I commend the authors on their efforts in this revised draft. Generally, the paper has been improved. I previously raised several major objections, and now some of them have been addressed, though others have not:

We sincerely appreciate your detailed feedback and your recognition of the improvements we have made to the revised draft. We have addressed each of the remaining concerns thoroughly in the following responses with supplementary analyses, revised validations, and new product developments as outlined below.

Improper Validation: I am very pleased to see the authors have added a model-based OSSE. However, the implementation was unexpected and likely not as helpful as I hoped. The point of such a model reconstruction is usually to assess how well the approach works when applied in the same way it is being applied in the real world. Put another way, this should be a test of how well sparse data and machine learning can be used to estimate values where they are desired. Therefore, for this test, I would have expected the analysis to subsample the model at the locations and times where measurements are available (interpolating between data points and/or times as necessary), train an algorithm, and then use that algorithm to estimate the model values at the locations and times where the new dataset is reporting estimated values. Alternatively (or supplementally), a more general test could be done using the same first two steps but then reproducing the full (Atlantic) model distribution over time. Instead, this analysis seems to sub-sample the model in an idealized and homogeneous grid. This is really only a test of what could have been done had we implemented an entirely different, and much more expensive, data collection effort over the last few decades. I'm therefore not sure much is learned.

R: Thanks for your feedback. We apology for misinterpreted your original suggestion. We agree with your perspective that model-based OSSE validation should closely mirror the real-world application of our method, specifically to assess how effectively sparse observational data and machine learning can estimate values at target locations and times. To address this, we have now taken a validation approach to align with your core recommendation as much as we can. We now subsample the model data exclusively at the spatial locations of the collected cruises used in our training, validation, and test sets, as well as the GLODAP dataset employed for prediction. Given the model data only spans 1990–2002, we mapped observational data from other decades (e.g., 2000–2009 observations to 1990–1999 model data) to maintain a sample size comparable to the full observational dataset. This design exactly replicates the sparse, heterogeneous distribution of real oceanographic observations, eliminating the idealized grid-based subsampling you mentioned above.

The results from our revised training, validation, and test sets demonstrate strong performance (Figure R1): the training set achieves an R^2 of 0.999 and an RMSE of 0.009‰, the validation set shows an R^2 of 0.9985 and an RMSE of 0.012‰, and the test set yields an R^2 of 0.9950 and an RMSE of 0.025‰. These metrics confirm that the model can learn robustly from sparse, observation-like data and generalize effectively to unseen cruises.

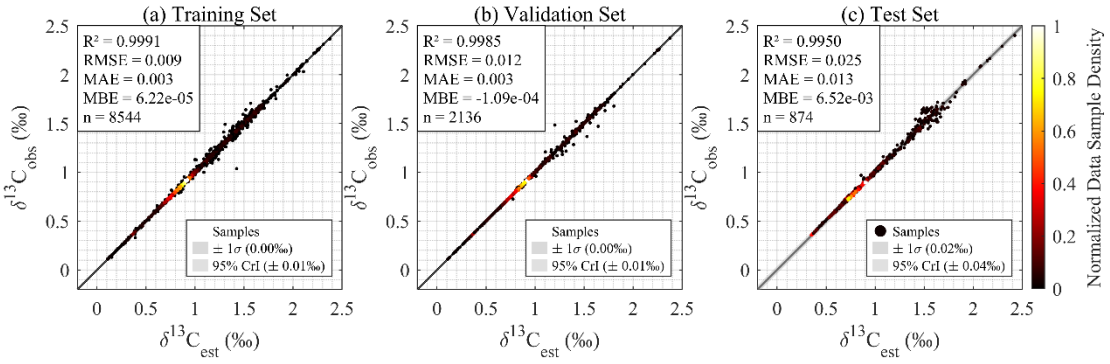

Figure R1. GPR model evaluation for numerical model $\delta^{13}C$ reconstruction: Density scatter plots comparing GPR model-estimated ($\delta^{13}C_{est}$) versus numerical model outputs $\delta^{13}C$ ($\delta^{13}C_{obs}$) values during (a) training, (b) validation, and (c) independent testing.

For the prediction phase, when estimating values at GLODAP locations, the reconstructed $\delta^{13}C$ values align closely with the model's values, as shown by an R^2 of 0.95 and an RMSE of 0.080‰, with high-density data points clustering tightly along the 1:1 line (Figure R2a). Some data points deviate from the 1:1 line, which is likely stemmed from inherent uncertainties and biases in the numerical model. Despite this, the density distribution plot confirms that our reconstructed values successfully reproduce the model distribution of $\delta^{13}C$, verifying the method's ability to capture large-scale spatial patterns from sparse observations. Additionally, the KDE analysis (Figure R2b) underscores the strong statistical congruence between the numerical model $\delta^{13}C$ values and reconstructed estimates, with the minor differences being negligible relative to the GPR model's success in replicating the core characteristics of the numerical model's native distribution. By subsampling the model at the actual locations of real observations, our revised OSSE now directly evaluates the method's performance in a real-world scenario, addressing your concern that the original idealized grid-based subsampling did not reflect practical data collection efforts. We believe these revised results provide meaningful evidence of the method's utility and robustness.

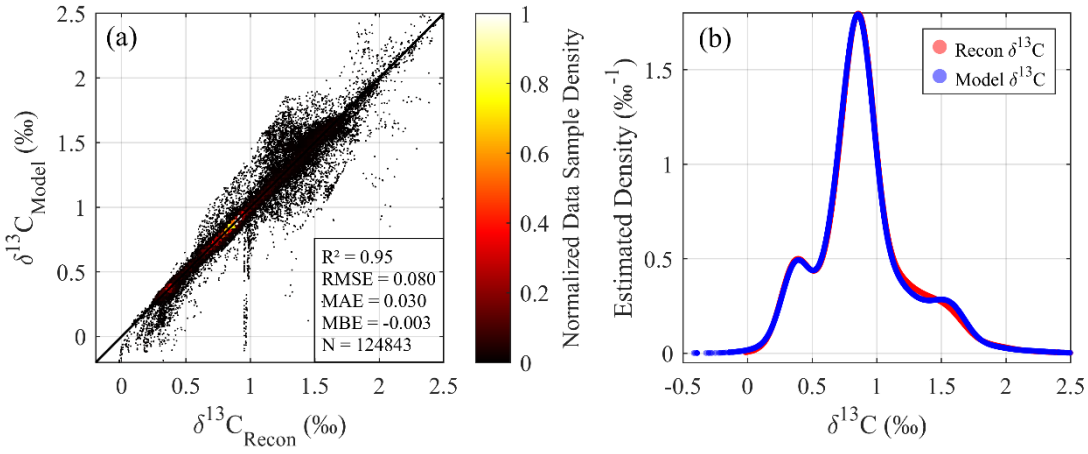

Figure R2. Comparison of model and reconstructed $\delta^{13}C$ values during prediction phase. (a) Density scatter plot of model versus reconstructed $\delta^{13}C$ values (n = 124,843). (b) Gaussian kernel density estimations (KDEs) for a comprehensive evaluation of model and reconstructed $\delta^{13}C$DIC values.

To further strengthen the integrity of our validation system and enhance the credibility of both

the GPR model and reconstructed data, we have supplemented the aforementioned content into the Appendix of the main manuscript. We believe these revised results and the supplementary Appendix validation provide meaningful and actionable evidence of the method's utility and robustness.

We have also attempted to address your supplementary suggestion to perform a more general test that reproduces the full Atlantic model distribution over time. Given the model's 1°×1° horizontal resolution, 50 vertical levels, and the resulting large sample size, which posed high computational memory demands, we applied a sparse sampling strategy to the gridded model data. We subsampled every other grid point in both the horizontal and vertical directions, and focused on the 1990–1999 period (120 months) to maintain computational feasibility. Using the same machine learning model trained on the observation-matched model data, we then predicted the gridded $\delta^{13}C$ values across the entire Atlantic domain and compared these predictions to the model's native "true" values. The results show good alignment (Figure R3), which demonstrates that the predictions align well with the model's native $\delta^{13}C$ distribution. The result achieved an R^2 of 0.8305, an RMSE of 0.193‰, an MAE of 0.074‰, and an MBE of -0.024‰ across 8,425,440 valid samples. High-density regions cluster tightly along the 1:1 line. The consistent large-scale distribution confirms our method generalizes effectively. It reproduces the full Atlantic $\delta^{13}C$ patterns over time. This test provides additional evidence that our approach leverages sparse observations to estimate large-scale oceanographic distributions. As limited by our computational capability, this part is preliminary and can be much improved in future work, we do not intend to add it to the Appendix.

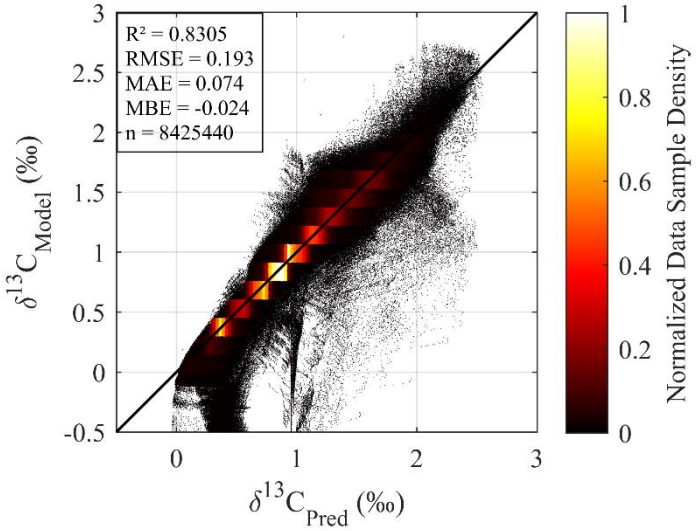

Figure R3. Density scatter plot of grid model versus grid reconstructed $\delta^{13}C$ values (n = 8,425,440) during prediction phase.

The k-fold validation experiment is also not attempted. However, the authors do point out something I'd missed previously, which is that the test data set is separated by cruise number. This is good news, though the statistics would still be still more trustworthy if they had used a k-fold validation for this step (note, there need not be the same number of cruises in each set... just an approximately comparable amount of training data with some number of whole cruises).

R: Thanks for your comment on k-fold cross-validation. Following your suggestion, we implemented 5-fold cruise-separated cross-validation for the training set (35 cruises in total,

excluding the independent test cruises 25 and 28). To address the small sample sizes of some cruises, cruises were allocated to folds using an iterative balancing procedure that minimizes disparities in total sample size across folds. The grouping results were: Fold 1: 2099 samples (6 cruises); Fold 2: 1959 samples (8 cruises); Fold 3: 2457 samples (6 cruises); Fold 4: 1836 samples (8 cruises); Fold 5: 2718 samples (7 cruises).

To avoid data leakage, each fold was standardized independently (mean and standard deviation calculated only from the training subset of the fold). The cross-validation yielded robust statistical results:

Per-fold validation RMSE: 0.121‰, 0.149‰, 0.099‰, 0.156‰, 0.097‰;

Average validation RMSE: 0.124‰ (±0.027‰, standard deviation);

Per-fold adjusted R^2: 0.749, 0.756, 0.848, 0.740, 0.916;

Average adjusted R^2: 0.802 (±0.077, standard deviation).

By grouping entire cruises and avoiding data leakage between folds, this validation fully accounts for the heterogeneity of observational conditions across different cruises. Although the cross-validation results (average adjusted $R^2 = 0.802 \pm 0.077$, average RMSE = 0.124 ± 0.027‰) are slightly less optimal than the independent test set, this is entirely reasonable. Each fold in the cross-validation includes cruises with diverse observational backgrounds, creating a more challenging and realistic test scenario that better reflects the GPR model's ability to adapt to unseen cruise data. These results confirm that the model is robust to inter-cruise variability and generalizes well to new observational datasets.

We have supplemented the above cruise-separated k-fold design details, grouping logic, and statistical results to the Method (paragraph 3 in Section 2.3) and Results (Section 3.1) sections of the manuscript. This supplementary validation significantly enhances the statistical trustworthiness of our model performance metrics as you requested.

Lack of Motivation: I'll begin here with an acknowledgement that at least one other reviewer did not seem to share my concern here.

That said, I don't believe the issues I raised about the motivation of the paper were well addressed. The authors provide some evidence from ongoing work, but their analysis actually demonstrates my reasons for concern better than I did. The ongoing work shows that eMLR faces significant problems when it is presented with reduced data density. However, eMLR is actually also an algorithm-based approach that exploits covariance between various measurements (much like the algorithm that they are employing in this manuscript). Therefore, it facing major problems when given fewer data is actually evidence that an algorithm cannot always effectively hide the limitations from sparse sampling. To make the point that the authors were trying to make, they would need to have a third set of panels where they have used an algorithm trained on the sub-sampled data to up-sample the cruise back to its original data distribution (this step is effectively what their paper is attempting to do)... and then use eMLR on that up-sampled data set to test whether the up-sampled data set can faithfully reproduce the canth estimates produced from the original data. If they were to do this, then it should go in the paper. However, this analysis leaves me worried that people using the up-sampled carbon-13 distribution will reach erroneous conclusions.

R: Thanks for your comments. We agree with your core concern about the need to verify the consistency between the algorithm-up-sampled data and the results derived from the original data.

We have supplemented a comprehensive four-scenario comparative analysis that not only includes the third set of panels you suggested but also adds an additional test (using up-sampled/reconstructed data for both 2013 and 2023 to estimate anthropogenic $\delta^{13}C$ changes) to further validate the product's utility for covariance-dependent analyses. The results (Figure R4) verify the effectiveness of our reconstruction algorithm in alleviating the limitations of sparse sampling, as they directly implement the verification process you proposed.

Figure R4a serves as the benchmark (full original 2013/2023 $\delta^{13}C$ data), reflecting the true $\Delta\delta^{13}C_{anth}$ characteristics without sampling limitations. Figure R4b simulates sparse sampling (2023 stations restricted to 2013 $\delta^{13}C$ latitudes), and its obvious deviation from Figure R4a. This aligns with your point that algorithms face limitations under sparse data, and directly confirms the necessity of our reconstruction work. Critical to your suggestion, Figure R4c implements the verification process you proposed: we used the reconstructed $\delta^{13}C$ data in 2023 combined with original 2013 data for $\Delta\delta^{13}C_{anth}$ estimation via eMLR. The results here are highly consistent with the Figure R4a, proving that our up-sampled data restores the original $\delta^{13}C$ distribution well enough to make eMLR reproduce full-data results, effectively alleviating sparse sampling limitations.

We further added Figure R4d (full reconstructed 2013/2023 $\delta^{13}C$ data) to test the algorithm's stability. Its strong consistency with Figure R4a demonstrates that our algorithm reliably completes sparse data across cruises, and the reconstructed data's covariance structure matches the original, supporting its suitability for covariance-dependent analyses like eMLR. Together, these results address your concern that up-sampled data might lead to erroneous conclusions, as the reconstructed product's consistency with original data is quantitatively verified.

This aforementioned result represents only a preliminary verification, as our research on anthropogenic carbon using the eMLR method based on $\delta^{13}C$ remains ongoing. The present analysis demonstrates that our reconstructed data can effectively capture signals of $\Delta\delta^{13}C_{anth}$ without producing erroneous conclusions. These preliminary findings have not been included in the manuscript, as they do not constitute our final research outcome and the primary focus of this manuscript focuses on the reconstructed $\delta^{13}C$ data itself. The complete results of this investigation will be presented in our future work.

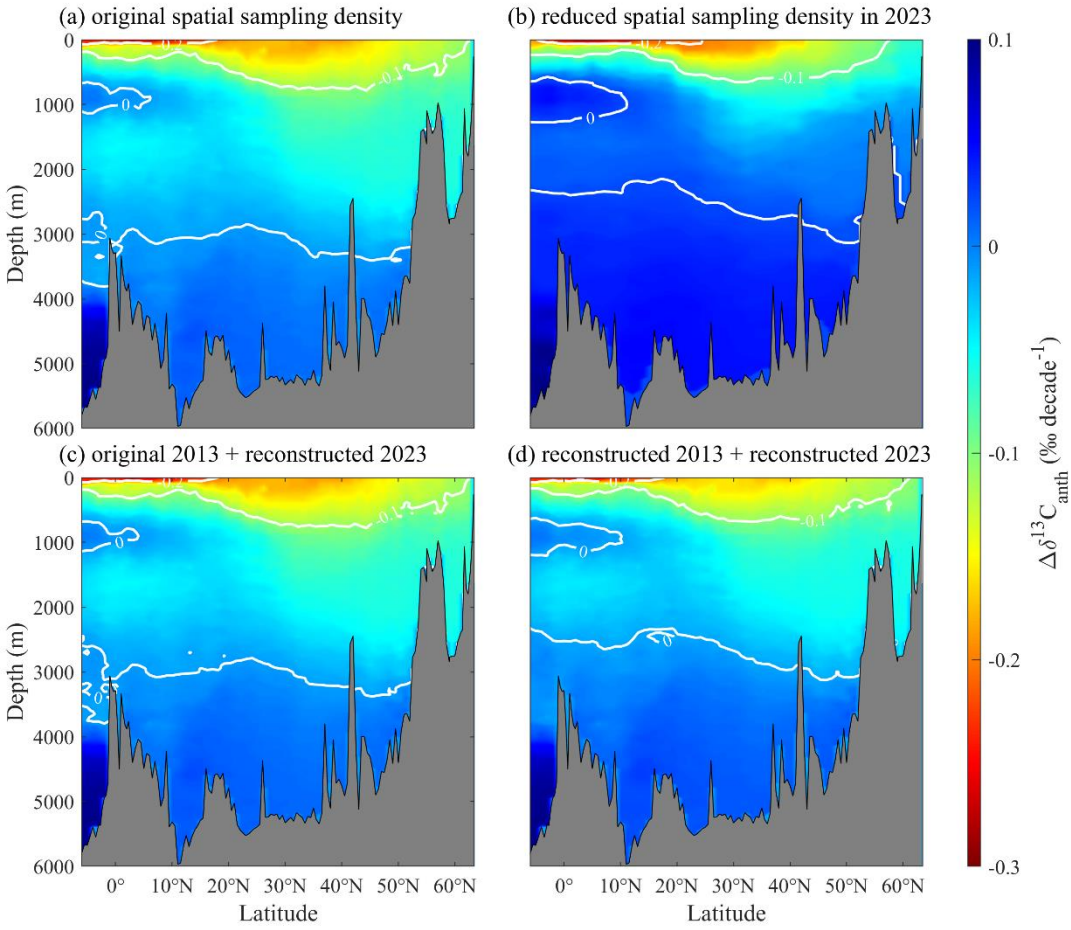

Figure R4. Anthropogenic $\delta^{13}C_{DIC}$ changes ($\Delta\delta^{13}C_{anth}$) along A16N between 2013 and 2023. (a) original spatial sampling density (full 2013/2023 $\delta^{13}C$ data), (b) reduced spatial sampling density in 2023 (2023 stations matching 2013 $\delta^{13}C$ latitudes), (c) original 2013 & reconstructed 2023 $\delta^{13}C$ data, and (d) reconstructed 2013 & reconstructed 2023 $\delta^{13}C$ data.

Errors in the measurement uncertainty sensitivity: my concerns in this area were addressed, thank you.

Re: GLODAP gridded product. I was suggesting using the authors' algorithm with the gridded T and S from GLODAP to produce a gridded field of d13C, not that GLODAP had themselves produced a gridded d13C product. I still think that this would be an easy and useful application for their approach.

R: Thank you for clarifying your suggestion regarding the GLODAP gridded product.

We accessed the climatological mapped GLODAP gridded product (Lauvset et al., 2016, Earth Syst. Sci. Data, 8, 325–340, https://doi.org/10.5194/essd-8-325-2016), which provides global 1°×1° gridded data for key environmental parameters. Following your suggestion, we utilized the gridded temperature, salinity, oxygen (used to calculate AOU), nitrate, silicate, and TCO_2 from this product, combined with the 1972–2013 atmospheric mean xCO_2, to reconstruct a climatological gridded $\delta^{13}C$ dataset for the Atlantic Ocean via our GPR algorithm.

This new gridded $\delta^{13}C$ product maintains full interoperability with GLODAPv2, adopting the

same spatial resolution (1°×1°), vertical depth layers (33 standard depths), and NetCDF format as the original GLODAP gridded data. Users can now seamlessly integrate the $\delta^{13}C$ grid with GLODAP's existing grids for basin-scale analyses. We have uploaded this climatological gridded $\delta^{13}C$ dataset to the Zenodo repository, alongside our existing datasets (https://doi.org/10.5281/zenodo.18481145). The updated Zenodo description clearly outline the product's origin, parameters, and compatibility with GLODAPv2, ensuring ease of use for the research community.

In summary, I'm not sure I would trust the new product enough to rely on it. Currently, if I were to want to create a 4D product, I would first go back to the original data and train an algorithm from those measurements rather than trying to chain algorithm estimates together. Generally, sparse data are not useful until they are brought into some kind of analysis, and I cannot think of an analysis that would benefit from starting out with sparse, fixed empirically estimated values. This would only be a way to give users a false sense of confidence when the estimated data confirm the patterns found in the real data. This therefore leaves me still doubtful about the practical utility of this work.

R: We appreciate your feedback on the rigor of this work and the product's practical utility. To fully address your core concerns, we emphasize two empirically verified improvements that collectively validate the model's reliability and the product's practical value:

1) The revised OSSE validation (Figures R1-R3) strictly adopts the same data processing pipeline as the manuscript's observational analysis, subsampling numerical model data at the exact spatial locations of real Atlantic cruises, preserving the sparse, heterogeneous sampling characteristics of oceanographic observations. This directly verifies that our GPR model can robustly extract large-scale biogeochemical signals from sparse data, laying the foundation for the product's reliability.

2) The four-scenario $\Delta\delta^{13}C_{anth}$ verification (Figure R4) proves that our reconstructed data can faithfully reproduce anthropogenic carbon estimates from original dense data, validating its suitability for covariance-dependent analyses without leading to erroneous conclusions.

Together, these improvements address your concerns about false confidence and practical utility. We have revised the manuscript to highlight these interconnected validations, ensuring clarity on the product's unique value for the research community. We feel our intellectual exchanges during this review process greatly improved the rigor and impact of this work. We truly appreciate your comments and suggestions.

(Separately, I'd urge the authors to write shorter and more focused responses to reviewer comments.)