Responses to Reviewer 2 (reviewer comments in grey; our response in blue)

This manuscript presents ResNet-DIC, a machine-learning reconstruction of global interior-ocean DIC extending to nearly full ocean depth. I think the development of a full-depth, observation-based DIC reconstruction is timely and potentially valuable. In particular, the attempt to use discrete GLODAP observations rather than mapped biogeochemical climatologies is interesting.

However, I have several substantial concerns regarding the choice of predictors, the validation strategy, and the interpretation of changes in total DIC as changes in the ocean carbon sink. I also found the methods difficult to follow in their current organization. I think these issues need to be addressed before the robustness of the reconstructed temporal variability can be assessed.

We thank the reviewer for their thorough and constructive evaluation of our manuscript. We appreciate their assessment of the study and the helpful suggestions, which will allow us to improve this work.

Major comments

In my opinion, the methods need substantial restructuring. The current methods mix descriptions of the input datasets, neural-network architecture, justification relative to previous products, predictor-selection experiments, and validation results. This makes it difficult to determine exactly what data enter the reconstruction and how the final product is generated.

We agree that the organisation of the Methods made it unnecessarily difficult to follow the workflow and identify precisely how the final reconstruction was generated. In response, we will reorganise and streamline the Methods section to provide a clearer and more logical description of the methodology. In particular, we will aim to separate the descriptions of the input datasets and predictors, the ResNet-DIC model architecture and training procedure, the predictor-selection experiments, and the validation strategy. We will also review the section to remove or relocate material that interrupted the methodological narrative. Furthermore, we will rework figure 1b) so that it clarifies what data are used during each step.

At present, Section 2.1 moves from ResNet to SOM-FFN and back to ResNet, while Section 2.2 contains substantial discussion of MOBO-DIC and already begins presenting predictor-performance results. The reader mainly needs to know what was done and why. Comparison with previous reconstruction approaches could be shortened substantially here and moved to the Discussion.

We agree that the original organisation could be improved. We will therefore streamline Section 2.1 to focus more directly on the ResNet-DIC methodology and move much of the broader comparison with previous reconstruction approaches to the Discussion.

Nevertheless, we plan to retain a concise description of SOM-FFN. It provides important context for the development of machine-learning-based global pCO_2 mapping and is also used in MOBO-DIC for DIC mapping. We believe this context is useful for understanding the methodological differences between MOBO-DIC and ResNet-DIC, particularly that ResNet-DIC does not rely on the SOM step, which has been an important component of previous mapping approaches. We will also clarify this in the text to make this connection clearer while keeping the focus on the methods used in the present study.

This would also help clarify Figure 1b. At present it is not obvious that temperature and salinity both constitute predictor information and define the EN4 mapping grid. The manuscript states that the trained model is subsequently mapped onto EN4 temperature and salinity fields, but the schematic visually separates these concepts in a confusing way.

We agree and will revise Figure 1b) to make it much clearer.

I am not yet convinced that temperature and salinity are sufficient predictors for reconstructing temporal changes in total DIC. Temperature and salinity contain substantial information about water masses, circulation, mixing, and preformed carbon, but remineralization and other biogeochemical processes strongly affect interior DIC. I am therefore concerned that a reconstruction based primarily on temperature, salinity, location, depth, and atmospheric CO2 may reproduce the large-scale distribution of DIC (likely preformed increasing with CO2?) while failing to reproduce the processes responsible for its temporal evolution.

We agree that while temperature, salinity, depth, location and atmospheric CO_2 contain substantial information about water masses, their origins, circulation and ventilation history, they cannot be expected to capture all sources of DIC variability. In particular, we acknowledge that a reconstruction based primarily on these predictors may reproduce the large-scale DIC distribution more reliably than the temporal variability associated with biogeochemical processes.

Nevertheless, these variables also contain indirect information about biogeochemical processes through their relationships with water-mass properties, mixing and ventilation. A machine-learning model can exploit these statistical relationships to reconstruct DIC variability, even though it does not explicitly represent the underlying physical and biogeochemical processes. The choice of predictors therefore involves a trade-off between their information content and their observational coverage. Ultimately, the suitability of a given predictor set must be assessed by how well the resulting model reconstructs independent, held-out DIC observations.

Our objective is to establish the extent to which DIC variability can be reconstructed from comparatively well-covered hydrographic observations. This is particularly relevant given that several additional biogeochemical variables have more than an order of magnitude fewer observations. We believe there is considerable value in first establishing a robust reconstruction based on the more extensive hydrographic dataset, while recognising its limitations in capturing biogeochemical contributions to temporal variability. We will clarify the rationale for our predictor selection and the associated limitations in the revised manuscript.

The authors already show that adding oxygen reduces GLODAP RMSE from approximately 7.9 to 7.0 µmol kg-1, and adding oxygen, nitrate, and silicate reduces it further to approximately 5.2 µmol kg-1. I do not think the relatively modest difference in global RMSE is sufficient to demonstrate that these variables are unnecessary.

We agree that the improvement in GLODAP RMSE when including additional biogeochemical predictors demonstrates that these variables contain valuable information for reconstructing DIC. However, our intention was not to suggest that oxygen, nitrate or silicate are unnecessary, but rather to highlight the trade-off between predictive skill and observational coverage. Although oxygen and nutrient measurements are often available alongside DIC observations, their overall spatial and temporal coverage is substantially more limited than that of temperature and salinity. This is particularly important for the mapping phase, which requires sufficiently complete predictor fields to produce a global, full-depth reconstruction. Our approach therefore prioritises the more extensive temperature and salinity fields available from EN4, while recognising that this comes at the cost of excluding potentially valuable biogeochemical information. We discuss this limitation further in our response to the following comment.

The relevant test is whether the scientific conclusions about DIC variability and trends are robust to predictor choice. I strongly encourage the authors to train at least an alternative T/S/O2 model and compare, for example: vertical or basin-scale DIC sections; spatial patterns of DIC change; trends at HOT/BATS or other long-term observational locations; basin and depth-resolved inventory changes; and ideally, the reconstructed temporal evolution shown in Figures 11 and 13. If the T/S-only and T/S/O2 reconstructions produce similar temporal and spatial changes, that would provide much stronger support for the chosen predictor set. If they do not, then the interpretation of ResNet-DIC as a reconstruction of total DIC variability, rather than predominantly the component associated with hydrographic/water-mass changes, needs to be reconsidered. This concern is particularly important because the headline result of the manuscript is not simply reconstruction accuracy, but decadal variability in DIC storage.

We agree that further evaluation of the robustness of the reconstructed DIC variability to predictor choice would be useful. We will therefore extend our evaluation using

complementary sources of observational data and an alternative predictor configuration of temperature, salinity and oxygen. First, we will deepen our comparison of the reconstructed temporal variability with long-term observations from the HOT and BATS time-series stations, complemented by other suitable observations from the Global CO_2 Time-Series and Moorings Project.

Second, we will produce an alternative global reconstruction using temperature, salinity and oxygen as predictors, with the global oxygen field obtained from GOBAI-O_2 (Sharp et al. 2023). We will compare this reconstruction with our existing T/S-based reconstruction, focusing particularly on basin-scale and depth-resolved integrated DIC trends, including by recreating the Hovmöller and carbon inventory plots of Figures 11 and 13 with the new reconstruction.

Nevertheless, the caveat to this approach is whether the available global oxygen product is suitable for the global mapping step. While products like GOBAI-O_2 incorporate direct oxygen observations from GLODAP and BGC-Argo, its algorithms also rely on hydrographic information, including temperature and salinity, which are already used as predictors in our approach. Thus, the oxygen field is not independent of the information used in our reconstruction. Furthermore, GOBAI-O_2 is based on a relatively sparse observational dataset and is limited to the upper 2000 m. Finally, using a machine-learning reconstruction as an additional predictor introduces further uncertainties associated with the reconstruction itself, which will propagate from one machine-learning model into another. Consequently, differences between the T/S/O_2 and T/S reconstructions cannot solely be attributed to the additional information provided by oxygen, as they will also reflect uncertainties in the oxygen reconstruction used as input. It will therefore be important to also evaluate the resulting mapped product directly against independent observations.

To address this, we will further strengthen the evaluation using our main independent validation dataset: the GLODAP observations withheld from training for each ensemble member. In addition to evaluating model predictions at the original observation locations using the corresponding observed predictor values, we will evaluate both mapped DIC products directly against these observations. Specifically, we will collocate each withheld GLODAP observation with the nearest grid point in each global reconstruction at the corresponding time and depth – both the T/S/O_2 and T/S reconstructions. This provides a better assessment of the complete mapping procedure, including the effects of using the gridded predictor fields (EN4 temperature and salinity, and GOBAI-O_2 for the alternative reconstruction), rather than evaluating only the neural network predictions under the observed conditions at the validation locations.

Also, to me, the observational validation needs to be made more independent and more stringent. The manuscript withholds complete vertical profiles rather than individual measurements, which is preferable to a point-wise random split. However, profiles from the same cruises and nearby regions/times can still be

highly correlated. The authors themselves show that cruise-based withholding results in poorer predictive performance on the observational data. So for a global reconstruction intended to infer temporal variability, I think withholding entire cruises is a more appropriate test of observational generalization than randomly withholding profiles. I therefore find the statement that profile-based splitting is adopted because the objective is “global reconstruction rather than prediction for entirely unseen campaigns” unconvincing: predicting regions/times that are not directly observed is precisely the purpose of the reconstruction.

I would also strongly recommend that cruise-based validation be treated as a primary validation rather than an appendix sensitivity analysis.

We agree that cruise-based validation provides a more stringent assessment of the model's ability to reconstruct DIC in regions and periods not represented by the withheld cruises. We will therefore give greater prominence to these results and make both validation approaches available, allowing readers to assess reconstruction performance under the different validation conditions.

Our rationale for using profile-based withholding for the main validation was that it provides a more even distribution of withheld observations across the available cruises, regions and time periods, allowing the model to be trained on a more geographically and temporally representative dataset. Cruise-based withholding provides a more challenging validation test, as withholding entire cruises can leave larger gaps in the observational coverage of the validation dataset. This is reflected in the lower predictive performance obtained with cruise-based withholding. However, Figure A7 shows that reconstruction skill for the global ocean, albeit within the less complex model domain, remains similar between the cruise- and profile-based approaches. This supports our decision to retain profile-based withholding.

It would also be useful to show the spatiotemporal distribution of the held-out observations. A random percentage alone does not establish that the product is tested adequately across basins, depths, and decades.

The reconstruction is based on a 15-member ensemble, with each member using a different random split of the original GLODAP dataset into training and validation observations. Consequently, each ensemble member has a different distribution of withheld observations, rather than there being one fixed validation subset. The purpose of this approach is to assess model performance across multiple random partitions of the available observations, reducing the dependence of the validation results on any single random split. As the holdouts are randomly selected from the original GLODAP dataset, their combined spatiotemporal distribution across the ensemble is expected to reflect that of the original dataset. We therefore do not rely on the distribution of any individual holdout. In response to Reviewer 1, the rationale for the number of ensemble members will be further clarified in the manuscript.

The ECCO-Darwin and NorESM subsampling experiments are valuable and I would retain them, but they should not be described as equivalent to independent validation against observations. A neural network trained and evaluated within a model-generated ocean tests the consequences of sampling a known internally consistent system; it cannot establish that the corresponding relationships hold in the real ocean.This also makes the independent HOT, BATS, and Drake Passage comparisons therefore particularly important, but three locations cannot provide global observational validation.

We agree that the ECCO-Darwin and NorESM subsampling experiments cannot replace independent validation against observations. However, this was not their intended purpose. As described in the manuscript, these experiments are designed to investigate the effects of incomplete observational sampling on reconstruction performance within a known ocean system. We will clarify this distinction in the revised manuscript to avoid any ambiguity regarding their purpose and interpretation.

We would also like to clarify that our observational validation is not limited to the comparisons at HOT, BATS and Drake Passage. The main validation framework uses withheld GLODAP observations to independently evaluate model performance across the global ocean. We have tested several configurations, including multiple random profile-based splits as well as approaches withholding entire cruises. This allows us to assess model performance under different observational withholding strategies and the sensitivity of the results to these choices.

The comparisons at HOT, BATS and Drake Passage provide additional assessments against independent observational records at locations with long-term or repeated measurements. In the revised manuscript, we will further expand these comparisons by including additional time series from the Global CO_2 Time-Series and Moorings Project, and by further validating the reconstructions against the independent data as described above.

The distinction between total-DIC inventory changes and anthropogenic carbon uptake needs to be much sharper. This issue becomes increasingly important in Sections 4.2–4.3 and the conclusions. The manuscript correctly acknowledges that changes in total DIC are not equivalent to anthropogenic carbon uptake, because natural-carbon variability can contribute. However, the terminology subsequently shifts repeatedly toward “uptake,” “carbon sink,” and comparison with anthropogenic carbon-sink estimates. For example, the authors obtain a mean DIC accumulation of 2.64 Pg C yr-1 and compare this with an anthropogenic carbon sink estimate of 2.7 Pg C yr-1. Later, the reconstructed 2015–2020 accumulation reaches approximately 5.9 Pg C yr-1, substantially exceeding contemporary estimates of anthropogenic uptake. The paper interprets much of this variability in terms of weakening and strengthening of the ocean carbon sink. I think the manuscript needs to distinguish consistently between change in total ocean DIC inventory;

anthropogenic carbon accumulation; and net air–sea CO2 uptake. Agreement between their long-term magnitudes is interesting, but they are not interchangeable quantities. This is especially important on the five-year timescales emphasized here, where redistribution and natural carbon-cycle variability may matter. So I would therefore temper statements that ResNet-DIC provides an “independent estimate of the ocean carbon sink” unless the authors can demonstrate that the reconstructed total-DIC changes can reliably be interpreted that way.

We agree that we have used terms such as "uptake" and "carbon sink" too loosely, particularly when discussing the reconstructed inventory changes and comparing them with estimates of the anthropogenic ocean carbon sink. We will therefore revise the terminology throughout the manuscript and the conclusions to distinguish more clearly between changes in total DIC inventory and anthropogenic carbon uptake. We will also clarify that the comparison between our long-term mean DIC accumulation and published estimates of anthropogenic carbon uptake is intended as a comparison of magnitudes, rather than an equivalence between the two quantities.

We believe that reconstructing the total DIC pool remains scientifically valuable in its own right. By providing a full-depth reconstruction, ResNet-DIC offers a basis for further investigations into the respective contributions of anthropogenic carbon accumulation and natural carbon-cycle variability to changes in ocean DIC storage. We will revise the manuscript to make this distinction and the scope of the present reconstruction clearer.

Finally, the robustness of the temporal variability (the central result) needs stronger demonstration. The most interesting result is the slowdown around the early 2000s followed by a strong acceleration, yet this is also where several methodological concerns converge.The manuscript acknowledges temporal edge effects, sparse observations at the beginning and end of the reconstruction, increased ensemble spread after approximately 2016, and a sharp decline in available data toward the end of the period. The inferred 2015–2020 accumulation rate is also extremely large relative to established estimates. I think the paper would benefit from explicit sensitivity analyses demonstrating that these temporal features are not artifacts of changing GLODAP sampling density or changes in EN4 observational coverage; also the use of atmospheric CO2 as the temporal predictor or predictor selection; and temporal extrapolation near the ends of the training period.

Eg. a useful experiment could be pseudo-hindcasting: truncate the training observations at an earlier year and test whether the model reconstructs subsequently observed DIC evolution. The existing model-subsampling experiments are informative, but I would like to see stronger observational evidence before attributing the reconstructed peaks to real changes in the ocean carbon sink.

We agree that the robustness of the reconstructed temporal variability deserves further consideration. However, we would like to clarify that ResNet-DIC is not a time-series model in the conventional sense. Rather, it reconstructs DIC from a combination of

spatial and environmental predictors, with atmospheric CO_2 providing an additional temporal constraint. As demonstrated by the Shapley values (Fig. A14), atmospheric CO_2 has a relatively small influence on the model predictions compared with the other predictors. Thus, the reconstruction does not rely primarily on temporal prediction, but on the relationships between DIC and the combination of all available predictors.

Restricting the training period, as proposed in a pseudo-hindcasting experiment, would also reduce the number of available observations and the range of environmental conditions represented in the training dataset. Consequently, any reduction in reconstruction performance could reflect not only the model's ability to predict DIC beyond the training period, but also the reduced size and coverage of the training data. This makes it difficult to isolate data availability across time as a factor in such an experiment.

Nevertheless, uneven temporal sampling may influence the reconstruction if the available training observations do not adequately represent the range of environmental conditions and predictor combinations occurring during a given period. To assess these potential effects, we will strengthen the observational evaluation by predicting both the withheld GLODAP observations and time series datasets like HOT or BATS as a function of time, examining changes in bias and RMSE alongside the number and geographical distribution of validation observations, and the availability of both GLODAP and EN4 data. In particular, we will investigate whether reconstruction errors increase towards the beginning and end of the reconstruction period, where temporal edge effects and reduced observational coverage may leave the reconstruction less well constrained. These analyses will help assess the robustness of the reconstructed temporal variability, particularly towards the temporal boundaries.

Minor and specific comments

Abstract: “An independent method to study the uptake uses interior observations to calculate how dissolved inorganic carbon (DIC) is changing.” This is awkward and non-specific. Perhaps: “An alternative, observation-based approach is to infer carbon storage changes from changes in interior-ocean dissolved inorganic carbon (DIC).”

We agree and will adopt the suggested phrasing. Thank you.

Abstract: “global, full-depth reconstructions … have not been produced” should be nuanced. There are existing interior carbon/DIC reconstruction efforts, including MOBO-DIC and TRACEv1, although with different depth ranges, target variables, and methodologies. The novelty should be stated precisely rather than implying that no comparable products exist.

We agree and will revise the wording in the Abstract to better acknowledge existing efforts, including MOBO-DIC and TRACEv1 and to avoid implying that no previous global interior DIC or carbon reconstructions exist.

Abstract/throughout: Is "ResNet-DIC" the name of the machine-learning architecture/method, or the resulting reconstruction/product? Please use the term consistently.

We will revise the manuscript to use ResNet for the architecture and ResNet-DIC for the resulting product throughout.

L26: If the diagnosed inventory is similar to previous estimates, I would state this explicitly rather than simply giving the value.

We will clarify that the inventory is comparable to previous estimates.

L42 onward: The p in pCO2 should be italicized consistently throughout.

Thanks. This will be corrected throughout the manuscript.

L56: "If the variability it is real" → delete "it".

We will correct this.

Around L65–75: I suggest first acknowledging existing interior-DIC products and then clearly explaining what limitation the present work addresses. I also do not understand why the use of remote-sensing information by previous studies is presented as a disadvantage in itself. Surface observations could provide useful information about processes propagating into the interior; indeed, later in the manuscript the authors acknowledge that satellite information could improve near-surface predictions.

We agree and will revise the paragraph to better distinguish the contributions and limitations of previous products. We will also remove the implication that using remote-sensing information is inherently disadvantageous.

End of Introduction: The final two paragraphs still read largely as background/methodological justification. I suggest replacing them with a more conventional "Here we…" paragraph stating the objectives, approach, and principal questions addressed.

We agree and will revise the final paragraphs of the Introduction to provide a clearer statement of the study objectives, approach and main research questions.

L95: This appears inconsistent with the earlier suggestion that full-depth/global DIC reconstruction products do not exist. Please sharpen the distinction among total DIC, anthropogenic carbon, depth coverage, and geographic coverage.

We will revise the text to clarify how ResNet-DIC differs from previous products in terms of the carbon variable reconstructed and its depth and spatial coverage.

L102: A ResNet is a neural network; specifically a deep neural network with residual/skip connections. The wording should make this explicit rather than presenting the two almost as separate model types. The architecture in Figure 1a indeed looks like a neural network because it is one.

Thank you for this comment. We did not intend to distinguish ResNet from neural networks, but rather from a conventional feed-forward neural network (FFN), such as that used in SOM-FFN. We will revise our wording to make this clear.

L111: Replace "recent version" with the explicit version number, GLODAPv2.2023.

We will replace this.

L117: The reconstruction begins in 1990. Please discuss the implications of the history of certified reference material (CRM) use for the comparability/quality of early carbonate-system observations. Is 1990 an appropriate starting date for DIC training data, or would a later start produce a more homogeneous observational constraint? A sensitivity test could be useful.

We thank the reviewer for highlighting this issue. We will add information on the development of certified reference materials (CRMs) for seawater DIC measurements and the implications for the comparability of early observations. Our reconstruction relies on the quality control applied in GLODAP, which includes adjustments to improve consistency between cruises and results in relatively few observations from before 1990 being retained. We will additionally perform a sensitivity test using a later starting date for the training data to assess the influence on the reconstruction.

Figure 1b: I found this confusing. Temperature and salinity are predictor variables but are also supplied through the EN4 mapping grid. The diagram should distinguish clearly between variables in the GLODAP training dataset and gridded predictor fields used for prediction.

We agree that this figure needs reworking to make clearer what variables are used where. We will include a new version in the revised manuscript.

L140–145: The discussion that salinity affects alkalinity does not establish that salinity can represent temporal variations in DIC due to biological processes. Please avoid implying a mechanistic adequacy that has not been demonstrated.

We agree and will revise the text to avoid implying that salinity can represent DIC variations driven by biological processes.

L146: Which atmospheric CO2 product/quantity is used exactly? The Methods later indicate deseasonalized Mauna Loa atmospheric CO2. Please specify whether this is mole fraction (xCO2) or atmospheric partial pressure and use the terminology consistently.

We will clarify that the predictor is monthly deseasonalised atmospheric CO_2 mole fraction (xCO_2) from Mauna Loa and will make the terminology consistent throughout the manuscript.

Section 2.2: I would separate methodological choices from results. Figure 2 and the RMSE/bias comparison are effectively a predictor-selection experiment and could be presented in a dedicated subsection, rather than interwoven with the description of predictor variables.

We agree and will move the predictor-selection experiment, including Figure 2 and the associated RMSE and bias comparisons, into a separate subsection. This separates the description of the input predictors from the evaluation of their contribution to model performance.

Around L173: "Compared to GLODAP" is unclear. Presumably this means predicted DIC versus withheld GLODAP DIC observations. Please specify.

We will clarify that this refers to comparisons between predicted DIC and withheld GLODAP DIC observations.

Figure 2 / predictor testing: Because the models including oxygen/nutrients use a different subset of observations, please explain how the comparisons are made fairly. Ideally, predictor sets should be evaluated on the same observations; otherwise differences in RMSE may partly reflect differences in sample composition.

We agree that the differences in observational coverage need to be considered when interpreting these comparisons. Although adding oxygen and nutrients improves predictive performance, these variables have lower data density than hydrographic variables. This presents two related challenges: fewer observations may be available for model training, and the limited spatial and temporal coverage of these variables makes it more difficult to apply the trained model consistently across the global ocean. We will further clarify this trade-off in the manuscript and will additionally compare predictor sets using the same subset of DIC observations to isolate the effect of predictor choice from differences in training data availability.

L205: Explain the data partitioning more explicitly rather than simply referring to a "15% train/test split." In machine-learning terminology, the manuscript actually

appears to use training, validation, and a further hold-out test set, so the terminology should be made consistent.

We will clarify the terminology throughout the Methods section.

Figure 4: The overall relationship is strong, but I would investigate the scattered high-error observations rather than relying primarily on R2. Are they concentrated in particular basins, depths, water masses, years, or hydrographic regimes? Those points may be disproportionately informative about processes that the predictor set misses.

The spatial distribution of reconstruction errors across basins and the two depth layers is shown in Figure 7. Figure 4 displays the density of observations on a logarithmic scale, meaning that points further from the main distribution generally represent individual observations or small groups of observations. We will clarify this in the text and add a reference to Figure 7 to guide readers to the spatial context of the larger reconstruction errors.

Section 3.3: The representation uncertainty of 16 µmol kg-1 is adopted from an estimate for total alkalinity. This is a substantial component and apparently dominates the final uncertainty. Please provide stronger justification for transferring this value to DIC, or test sensitivity to this assumption.

We have adopted the same approach as Keppler et al. (2023), who used the estimate of representation uncertainty for surface ocean total alkalinity from Gregor and Gruber (2021) as a proxy for DIC. This was justified by the similar density and spatial distribution of total alkalinity and DIC measurements in GLODAP, as well as the tendency for regions with high spatiotemporal variability in total alkalinity to also exhibit high variability in DIC. As DIC measurements are, on average, 10% more numerous than total alkalinity measurements in most years of GLODAP, the adopted value is likely to provide a conservative estimate of DIC representation uncertainty. We will clarify this justification in the revised manuscript.

Section 3.3 / uncertainty: EN4 uncertainty is explicitly excluded from the propagated reconstruction uncertainty even though EN4 supplies the gridded temperature and salinity fields on which the reconstruction depends. This should be made especially clear when interpreting inventory uncertainties; the quoted ±274 Pg C should not appear to represent all relevant sources of uncertainty.

We agree and will clarify that uncertainties in the EN4 temperature and salinity fields are not included in the propagated reconstruction uncertainty. Consequently, the reported ±274 Pg C represents the uncertainty contributions considered in our analysis, rather than the total uncertainty associated with the reconstructed DIC inventory. We will add this qualification to the manuscript and emphasize it when interpreting the inventory estimates.

Section 4.3: I was surprised by the replacement of predicted DIC in semi-enclosed seas with a global mean depth profile for the inventory calculation. Even if the global impact is small, this is an ad hoc correction. Please identify exactly which regions are replaced, justify the procedure, and provide the inventory both with and without this modification.

We agree that this correction requires further justification and that the affected regions should be clearly identified. These semi-enclosed seas have very few available observations, resulting in poorly constrained reconstructions and particularly large DIC biases. We therefore replace the predicted DIC in these regions with the global mean depth profile to avoid allowing these poorly constrained estimates to distort the global inventory. To make this procedure more transparent, we will include plots showing the regional biases that motivated the correction, the geographical extent of the affected regions, and the differences in the resulting DIC inventory. We will also report the global inventory both with and without the correction to quantify its impact.

Around L430 onward: Calling the diagnosed DIC inventory change an “uptake rate” risks conflating inventory redistribution/storage changes with air–sea CO2 uptake. I suggest using “DIC accumulation rate” unless referring specifically to air–sea carbon uptake.

We will revise the terminology accordingly throughout the section.

Late-period trend: The 2015–2020 estimate of ~5.9 Pg C yr-1 is extremely high and the authors themselves note both increased ensemble spread and declining data availability. I think this result should be treated much more cautiously in the abstract/conclusions unless additional validation demonstrates its robustness.

We agree and will revise the discussion and conclusions to emphasise the uncertainty associated with the late-period estimate. In addition, we will change the temporal averaging from 5-year to 10-year running means to provide a more robust representation of the longer-term trend and reduce the influence of short-term variability.

Figure 13: The caption states that the reconstruction extends from 1990 to 2023, whereas the abstract and main description state 1990–2019. Please clarify the actual product period and the period used for scientific interpretation.

We will correct the caption to state 1990–2019, consistent with the actual reconstruction period.

Around L444: “GLODAPv2.2013” appears to be a typo for GLODAPv2.2023.

We will corrected this to GLODAPv2.2023.

Around L390: There are several unit/typographical errors such as “Pg Pg C/yr” and “Pg C Pg C/yr”.

We will correct the units and typographical errors throughout.

Summary and outlook: The statement that the reconstruction “partly corroborates” variability inferred from surface pCO2 products seems too strong given that the manuscript subsequently acknowledges that this variability may be an artifact of sparse observations. I would phrase this as qualitative similarity rather than independent corroboration unless the robustness tests above support the stronger interpretation.

We agree and will soften the wording to describe the variability as qualitatively similar, rather than implying stronger corroboration between the two estimates.

Summary and outlook: “ResNet-DIC offers an independent observation-based reconstruction suitable for assessing long-term variability and trends in the ocean carbon sink” similarly seems stronger than justified at present. It clearly provides a reconstruction of total interior DIC; whether that reconstruction can quantitatively diagnose variability in the ocean carbon sink is one of the central questions the paper still needs to establish.

We agree that the distinction between reconstructing total interior DIC and quantifying variability in the ocean carbon sink should be clearer. We will revise the statement to emphasise that ResNet-DIC provides an independent observation-based reconstruction of total interior DIC, which may serve as a basis for further investigating long-term changes in ocean carbon storage, rather than presenting it as a direct measure of ocean carbon sink variability.

References:

Sharp, J. D., Fassbender, A. J., Carter, B. R., Johnson, G. C., Schultz, C., and Dunne, J. P.: GOBAI-O_2: temporally and spatially resolved fields of ocean interior dissolved oxygen over nearly 2 decades, Earth Syst. Sci. Data, 15, 4481–4518, https://doi.org/10.5194/essd-15-4481-2023, 2023.

Gregor, L. and Gruber, N.: OceanSODA-ETHZ: a global gridded data set of the surface ocean carbonate system for seasonal to decadal studies of ocean acidification, Earth Syst. Sci. Data, 13, 777–808, https://doi.org/10.5194/essd-13-777-2021, 2021.

Keppler, L., Landschützer, P., Lauvset, S. K., & Gruber, N. (2023). Recent trends and variability in the oceanic storage of dissolved inorganic carbon. *Global Biogeochemical Cycles*, 37, e2022GB007677. https://doi.org/10.1029/2022GB007677