Responses to Reviewer 1 (reviewer comments in grey; our response in blue)

Dear Editor,

I believe the reviewed work is scientifically significant and presents necessary development in the field of ocean biogeochemistry reconstruction. I believe the methods largely sound and the work well presented. I believe the reviewed manuscript should proceed to publication with minor revisions. I include my comments below.

We thank the reviewer for their positive assessment of the scientific contribution and methodology of our work, and for their constructive comments and suggestions. We have carefully considered each comment and suggest revisions to the manuscript in our responses.

Scientific Comments:

> The authors analyse decadal variability in the manuscript but are limited to a time period of 30 years by the availability of predictor data. The authors also express that the start and end of the 30-year prediction should be treated with caution due to temporal edge effects (lines 393-398).
>
> Furthermore, it is noted that the author's describe a negative bias in section 4.2 at lines 387-388 and reference section 3.1, but in the referenced section, appear to describe the evolution of the prediction bias in time (Fig. 5 & 6) rather than directly addressing the bias in the trend itself.
>
> Overall, the reviewer is not confident that the product describes a clear decadal variability (line 477).
>
> Though the reviewer does agree that the product is able to capture the observed and modelled trends over the full prediction period.
>
> *We thank the reviewer for highlighting this limitation. We will moderate our statements regarding decadal variability and change the temporal averaging from 5-year windows to a 10-year running mean. This retains the main temporal variations while reducing the emphasis on shorter-term fluctuations, which are less well constrained by the available observations. We will also clarify the distinction between the temporal evolution of prediction bias and its potential influence on the diagnosed trends.*

- The authors do not describe the treatment of predictor data to the satisfaction of the reviewer. Given the 4-dimensional nature of the prediction; how is 3-dimensional atmospheric CO_2 data treated? The authors do not appear to acknowledge the lack of spatial heterogeneity in atmospheric CO_2 by using a monthly mean value from Mauna Loa globally. Furthermore, are the GLODAPv2.2023 observational data binned (though the duplicates described at line 251 suggests data were not binned) or otherwise processed prior to training?

  We will clarify the treatment of atmospheric CO_2, which is represented by a single monthly, deseasonalised Mauna Loa value applied uniformly across all spatial locations. It therefore acts primarily as a one-dimensional temporal predictor rather than representing spatial variations in atmospheric CO_2. We will also clarify how the GLODAPv2.2023 observations are processed before training, including that all observations are not binned and instead retained at their original locations.

- At line 162, the authors address the greater availability of temperature and salinity data in the EN4 dataset compared to the availability of oxygen, which is a reasonable argument. However, the authors do not appear to put this in context of the even lower availability of dissolved inorganic carbon data in GLODAPv2.2023, which is also described to be $O(10^5)$ (Lauvset *et al.*, 2024). This context would appear to weaken the justification for not using additional biogeochemistry predictors due to data availability, especially given the importance of biogeochemistry predictors indicated in Figures 2 and A14. The reviewer would like the authors to expand and clarify this justification.

  Thank you for this comment. We recognise that our original discussion did not sufficiently distinguish between the availability of observations for model training and the availability of predictor fields for global reconstruction. While the number of DIC observations in GLODAP is an important consideration for model training, particularly because oxygen measurements are often available alongside DIC observations, the limitation we intended to emphasise concerns the availability of a suitable global oxygen field when applying the trained model to reconstruct global DIC.

  Thus, the main limitation of including oxygen is not necessarily the number of available training samples, but the availability of a suitable global oxygen field for reconstruction. Existing options are limited: observational climatologies, such as those based on GLODAP or WOCE data, are time-invariant and do not resolve temporal variability. They also introduce relatively smooth climatological fields into the model. Machine-learning-based reconstructions (e.g. Sharp et al., 2023; Ito et al., 2024) have their own uncertainties and rely

primarily on temperature and salinity, which are already used as predictors in our approach. We will clarify these limitations and their implications for the choice of predictors in the revised manuscript.

- The authors cite Turner *et al*. (2023) as justification that temperature and salinity are well-suited for reconstructing dissolved inorganic carbon. However, Turner *et al*. (2023) conclude that temperature and salinity may be insufficient and with increasing errors below 1000m depth. The authors do not appear to acknowledge this limitation despite the prediction being presented as performing better at depths greater than 1000m throughout the manuscript. The reviewer would like the authors to address how much of this predictive success may be due to the slow-varying concentrations in the deep ocean as opposed to the predictive power of temperature and salinity.

  Thank you for pointing this out. We will moderate the statements about predictive performance at depth and expand the discussion of the factors contributing to the lower errors. In particular, we acknowledge that the relatively low variability of deep-ocean DIC may contribute to the apparent skill, rather than attributing it solely to the predictive information in temperature and salinity. We will also clarify the limitations of the approach in light of increasing errors below 1000m depth in Turner et al. (2023).

- The authors describe validation of the machine learning model in section 3.1 and highlight "*any model likely represents a system of lower complexity than the real ocean*" (lines 260-261). The authors proceed to describe a marginally poorer prediction with larger magnitudes of RMSE and bias for NorESM and ECCO-Darwin model data than for the real-world observations being validated. The reviewer acknowledges the successful validation but finds that the authors weaken their own validation as the machine learning model could be expected to have greater predictive power for a less complex system. The authors do not appear to address this point. A potential explanation for why this is the case may counteract the weakening by raising these points.

  To clarify: the RMSE values for NorESM and ECCO-Darwin represent errors across the global ocean, whereas the observational validation is based on the more limited spatial and temporal coverage of GLODAP. The two sets of metrics therefore do not represent directly comparable validation domains. We will revise the discussion to make this distinction clear and will further include the equivalent results.

- The reviewer is concerned by the spread of ensemble members described in section 4.2 (lines 393-398). The reviewer is not familiar with Residual Neural Networks but it is not uncommon for other machine learning reconstruction methods to tend towards the mean and therefore produce a smaller ensemble spread around the mean year (e.g. 2005) but here the authors find a reduced ensemble spread around the year 2000 and also note that this is a year with low training data availability. The reviewer finds the authors explanation that "*there may be additional uncertainties introduced through temporal edge effects*" to be insufficient explanation, especially when taken in addition to the previous concerns expressed about the impact of temporal edge effects on the decadal variability conclusions over a 30-year prediction period. The reviewer invites the authors to expand on the explanation of why the ensemble spread is smaller around a year with low data availability.

  We acknowledge that the reduced ensemble spread around 2000 is not fully understood. ResNet-DIC does not directly model the temporal evolution of DIC, but reconstructs DIC from relationships with the available environmental predictors. Time (or rather atmospheric CO_2 concentration as a proxy for time) is just one of the predictors. Consequently, the ensemble spread in the resulting inventory time series does not necessarily correspond directly to the number of observations available in each year. The reduced spread around 2000 may reflect factors beyond the temporal availability of training data, but we cannot currently identify its precise cause. We will therefore revise the discussion to avoid overinterpreting this feature.

- The authors raise the point that the Residual Neural Network method employed is highly computationally efficient (lines 492-494) but elected to produce an ensemble of 15 members, lower than typical statistical standards. The reviewer queries this choice.

  The ensemble size of 15 was selected based on recommendations from previous studies, but also due to limited storage for the reconstructions and giving consideration to the carbon emissions associated with large-scale data storage. However, given the computational efficiency of the ResNet architecture, we agree that the sensitivity of the results to ensemble size should be assessed. We will therefore perform additional experiments with increasing numbers of ensemble members and evaluate the convergence of the ensemble mean and standard deviation.

Language/ Text Comments:

- Line 56: "If the variability it is real…" grammatically incorrect.

We will correct this.