We are grateful to Dr. Patrick Rafter for dedicating his time and providing constructive comments, which are instrumental in refining this manuscript. Below, we have thoroughly addressed every comment, and the original review text is presented in italics.

The manuscript "Reconstruction of d13CDIC in the Atlantic Ocean…" as reviewed by Patrick Rafter

First, I'd like to thank the other (anonymous) reviewer for their careful and useful review of this manuscript. If I were the author of this manuscript, I would greatly appreciate the many meaningful and well-informed comments. I don't fully agree with all their suggestions, but it is undeniably a high-quality review.

*For example, I think—for the most part—this study needs less additional work than the other reviewer. The suggestion to implement the ML method in a model environment would be a very interesting and valuable addition to this work, but I predict the authors' response will be "outside the scope of the current study". It sounds to me like a huge amount of new work, but I may be incorrect in this (or it may just be a huge amount of work for *me* and not someone else (it almost surely is)). Note that I do not have the experience in this space to comment on whether this model environment application is "now common practice", but I will say that this would have been a novel (to me), interesting, and seemingly robust application of the methods developed here. But I would like to note that if this manuscript / dataset were to follow the reviewer's advice, it would boost my score for the "significance" and "data quality" categories into and above the 'Excellent' category. As of now, I have scored these as 'good'.*

I also think the motivation is appropriate for this specific study and that the decadal trends in the Kernel Density Estimates (see Fig. 8) are an interesting outcome from this study (as it exists now).

R: Thank you for your positive and constructive feedback on our manuscript. We greatly appreciate your recognition of the study's value and your thoughtful reflections on the other reviewer's comments.

Regarding the suggestion to validate the Machine Learning method in a model environment (raised by the other reviewer), we are pleased to inform you that we implemented this supplementary validation, and it was indeed feasible within the scope of the current study. As noted in our response to the other reviewer, the model dataset they referenced (https://doi.org/10.5194/gmd-17-1709-2024) does not provide the required carbon isotope data. Instead, we adopted the well-validated model data from Claret et al. (2021), which includes comprehensive carbon isotope simulations ideal for this validation purpose. Following the proposed workflow, we subsampled the model outputs across time and space, reconstructed the 4D $\delta^{13}C_{DIC}$ distribution, and thoroughly evaluated the model's performance.

This supplementary validation not only confirms the method's ability to accurately reconstruct spatiotemporal patterns from sparse and noisy data but also reveals its strengths in mitigating sampling biases, effectively addressing the limitations of validating solely with sparse observations. All details of this model-based validation, including data processing steps, evaluation metrics, and key results, have been added to the Appendix of the revised manuscript for transparency and reference.

We are grateful for your note that this additional work would enhance the study's "significance" and “data quality” categories. By incorporating this model-based validation, we aim to strengthen the scientific rigor and reliability of our research as you suggested. Thank you again for your valuable input and support. Your feedback has been instrumental in refining our work.

Claret, M., Sonnerup, R. E., and Quay, P. D.: A Next Generation Ocean Carbon Isotope Model for Climate Studies I: Steady State Controls on Ocean 13 C, Global Biogeochemical Cycles, 35, e2020GB006757, https://doi.org/10.1029/2020GB006757, 2021.

Where I agree with the anonymous reviewer is that I think the new “reconstructed” dataset could be (I think): (1) expanded spatially using the GLODAP gridded product and (2) that this would be a very useful addition to our community. I am assuming these are “minor revisions” as the ML model is already built and I assume the application to the gridded product will be straightforward (and worth the time for the community to use!). I would also urge the authors to consider the other options listed by the anonymous reviewer to expand the ML methods temporally, although I am unfamiliar with the reviewer’s specific suggestions and cannot comment on the time requirements for such new applications.

Likewise, the other reviewer makes strong comments about the dataset itself. I agree that adding the reconstructed dataset as its own column (with -999 for other basins) to the existing GLODAP data would be very useful for the community. Even better would be for the community to have a gridded product!

R: Thank you for your valuable feedback and recognition of the community utility of our reconstructed dataset. We fully agree with your suggestions regarding spatial expansion and dataset compatibility.

Regarding your suggestion to expand spatially using the GLODAP gridded product, we have thoroughly checked the official GLODAP repository but have not found an official gridded version of the dataset. We fully acknowledge the value of a gridded $\delta^{13}C_{DIC}$ product for the community and would be pleased to supplement our reconstruction with a corresponding gridded product if an official GLODAP gridded product becomes available. However, given the current spatiotemporal sparsity of the underlying $\delta^{13}C_{DIC}$ observations, we cautiously note that direct gridding at this stage may introduce additional uncertainties, including over interpolation in data-sparse regions and potential misrepresentation of true biogeochemical variability. This is a key consideration for maintaining the scientific rigor of the product, as our priority is to provide a reliable dataset that reflects the actual constraints of available observations.

We also sincerely appreciate your suggestion to enhance compatibility with GLODAP. However, as GLODAPv2 provides separate, official datasets for individual ocean basins (e.g., Atlantic, Pacific, Indian, Arctic), we have retained our product’s focus on the Atlantic Ocean, which is consistent with this basin-specific framework, rather than adding the reconstructed $\delta^{13}C_{DIC}$ to the full global GLODAPv2.2023 dataset (with non-Atlantic basins set to -999). This approach avoids unnecessary redundancy, as users can already access GLODAP’s global or other basin datasets directly from the official repository and seamlessly merge them with our Atlantic product as needed. To ensure clarity and interoperability, we have updated the Zenodo archive with a detailed README file. This document explains the dataset structure, labels for new fields, and step-by-step guidance for merging our product with GLODAPv2’s global or basin-specific datasets. We have also clarified citation requirements in both the manuscript and Zenodo metadata, emphasizing that users should cite GLODAPv2 for native variables and our work for the reconstructed $\delta^{13}C_{DIC}$.

Regarding your suggestion to expand the Machine Learning method temporally, we fully recognize the value of a spatiotemporally continuous $\delta^{13}C_{DIC}$ product and view this as a critical next step. As noted in our response to the anonymous reviewer, the current study prioritizes addressing spatial sparsity, an urgent gap given the extreme paucity of $\delta^{13}C_{DIC}$ observations. Temporal extension requires robust constraints on seasonal/interannual variability, which are currently limited by uneven temporal coverage of existing observations (most concentrated in summer). To advance this, we plan to integrate long-term time-series data from programs like BIOS (Bermuda) and HOT (Hawaii) to calibrate the ML model for temporal dynamics, building on the validated spatial reconstruction framework. We will also more fully use the numerical model data to validate the future work.

Again, thank you for your constructive suggestions. Your suggestions have helped us refine the utility and transparency of our dataset, and we remain committed to enhancing its value for the community in future work.

Below I have listed notes I made on the manuscript as I read through it.

Line by line notes

27: need to define delta notation

R: We appreciate your comment pointing out the need to define delta notation. In the revised manuscript, we have supplemented this definition when $\delta^{13}C$ is first introduced: “The stable carbon isotope ratio, $\delta^{13}C$ (expressed via the standard delta notation: $\delta^{13}C=((^{13}C/^{12}C)_{sample}/(^{13}C/^{12}C)_{standard}-1)\times10^3$, with the international reference standard usually the Vienna Pee Dee Belemnite ([V]-PDB) fossil), has been widely applied as a tracer in marine carbon research, providing valuable insights into various processes within the oceanic carbon system.”

79+: I don’t see a need to shorten “Section” here

R: Thanks for your comment. We agree with your view that there is no need to abbreviate "Section" here. To align with your suggestion and enhance readability, we have revised the original text by restoring all abbreviated "Sect." to the full term "Section".

100: I like the previous paragraph

R: Thanks for your comment.

132: what exactly does “exhibit high internal consistency” mean? Are there statistics to support this statement?

R: Thanks for your comment. To clarify, this phrase aligns with the definition used in Becker et al. (2016) and refers to the quantifiable agreement between overlapping data points within the dataset, ensuring no contradictory or anomalous deviations that would compromise reliability. Its core lies in verifying the consistency of data within the dataset through quantitative calculations. The specific explanation and statistical support are as follows.

Here, "high internal consistency" refers to a high level of coordination and reliability among the various data points within the final dataset, with no significant contradictions or abnormal deviations. This consistency is not a subjective judgment but a conclusion drawn from quantitative calculations of the "offsets" at "crossovers" in the dataset, ensuring the dataset has logical stability internally.

The statistical calculation for this conclusion refers to the method proposed by Tanhua et al., 2010, with specific steps as follows: The "Weighted Mean (WM)" is used to quantify the internal consistency of the dataset. The weight is determined by the offset of each crossover and its standard deviation, which emphasizes the influence of more reliable data on the result.

$$WM = \frac{\sum_{i=1}^{L} D(i)/(\sigma(i))^2}{\sum_{i=1}^{L} 1/(\sigma(i))^2}$$

Parameter Definitions:

L: Represents the total number of crossovers in the dataset.

D(i): Refers to the respective offset of the i-th crossover (i.e., the numerical difference of different data at that crossover).

σ(i): Denotes the standard deviation of the offset of the i-th crossover, which is used to measure the degree of dispersion and reliability of that offset.

Becker, M., Andersen, N., Erlenkeuser, H., Humphreys, M. P., Tanhua, T., and Körtzinger, A.: An internally consistent dataset of δ13C-DIC in the North Atlantic Ocean – NAC13v1, Earth System Science Data 8: 559-570, https://doi.org/10.5194/essd-8-559-2016, 2016.

Tanhua, T., van Heuven, S., Key, R. M., Velo, A., Olsen, A., and Schirnick, C.: Quality control procedures and methods of the CARINA database, Earth Syst. Sci. Data 2: 35-49, https://doi.org/10.5194/essd-2-35-2010, 2010.

139: Is GPR an acronym? Perhaps not relevant, but I wanted to know

R: Thank you for your question. GPR is the acronym for Gaussian Process Regression, which first appears in the Introduction (Line 74).

161: Repeated text

R: Thanks for your comment. We have deleted these sentences and reorganized the sentences in Section 2.2 (Line 133-135) as: "After applying additional adjustments, the $\delta^{13}C_{DIC}$ data for the remaining 37 cruises exhibit high internal consistency. These 37 cruises do not include 13 cruises without deep-water crossover stations (Table 1) and cruise 64TR19900417, which were excluded to ensure data reliability as their uncertainties cannot be objectively quantified. Collectively, these excluded cruises accounted for less than 3 % of total $\delta^{13}C_{DIC}$ measurements."

Fig. 2: I like the figure, but as the other reviewer noted, it would be better to use completely independent cruise datasets for the validation as well as the "independent" tests

R: Thank you for your feedback. We would like to clarify that our existing validation framework design, which aligns with this core principle while balancing statistical robustness and practical feasibility for sparse oceanographic data.

As detailed in our response to the first reviewer, our independent test set was intentionally selected to be fully decoupled from the training/validation pool, with no overlap in cruises, spatial regions, or temporal coverage. This means measurements from any cruise are entirely confined to either the training/validation set or the independent test set, ensuring the final performance evaluation (reported RMSE and R^2) is based on completely unseen cruises, which directly addresses the need for independent cruise-based testing. The 10-fold cross-validation within the training set was solely for hyperparameter tuning, not for final performance assessment, so it does not compromise the independence of the test phase.

We maintained this design because many of the 51 cruises in our dataset have small sample sizes. Splitting the training/validation set by cruise would result in highly imbalanced folds, leading to unstable hyperparameter tuning and biased cross-validation results, undermining the statistical rigor of the model development process. By using random splitting within the training/validation pool, we preserve the natural spatiotemporal variability of $\delta^{13}C_{DIC}$ data, ensuring the model is tuned to generalize across diverse oceanic conditions rather than specific cruises. This approach is consistent with established practices in oceanographic ML studies (e.g., Lima et al., 2023; Regier et al., 2023; Wu et al., 2025), as cited in our response to the first reviewer.

To enhance clarity on the independence of the cruise datasets, we have revised the relevant paragraph to explicitly highlight that the independent test set comprises completely separate cruises from the training/validation pool: "The dataset was randomly split into a training set (80%) and a validation set (20%), with model training and hyperparameter tuning performed using 10-fold cross-validation within the training set to mitigate overfitting. An independent test set was reserved for final performance evaluation, selected to ensure no overlap with the training/validation set in cruises, spatial regions, or temporal coverage. We opted for random splitting over cruise-separated k-fold cross-validation to balance robustness and feasibility: many of the 51 cruises have small sample sizes, and cruise-separated splitting would cause imbalanced folds, leading to unstable hyperparameter tuning and biased results. Random splitting also preserves the natural spatiotemporal variability of $\delta^{13}C_{DIC}$, tuning the model to generalize across diverse oceanic conditions rather than specific cruises. This framework aligns with established practices for sparse oceanographic datasets (Lima et al., 2023; Regier et al., 2023; Wu et al., 2025).". This revision ensures the manuscript clearly conveys that our validation strategy incorporates fully independent cruise datasets for the critical final evaluation, while the training-phase cross-validation design prioritizes practical feasibility and stable model tuning.

Lima, I. D., Wang, Z. A., Cameron, L. P., Grabowski, J. H., & Rheuban, J. E.: Predicting Carbonate Chemistry on the Northwest Atlantic Shelf Using Neural Networks. Journal of Geophysical Research: Biogeosciences, 128(7), e2023JG007536. https://doi.org/10.1029/2023JG007536, 2023.

Regier, P., Duggan, M., Myers-Pigg, A., & Ward, N.: Effects of random forest modeling decisions on biogeochemical time series predictions. Limnology and Oceanography: Methods, 21(1), 40-52, https://doi.org/10.1002/lom3.10523, 2023.

Wu, Z., Lu, W., Roobaert, A., Song, L., Yan, X.-H., and Cai, W.-J.: A machine-learning reconstruction of sea surface p CO2 in the North American Atlantic Coastal Ocean Margin from 1993 to 2021, Earth Syst. Sci. Data, 17, 43–63, https://doi.org/10.5194/essd-17-43-2025, 2025.

192: I wonder if other Earth scientists would be as surprised to learn of Mean Absolute Error and Mean Bias Error. I think they might and it might therefore be useful to use a sentence or two describing why these additional metrics are useful to the study

R: Thank you for your suggestion. As recommended, we have supplemented the original sentence about model accuracy evaluation with 2-3 sentences (Line 205-209: "Among these metrics, MAE and MBE are valuable for evaluating the performance of the machine learning models. MAE quantifies the average absolute deviation between observed and predicted values; its insensitivity to outliers makes it ideal for handling the potential noise in $\delta^{13}C_{DIC}$ observational data, ensuring a robust measure of overall prediction error. MBE, by retaining the sign of deviations, identifies systematic biases (e.g., consistent overestimation or underestimation of $\delta^{13}C_{DIC}$), which is critical for refining the machine learning model.") explaining MAE and MBE, with a specific focus on their application in machine learning, to enhance the manuscript's clarity for the broader community.

202: Propagated error?

R: Yes, the total uncertainty of the reconstructed $\delta^{13}C_{DIC}$ is a propagated error. As detailed in the revised manuscript, we assumed independence between the three uncertainty sources (u_{obs}, u_{inputs}, u_{map}) and calculated the total uncertainty using the standard error propagation method (root-sum-of-squares synthesis), as supported by the cited references (Hughes and Hase, 2010; Taylor, 1997). We have refined the text to explicitly emphasize the error propagation approach and its implementation, ensuring clarity on this point.

Revised Text in Manuscript:

The comprehensive uncertainty of the reconstructed $\delta^{13}C_{DIC}$ was derived via error propagation, assuming independence between distinct uncertainty sources. These sources of uncertainties include: the direct $\delta^{13}C_{DIC}$ measurement uncertainty from observations (u_{obs}), the uncertainty accumulated from the input variables (u_{inputs}), and the uncertainty induced by the mapping function (u_{map}). Following standard error propagation protocols (Hughes and Hase, 2010; Taylor, 1997), the comprehensive uncertainty of our estimated $\delta^{13}C_{DIC}$ product, $u_{\delta^{13}C_{DIC}}$, was calculated as the root sum of the squares of the individual uncertainties:

$$u_{\delta^{13}C_{DIC}}=\sqrt{u_{obs}^2+u_{inputs}^2+{u_{map}}^2}$$

212: perturbed not perturbs

R: We are grateful for you noticing this typo. The error has been fixed in the corresponding section of the revised manuscript.

230: I'm unsure where the 10-fold cross-validation comes from

R: 10-fold cross-validation was selected based on its availability as a standard option in MATLAB's Machine Learning Toolbox, which is widely used for model training in our field (e.g., Wu et al., 2025), and its suitability for balancing computational efficiency and generalization performance with our dataset. We have added this clarification to the manuscript to enhance transparency (Line): "During the training phase, we leveraged a 10-fold cross-validation approach, selected as it is a standard pre-implemented option in MATLAB's Machine Learning Toolbox. This approach balances computational efficiency and robustness, reducing overfitting by iteratively splitting training data into 10 folds: 9 for training and 1 for validation per iteration, with results averaged across iterations to ensure stable performance. Finally, the model achieved an R^2 of 0.92, an RMSE of 0.083 ‰, an MAE of 0.056 ‰, and an MBE of −0.0003 ‰ (**Fig. 3a**)."

Wu, Z., Lu, W., Roobaert, A., Song, L., Yan, X.-H., and Cai, W.-J.: A machine-learning reconstruction of sea surface p CO2 in the North American Atlantic Coastal Ocean Margin from 1993 to 2021, Earth Syst. Sci. Data, 17, 43–63, https://doi.org/10.5194/essd-17-43-2025, 2025.

249: This text is also somewhat a repetition of earlier text

R: Thanks for your comment. We revised this sentence as: "To assess the product's ability to capture $\delta^{13}C_{DIC}$ spatial patterns and quantify biases, we utilized the $\delta^{13}C_{DIC}$ distribution from independent test cruises 33MW19930704 and 33RO20050111 (**Fig. 4**)."

259: larger?

R: Thank you for pointing out this typo. We have corrected it in the revised manuscript.

272: Incredibly / unbelievably low input variable uncertainty (Uinputs). I wonder if this is a propagation of the input variable uncertainties or an error has been made along the way.

R: Thank you for drawing attention to the unusually low u_{inputs}. Upon thorough rechecking, we confirm that the initially reported value stemmed from a computational error in the Monte Carlo simulation workflow. We have corrected this issue, and the revised u_{inputs} is 0.0087 ‰, with contributions decomposed as follows: temperature (4.96×10^{-5} ‰), salinity (3.62×10^{-4} ‰), nitrate (0.004 ‰), silicate (0.002 ‰), DIC (0.005 ‰), AOU (0.004 ‰), and xCO$_2$ (6.52×10^{-4} ‰). This revised value is consistent with the expected magnitude of input-related uncertainty for $\delta^{13}C_{DIC}$ prediction in marine biogeochemical studies, resolving the counterintuitive result noted in your comment.

The corrected u_{inputs} and detailed uncertainty decomposition have been updated in the manuscript to ensure transparency and accuracy.

295: Maybe this is not important, but lower case "n" is typically used to describe the sample size

R: Thank you for pointing out this notation consistency issue. As recommended, we have revised all instances of the uppercase "N" (previously used for sample size) to the lowercase "n" throughout the manuscript to align with academic norms.

302: Is it expected that there would be a model smoothing tendency?

R: Yes, the model's tendency to smooth extreme values is expected. This behavior is inherent to the Gaussian Process Regression (GPR) model and aligned with the study's goal of reconstructing a spatially continuous, reliable $\delta^{13}C_{DIC}$ product for the Atlantic Ocean. Below we elaborate on the key reasons: 1) GPR's intrinsic smoothing property: As a non-parametric model based on Gaussian kernel functions, GPR inherently weights the influence of neighboring data points to produce continuous predictions. This kernel-based mechanism naturally mitigates the impact of extreme values (which are often sparse in observational data) to avoid overfitting to isolated outliers or sampling noise. 2) Goal of spatial reconstruction: Our study aims to capture the large-scale, intrinsic spatial patterns of $\delta^{13}C_{DIC}$ rather than replicate rare local anomalies. Smoothing extreme values helps filter out noise from discrete observations and enhances the spatial consistency of the reconstructed product. Thus, the intrinsic regularization of the GPR leads to reduced sampling noise, a sharper central peak and narrower tails in the reconstructed KDE compared with the empirical KDE from observations.

We have supplemented the manuscript to clarify that this smoothing tendency is expected and its rationale, as detailed below: "Consequently, the reconstructed values display a slightly sharper central peak and narrower tails than the observations, indicating a tendency of the model to smooth extreme values, which is expected given the intrinsic properties of the GPR model and the study's objectives. Specifically, relying on Gaussian kernel functions, GPR naturally weights neighboring data points to produce continuous, spatially consistent predictions, which mitigates overfitting to sparse extreme values often linked to sampling noise or local transient perturbations."

397: Is there an expectation that the model output would closely align with the observed data? Wasn't the 2023 data used to predict the "reconstructed data"? I'm not diminishing the work—I honestly think this is an expected outcome of using machine learning.

R: Yes, the machine learning model outputs are expected to align with observed data when predictors are reliable. We would like to clarify that our workflow consists of two distinct, independent phases: model training/testing and prediction (detailed in Fig. 2). Specifically, during the training/testing phase, we utilized all available Atlantic cruise datasets containing $\delta^{13}C_{DIC}$ observations (including 2023 data along A16N) to train the model. We then validated and tested the model's fitting performance through rigorous procedures, ensuring its robustness in capturing the relationship between input variables and $\delta^{13}C_{DIC}$. In the subsequent prediction phase, we applied this pre-trained and validated model to the input variables from the GLODAPv2.2023 Atlantic dataset to generate the reconstructed $\delta^{13}C_{DIC}$ data. Importantly, this reconstructed $\delta^{13}C_{DIC}$ dataset is entirely independent of the original $\delta^{13}C_{DIC}$ observations in GLODAPv2.2023; they are two separate datasets.

The reconstructed $\delta^{13}C_{DIC}$ data for 1993, 2003, and 2013 mentioned in this paragraph all come from this GLODAPv2.2023-driven prediction. Due to the fact that the observational data along A16N in 2023 not included in the GLODAPv2.2023 dataset, we used the same pre-trained and validated model, relying solely on the 2023 observational input variables (e.g., T, S, nutrients) to produce the predicted values. This allows the model to independently predict 2023's $\delta^{13}C_{DIC}$ based solely on the spatiotemporal and environmental patterns it learned during training. This design confirms the alignment between the 2023 reconstructed and original observational data reflects the model's genuine predictive ability, rather than overfitting or circular reasoning.

To clarify, we revised this paragraph as: "Besides horizontal distributions, the reconstructed $\delta^{13}C_{DIC}$ dataset also provides valuable insights into vertical variability. The depth profiles along the North Atlantic A16N section in 1993, 2003, 2013, and 2023 (**Fig. 9**) show that the reconstruction substantially improves vertical resolution and continuity, especially for years with sparse measurements. For instance, the $\delta^{13}C_{DIC}$ samples were increased from 493 to 1,618 in 1993, 38 to 2,395 in 2003, and 473 to 2,787 in 2013, respectively, enhancing data coverage across depths and latitudes, facilitating the detection of temporal trends associated with ocean carbon uptake and redistribution (**Fig. 9**). The reconstructed $\delta^{13}C_{DIC}$ for 1993, 2003, and 2013 was generated by applying the pre-trained model to input variables from the GLODAPv2.2023 Atlantic dataset. Notably, as the observational data along A16N in 2023 not included in the GLODAPv2.2023 dataset, we used the same pre-trained and validated model, relying solely on the 2023 observational input variables to produce the reconstructed values. The close alignment between 2023's reconstructed and observed data (**Fig. 9d** vs. **9h**) not only reflects the model's reliability but also validates its ability to generalize, strengthening confidence in reconstructions for years with sparse measurements (e.g., 2003 with only 38 observations)."

485: quality-controlled (?)

R: Thank you for catching this typo. It has been corrected in the revised manuscript.