

Referee #3

We thank the reviewer for their comments. In response, we have revised the manuscript to better explain our methodology and avoid potential confusion.

The authors have applied algorithms to generate estimates of biogeochemical data and then used an inverse model to attempt to extract an estimate of anthropogenic carbon from the estimated data.

I'm taking an unusual step and recommending this paper be rejected without an opportunity for resubmission as I don't see a way to fix the core idea. The estimation algorithms used by the authors work on the assumption that the covariance between measurable ocean properties is static and well-represented by the neural networks encoded into the associated routines. In the case of ESPERs, the temporal component is intentionally removed (for DIC and pH) before the algorithms are fit, because it was understood that anthropogenic changes would challenge the idea that the underlying relationships are static. The anthropogenic changes are then re-integrated in a rough adjustment that the authors say should not be used for the purposes of anthropogenic carbon estimation: (a quote) "There are uncertainties associated with the assumptions underlying both the 2002 gridded Cant data product and the transient steady state approach—particularly in regions where there are limited measurements of chlorofluorocarbons and other tracers used to calibrate the TTD approach. We therefore assert that {ESPER's Cant strategy} should not be used to estimate Cant distributions for any application where Cant is of primary interest." The ESPER solution to this challenge is also viewed by the authors as an improvement over the CANYON-B/CONTENT approach (and previous approaches) that it updates. In the case of CANYON-B, a temporal predictor allows for estimates to change with time in a simple but opaque way that becomes incorporated into the neural networks. In practice, this approach ends up projecting apparent trends that have been shown to be spurious in many places and not always sensibly related to the apparent Cant accumulation (e.g., significant pH decreases in the deep North Pacific). The phi-naught technique then takes prescribed, but different, versions of these covariances and attempts to adjust-away the exact natural variations that the algorithms are intended to reproduce to distill the underlying Cant.

We believe that there was a misunderstanding regarding which neural networks we used to estimate the carbonate variables in our study. We did not use ESPER_NN to estimate DIC nor alkalinity (A_T). Following Asselot et al. (2024), ESPER_NN was only used to estimate the macronutrients (phosphate, nitrate, and silicate). For the carbonate variables, we used CANYON-B and its associated CONTENT routine to get A_T and DIC, which are then used in the back-calculation method to compute C_{ant} . We have clarified this point in the revised manuscript to avoid confusion.

We agree with the reviewer that ESPER-derived DIC and C_{ant} should not be used when C_{ant} is the primary variable of interest, and this is precisely why we did not use DIC ESPER_NN predictions in this study. The statement quoted by the reviewer therefore does not apply to our C_{ant} estimates, which are based on CANYON-B/CONTENT rather than ESPER_NN.

Nevertheless, we agree with the reviewer on the broader issue that neural-network estimates of carbonate variables may have limitations, particularly when applied outside the temporal and spatial range represented by the training observations. CANYON-B explicitly includes time as a predictor, meaning that its estimates contain a temporal dependence coming from the training dataset. We now explicitly acknowledge this

limitation in the revised manuscript (new Section 2.4 *Evaluation of neural network predictions against observations*). We particularly assessed the performance of the CANYON-B/CONTENT DIC estimates for the post-training period and found that DIC was underestimated by a mean value of $2.74 \mu\text{mol kg}^{-1}$, but this value was comparable to that found for the training period ($2.28 \mu\text{mol kg}^{-1}$). Both estimates represent lower biases than those reported at the global scale in the original CANYON-B/CONTENT paper by Bittig et al. (2018). Because C_{ant} is derived from the estimated DIC, an underestimation of DIC propagates into an underestimation of C_{ant} of approximately the same magnitude, all else being equal. For example, a 5 mol m^{-2} underestimation of DIC (all other variables being equal) results in an approximately 5 mol m^{-2} lower C_{ant} inventory. However, we found that biases in input variables compensate for each other, resulting in a mean negative bias in C_{ant} of only $-0.48 \mu\text{mol kg}^{-1}$.

Finally, to assess the robustness of the temporal evolution derived from CANYON-B/CONTENT, we compared our Argo-based C_{ant} inventories with an independent scaling approach based on the “atmospheric perturbation” concept (see response to reviewer #2). The two approaches rely on different assumptions: the scaling approach represents C_{ant} accumulation associated with the atmospheric perturbation, whereas the Argo-based approach uses the observed temporal evolution of T, S, and O_2 to infer DIC and thereby incorporates changes in water-mass properties and physical structure. We have expanded the discussion of these complementary approaches and their respective limitations in the revised manuscript.

Ultimately, the I fear the authors’ estimated Cant information is, at best, going to reflect the temporal component of the CANYON-B algorithm, but in reality that signal will have been muddled and confused by the combination of multiple algorithms, any data quality issues, and a Cant extraction technique that itself has had its own challenges over the somewhat controversial years of its use. Some time perhaps could have been saved by using the CANYON-B algorithm on a fixed set of seawater properties with a variable time input and discussing the changes over time as though they reflected Cant accumulation... though even that is almost the equivalent of using eMLR and assuming the decadal accumulation rate would continue on indefinitely. It would tell us little, and arguably nothing that isn't already "known" by the CANYON-B/CONTENT neural networks.

We agree that the use of multiple neural networks and the C_{ant} back-calculation method introduces uncertainties that need to be considered. We have evaluated these uncertainties via a Monte Carlo approach (see supplementary materials) and reported these values in the manuscript. However, we would like to clarify that our approach does not interpret the temporal predictor of CANYON-B alone as a measure of C_{ant} . The neural-network estimates are driven by the observed Argo hydrographic properties, and therefore respond to changes in the physical characteristics of the water column. C_{ant} is calculated from DIC and A_T estimated from Argo observations, and the temporal evolution of the resulting C_{ant} estimates is then analyzed. Thus, the temporal variability in our C_{ant} estimates results from the combined evolution of the observed T, S, and O_2 properties and the temporal component considered in CANYON-B/CONTENT.

We agree that the temporal component of CANYON-B/CONTENT is a potential source of bias, particularly because most of our study period lies beyond the temporal range of the training observations. However, as explained above, we have now performed an extensive assessment comparing the CANYON-B/CONTENT DIC estimates with

GLODAPv3 observations in-training vs post-training period (see new Section 2.4). We show that the NN performance is largely maintained beyond the training period, with performance comparable to that observed during training, which gives us confidence in the robustness of our results.

We kindly invite the reviewer to read the revised version of the Data and Methods section, which now includes a new Section 2.4, “*Evaluation of Neural Network Predictions against Observations.*” This additional section provides a clearer presentation of the associated limitations associated with the use of neural networks. We hope that these revisions adequately address the reviewer’s concerns and clarify the methodological aspects of our study.

The authors should be commended on parts of their presentation, and for their hard work making the material understandable. Unfortunately, I believe entertaining this study does more harm than good for the community since it encourages the use of estimated data to capture variability that is only captured by measurements, and this paper comes at a time when funding agencies are looking for opportunities to defund observing and move to a cheaper estimation-algorithm-forward approach. It is unfortunate that the flaws in the method were not identified when the method was first put into the literature, but allowing this study to proceed on those grounds would only compound the error.

We fully agree and support that autonomous observations, such as those provided by Argo, and ship-based measurements in particular, remain essential. They provide the gold standard for the biogeochemical measurements. We do not intend, with our study, to suggest that estimated data can replace observing programs. On the contrary, one of the objectives of this study is to evaluate what can, and cannot, be learned from applying existing neural-network approaches to the Argo-O₂ dataset. The last two sentences of the Conclusion section explicitly acknowledge this point. We regret that the reviewer had not considered the manuscript in its entirety, as doing so would have provided the full picture for assessing the methodology, the presentation and validation of the results, and, importantly, the complementarity of our approach within the broader ocean observing system.

Our objective is not to present Argo-NN based C_{ant} estimates as equivalent to C_{ant} estimates from direct measurements only. This study rather investigates whether the much higher spatio-temporal coverage of the Argo-O₂ network can provide useful and additional information on the temporal and spatial variability of C_{ant} that cannot be resolved from sparse ship-based observations. With our Monte Carlo approach, we explicitly acknowledge that the C_{ant} estimates are subject to uncertainties in the input data, the neural networks and the C_{ant} back-calculation. We also acknowledged that the neural network performance deserved a dedicated section (new Section 2.4), in the interest of transparency regarding the limitations of the approach. At the same time, we believe that assessing, quantifying, and acknowledging these limitations is itself an important contribution to the community.

Rather than advocating a “neural-network” replacement for observations, our study highlights both the potential and limitations of using Argo-based estimates to complement existing observing programs. In this context, we think that the comparison with OVIDE observations and with an independent scaling approach provides an assessment of the robustness of the inferred C_{ant} evolution and helps identify where NN-based estimates may or may not provide reliable information.

Line by line comments:

29: remove the word “capacity” since it makes the sentence less true and the meaning is more clear without it.

Removed.

41: wintertime convection is captured by summertime cruises, since the convection changes the properties that are later observed the next summer. These are inventory measurements, not rate measurements.

We removed “that are not captured by summer cruises.”

50: not even half of ocean anthropogenic carbon is stored in the Atlantic. “Most” is incorrect.

The plain summary language has been removed since Biogeosciences doesn’t require one.

65: language issue “is commonly distinct” “is commonly called”

Changed.

75: amplitude should be magnitude

Changed.

Stopped reading after the methods.

We regret that the reviewer did not consider the manuscript in its entirety. We have carefully considered the concerns raised regarding our methodology and have substantially revised the manuscript to clarify the workflow and quantify the limitations of the neural-network approach. We hope that these revisions will make the scientific objectives, results, and contributions of the study clearer.