment" data from above? This may not be obvious to some readers because "assessment" or "validation" data are often used for model tuning, while a separate "test" set is withheld to calculate the final accuracy.

   I'm having trouble understanding the hyper-parameter optimization and validation design described on this page. Can you confirm that separate models were built and tuned for each fold, and were also validated using the same design, and that the out-of-bag (OOB) samples were not used anywhere in this procedure? If that is the case, why not use the OOB samples to tune hyper-parameters, then use the folds for validation? Tuning hyper-parameters on the validation data can slowly lead to a form of overfitting.

8) **Page 22, L 615**. Sorry, but in the same vein the terminology is confusing to me here. You write that:

   > *"Overall model performance metrics (RMSE and R2) were then calculated using the predictions across all CV folds with optimal hyperparameters and the last-fit; while predictor variable importance was calculated by fitting an additional model across all training data using optimal tuning parameters and the importance calculated through permutation."*

   Does this imply that, while performance was calculated using CV, the variable importance was calculated using the training data (i.e., the "analysis" data)? When you say permutation,

does this mean you permute the predictor variables and measure the impact of the model predictions on the training fit then? Or are the OOB data being used here? How were these "optimal tuning parameters" selected if hyper-parameter tuning was performed per fold?

9) **Page 22**. These paragraphs appear to imply that you predicted each of the 10 models from the CV across the entire domain, correct? Or was it a single model fitted "*across all training data*" from above? In either case, what does it mean that the data were randomly split into 150 samples, predicted, then merged to produce the final raster layer? What data were split and why is it random? Please clarify this.

10) **Page 23, L 658**. "Fourth", not "forth".

11) **Page 24, L 675-678**. Does this mean that the final bulk density estimates were the mean of the 14 individual estimates? How did you get standard errors after aggregating these predictions? Did you just use the standard deviation of the 14 predictions? How then can you get a 95% confidence interval? This needs clear explanation. As far as I can tell, you have a 95% CI for four of the transfer functions… how could you possibly pool these, in addition to the other estimates?

12) **Page 24, L 683-685**. You use the 95% CI bounds of the %OC predictions and the MAR predictions to calculate a 95% CI for OC accumulation rates. Why is that appropriate? For example, the upper bound of the %OC and MAR intervals implies you are 97.5% sure the mean value at a given location should be less than that value. Your calculation implies that, given 97.5% confidence that the mean OC% is < x, and given 97.5% confidence that the mean MAR is < y, you are 97.5% sure that the mean OC accumulation rate is < xy. It may be worthwhile to consult a statistician here, but I don't think that's the proper way to calculate a confidence interval. Would you not need some information about the joint distribution of these variables? I'm not a statistician; if I am wrong, please provide some source for this approach.

    This has very important implications for your calculations across the entire model domain. Is it appropriate to take the upper and lower confidence bounds of multiple variables, multiply them together, and aggregate those to arrive at 95% confidence that the total OC content across the entire shelf is within a certain range?

13) **Page 26 and surrounding**. A minor comment, but it would be nice if the fonts of all plot elements (e.g., axis labels) were consistent across figures.

14) **Page 28**. The high importance of the predicted mud content raises an important question for me regarding the potential for "data leakage" in your validation. Can you confirm that none of the %OC measurements came from data points where the mud content was also measured and used for those models? It seems possible that this could have occurred since many of both data points were sourced from NRCan, and that there were many more mud samples than %OC in your dataset.

    If the mud data points were used to create map predictions at all locations where you have %OC measurements, yet the %OC measurements were sourced from data points that occur in the mud dataset, it is possible that the test data from your cross-validation folds are "leaking" information

to the training data via the predicted mud layer. This would occur due to expected correlations between the mud and organic carbon content if they were obtained from the same sample, and not assigned the same CV folds. It would potentially inflate both the estimates of importance of the mud layer, and the estimates of model performance, despite your substantial efforts at ensuring an appropriate CV design.

15) **Page 32**. Similar comment to previous. You are assuming that the 95% CI can be acquired for dry bulk density predictions simply by performing the calculations using upper and lower bounds of the mud predictions. I'm not sure that is the case. I think it is true to say that you are calculating the *likely dry bulk density given the 95% CI bounds of the mud predictions*, but I don't think that is the same as calculating the 95% CI of the dry bulk density.

16) **Page 35**. Same comments as previous regarding the propagation of confidence intervals.