Author Response

Dear Reviewer,

We are very grateful for your valuable feedback and your time and effort in reviewing our manuscript. We have carefully considered all the comments and have done our best to address each one.

Below is a detailed point-by-point response to the comments. We have attached a supplement with a list of the new and revised figures and how they differ from the first submission. All figures and Table numbers in our reply refer to the figures included in the attached supplement.

Yours sincerely and on behalf of all co-authors,
Sweety Mohanty

****

RC 01: *1) The method section - and in particular section 2.3 onward are difficult to read especially for folks that are not familiar with machine learning. This is the result of the many specific terms used (e.g. "merging at a higher height", "Ward variance", "Euclidian Distance", "Ward Linkage", ...). To make the methods section more accessible to the wider audience of the journal, I would suggest to provide less technical text in the main section and add the required detail and terminology to the appendix.*

AR 01: Indeed, the accessibility and comprehensiveness of the text shall be ensured. For that purpose, we rephrased the terms in such a way that they become more intuitive.

- The phrase 'merging at a higher height' implies that clusters are grouped together on the dendrogram's vertical axis.
- Euclidean distance is the distance between two points in space
- Ward variance aims to group data points so that the variance is minimized.
- Ward linkage implies the link generated after clustering data points after the application/ calculation of Ward variance.

The text in the revised manuscript will be updated accordingly.

RC 02: *2) I dont find the argument about the choice of an MLR that convincing. Figure A1 clearly shows that the relationships are not "perfectly linear". Furthermore, the arguments*

provided on lines 178-180 that the MLR is faster and easier interpretable are only to a certain extend true. Using e.g. a simple single layer FFN instead of the MLR could account for the slight divergence from linearity without compromising on speed. For me, the main argument is interpretability. The single weights of the MLR are easier to interpret and process than the more complex weight Matrixes of a FFN.

AR 02: Thank you for challenging the argument about the choice of MLR. Your objection that runtime performance is not a valid/convincing argument for relying on MLR instead of an FFN is a strong point. As you suggested, we redirected our focus to the interpretability of the weights of MLRs compared to FFN weight matrices.

RC 03: 3) This may be a misunderstanding on my end, but I am still puzzled why you need a FFN for the time variation in the biomes. I fully understand the approach and I endorse it, but would you not also get changing biomes by doing the MLR followed by the hierarchical clustering for each month/year separately? Thought he changing HC relationships, you would also get changes in the biomes, no?

AR 03: We thank you for bringing up this point. The challenge of how to track the carbon biomes, in the absence of an initial set of labeled clusters, is one to which we devoted considerable thought. Since our approach is fully unsupervised and clusters do not have pre-defined labels, performing a one-to-one mapping between biomes over months and years becomes difficult. Indeed, we initially also used the approach that you suggest. In Mohanty et al. (2023) we tested whether the carbon biomes could be tracked only based on the changes of the feature-driver relationships over time. Specifically, we ran the clustering algorithm for every month of each year and then calculated the Frobenius Norm, a distance metric, between the RCs of each cluster between one time step and the next. We subsequently matched those clusters exhibiting the lowest Frobenius distance. However, this approach was not optimal, since the cluster number was not fixed and the Frobenius Norm failed to match all clusters over time.

Hence, for this manuscript we decided to use a more flexible tracking method. We selected a random month/year to define labels (please see also reply #6 to Reviewer 1 for more details on this particular aspect) and trained a neural network to learn the associations between labels and the underlying regression coefficients (RCs). This allowed us to predict the locations of the same labels over time in an effective way. As shown in **Fig. H1**, the RCs within each biome are (with the exception of the strongly variable ICE biomes) substantially stable over all months and years.

This aspect is now explained more thoroughly in the revised manuscript.

RC 04: *A more general question I had that was not answered in the paper: is your approached that was designed from a single model easily adoptable for other models?*

AR 04: The tool (detection and tracking) is transferable to other ocean models. Still, the exact biome locations and model weights and parameters (linear regression, clustering, neural networks) are not, as they are specific to the ocean model experiments. Ideally, the tool should be run from scratch based on a specific model output or observational data set. Requirements for this include saving surface CO_2 fugacity or partial pressure, DIC, alkalinity and temperature at a resolution sufficient to resolve the seasonal cycle. It should be noted that if the data set has a coarse resolution, bigger boxes will be needed to construct the spatial target-driver relationships, with the drawback of achieving a less refined picture, of cutting through sharp current, or of averaging out different regimes.

This study does not provide the exact locations of static biomes, but rather - due to the temporally-varying nature of the ocean - a probability that a certain biome will be found in a certain place (**Fig. 5**). We acknowledge that it is practical for model evaluation or intercomparison studies to define a set of static biomes over which different model diagnostics can be averaged (as done in e.g., DeVries et al., 2023). However, the usage of static biomes comes with disadvantages owing to time-varying and model-specific locations of ocean fronts (Fay and McKinley, 2014). We suggest that future research could use the detection and tracking tool proposed here across different data classes to test the spatial homogeneity of the carbon biomes as well as to better understand the specific dynamics of each model.

RC 05: line 20: please add "annual" to the 25% (the number refers to the present day uptake rate - historically, over the industrial period, the ocean uptake was larger)

AC 05: Added.

RC 06: *line 130: Please provide more detail how the outlier removal was done*

AC 06: We removed data points with pre-industrial DIC below 1500 micro-mol/kg, alkalinity below 1700 micro-mol/kg, and pre-indusrial fCO_2 above 500 micro-atm.

RC 07: *lines 255-270: The architecture of the NN are provided by no justification to why. Have you done some optimalization testing (e.g. on the optimal number of neutrons), or are these subjective choices?*

AR 07: We did not conduct a defined set of tests to choose the number of layers and neurons (which affect the model complexity), and we tried a very small batch of NN

architectures. We focused rather on optimizing different hyperparameters of a selected FFN architecture

In the revised manuscript, we now provide more detail on the optimization testing, as detailed below.

- First, we fixed an architecture with the number of layers to six (a relatively small number given the number of input features, i.e., three slopes) and no. of neurons per layer with 64, 128, 256, etc. (a standard practice in ML community to choose neurons as a power of 2) to conduct sensitivity analysis for the hyperparameters (optimizer, learning rate, batch size, no. of epoch, etc.) used in the neural network architecture. We have utilized a 5-cross validation technique for this analysis.
    - This can be found here: https://github.com/swemoh/Detection-and-Tracking-of-Carbon-Biomes/tree/main/jupyter_notebooks/output_files (the link for Gihub is available under Code Availability in the revised manuscript.)
    - We chose the parameters for which we received lower test loss, higher and comparable accuracy, precision, and recall and then trained the model to track the carbon regimes.
- Second, to understand how the model complexity could affect the outcome, we kept the hyperparameters constant and attempted to alter the neural network architecture by removing/adding a hidden layer, which resulted in a decrease/not significant increase of test loss and accuracy (not shown).

We could only check an NN architecture's accuracy, precision, etc., with different hyperparameters. Experimenting with different non-linear functions, individual weights, and biases for every layer of NN could be incredibly time-consuming. Neural networks can be opaque, making it hard to interpret precisely what each neuron has learned. This is especially true for deep networks with many layers. Besides, interpretability techniques like Layer-wise Relevance Propagation (LRP), SHAP, and LIME could have been applied to understand which features are essential and how neurons or layers contribute to the model's prediction capabilities. We suggest that this could be an outlook to be addressed in future research.

RC 08: line 471: "personality" is an odd choice of wording

AR 08: Updated to 'nature'

RC 09: lines 485-490 are a repeat from the introduction and can be removed in my view

AR 09: In the revised manuscript, we rewrote the sentence to make it more compact.

References

DeVries, T., Yamamoto, K., Wanninkhof, R., Gruber, N., Hauck, J., Müller, J.D., Bopp, L., Carroll, D., Carter, B., Chau, T.T.T. and Doney, S.C., 2023. Magnitude, trends, and variability of the global ocean carbon sink from 1985 to 2018. *Global Biogeochemical Cycles*, *37*(10), p.e2023GB007780.

Fay, A.R. and McKinley, G.A., 2014. Global open-ocean biomes: mean and temporal variability. Earth System Science Data, 6(2), pp.273-284.

Mohanty, S., Kazempour, D., Patara, L. and Kröger P., "Detection and Tracking of Dynamic Ocean Carbon Uptake Regimes Built Upon Spatial Target-Driver Relationships via Adaptive Hierarchical Clustering," 2023 IEEE 19th International Conference on e-Science (e-Science), Limassol, Cyprus, 2023, pp. 1-10, doi:10.1109/e-Science58273.2023.10254820.