

Referee #2

We would like to thank the reviewer for their fruitful comments that helped improve the clarity of the manuscript.

In this paper, DIC and TA are reconstructed from P, T, S, O₂, time in situ data (mostly ARGO-O₂ floats, but which miss the northern tip of the Irminger Sea based on Figure 1) by neural networks (the reconstruction is separately done for nutrients...) (Asselot et al., 2024). Then, back calculations are used to estimate Cant (Perez et al., 2008), similarly to what is done on the OVIDE cruise data. Nice and large use done on the Argo-O₂ network of floats in the North Atlantic subpolar gyre to extend view in space and time that could be reached from repeated cruises (such as Ovide cruises, Bajon et al (2026)).

The main paper results, which extend on what is presented in Asselot et al (2024), are

1: a small negative trend in Cant near 3000 m but only from 2016 - 2021 (with a decrease of a positive trend peaking near 1400m to 3000m), but a good Cant of 23.5 at 3000m, 27.1 at 2000 m and up to 37.4 at 1400m. At the surface in 2021 it reaches 53.2 micromol/kg.

2: For the integrated content to 2000m, a peak in 2016, and nearly constant after. These seem also to be present in the Ovide cruises (+ early 2015 with a larger inventory (from Fröb et al, 2016, 2018), which actually is not well captured in Fig.3) with a decrease both in top and lower layer between 2018 and 2021.

3: There are differences between the local trends in CANT along the Ovide sections, and the Irminger-basin average estimate probably outlining some spatial variability at this fairly regional scale that is captured by the neural networks.

DIC (and AT) here is based on the neural networks of Bittig et al (2018). Thus, the time dependence for DIC in these neural networks comes from using the GLODAP2V2 data base with a decimal variable time, (and T, S, O₂, which implicitly includes seasonal variability, not explicitly introduced) (CANYON-B + CONTENT, which homogenizes the 4 variables of the carbonate system). The data base GLODAP2V2 in recent years in the Irminger Sea is largely based on some of the OVIDE cruises (not all, as well as a few others including the early 2015 Norwegian cruise, also presented here). Thus, in some ways, I am not too surprised that the multi-annual time evolution of DIC is well captured during this decade (at least below the surface) by this neural network (and CONTENT). My main question is which are the cruises that were included, and what are the ones which are not used in this version of CANYON-B.

The CANYON-B training dataset is based on GLODAPv2 and contains observations from 724 cruises going from January 1972 to December 2013. The last OVIDE cruise included in the GLODAPv2 dataset is OVIDE 2010.

This information was added in section 2.2 of the revised manuscript.

I don't know whether there have been any updates on the network since the first paper, and thus in particular whether the more recent OVIDE cruises are or not included. This is important as a large part of the time dependency and trends are based on the period that was fed. To a large extent, I expect (but does not know) that the more recent period (after the last data fed into it) is pure extrapolation. There is a big question on how this time dependency works in the post-training period and whether it is (for the pure Cant part) just some extrapolation with likely larger errors than during the training period. I see no obvious reason why it would work better (for the post-data training part) than any other model, such as the stationary model with scaled forcing adopted,

as a comparison. SO, altogether, I believe there is a need of some discussions, tests, in order to convince the reader.

Part of our study period extends beyond the temporal range of the observations used for training, which covers observations up to 2013 for CANYONB/CONTENT and up to 2019 for ESPER_NN. We agree that the temporal dependence of the neural network should therefore be considered when interpreting trends during the latter part of our time series. Bittig et al. (2018) explicitly evaluated CANYON-B and CONTENT against independent post-GLODAPv2 GO-SHIP observations, including 19 cruises between 2012 and 2017, demonstrating the ability of the method to generalize beyond the training dataset. The authors nevertheless emphasize that CANYON-B and CONTENT can only reproduce spatial and temporal variability represented in the training data, and that extrapolation beyond this range should be assessed (see below).

To further evaluate the performance of the CANYON-B/CONTENT reconstruction during the post-training period, we compared the DIC/ A_T estimates from CANYON-B/CONTENT with DIC/ A_T observations after 2013 (GLODAPv3 database). We added a new section in the revised manuscript (Section 2.4) to detail this analysis. We found, as reported in the original NN paper of Bittig et al. (2018) at the global scale, that DIC is underestimated by a mean value of $2.28 \mu\text{mol kg}^{-1}$, but this underestimation does not considerably increase for the post-training period ($2.74 \mu\text{mol kg}^{-1}$). Because C_{ant} is derived from the difference between DIC and the estimated natural DIC component, a priori a e.g. 5 mol m^{-2} underestimation of DIC might directly translate into an $\sim 5 \text{ mol m}^{-2}$ underestimation of the C_{ant} concentration, all the other biogeochemical variables (T, S, O_2 , nutrients, alkalinity) being equal. Instead, we found that biases in input variables compensate for each other, resulting in a mean negative bias in C_{ant} of only $-0.48 \mu\text{mol kg}^{-1}$.

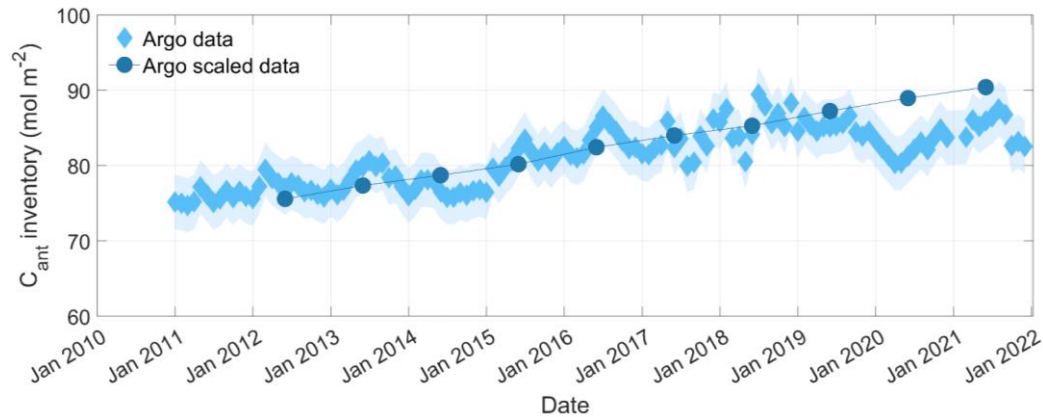
We have included this analysis and further implications in new section 2.4 of the revised manuscript. We also compared the temporal evolution of the Argo-based C_{ant} inventories with an independent scaling approach based on the “atmospheric perturbation” concept (see below), which provides an additional assessment of the robustness of the observed temporal evolution.

On the other hand, I would be more easily convinced that the part of the variability associated to changes in the vertical structure, mixing, ocean circulation that will impact T, S, O_2 , would be nicely reproduced by the neural network of CANYON-B. Is there a way to separate that part of the response in DIC from the changes induced by the increase in atmospheric $p\text{CO}_2$ and surface concentration, that are roughly modeled by the scaled forcing (at least for the vertical integrals).

To assess whether the temporal trends in our Argo-based C_{ant} inventories could be influenced by the temporal dependence of CANYON-B and CONTENT, we compared the Argo-based C_{ant} inventories with independently scaled C_{ant} inventories based on the “atmospheric perturbation concept” (see figure below). At the end of the study period (2021–2022), the Argo-based C_{ant} inventories decline and are lower than the scaled inventories. This decline is also seen in the OVIDE data implying that there must be a physical process (or a combination of processes) explaining this reduction rather than a negative bias between the NN-predicted DIC and observations.

Nevertheless, the difference between the two approaches also reflects their fundamentally different assumptions. The scaling method assumes that oceanic C_{ant} follows the atmospheric perturbation and therefore does not account for changes in vertical structure, mixing, or ocean circulation. In contrast, the Argo-based approach uses the observed temporal evolution of T, S, and O_2 and can therefore account for such changes in the physical structure of the water column

when estimating DIC. We have therefore revised the discussion to acknowledge both effects: the Argo-based estimates provide information on variability associated with physical changes that cannot be represented by the scaling approach, but potential temporal variations in DIC bias constitutes an additional source of uncertainty that must be considered when interpreting the magnitude of the C_{ant} trends.



Legend: C_{ant} inventory between the ocean surface and 2000 dbar. The blue diamonds represent our Argo-based C_{ant} inventories computed with the neural networks and the back-calculation method. The blue circles represent scaled Argo C_{ant} inventories. This scaling is done with the “atmospheric perturbation concept”.

There might be also an other issue related to the non-explicit consideration of the seasonal variability during the training phase (well as argued, it goes with the changes in near surface T, maybe also through the time variable in the model which is decimal). In particular, I worry about how the winter vertical mixing is modeled, as the GLODAP2V2 base is largely based on non-winter data in the Irminger Sea (with the exception of the early 2015 Norwegian cruise).

Seasonal variability can be indirectly captured through the biogeochemical variables, such as temperature, salinity, and oxygen, which reflect the seasonal evolution of the water column. In addition, Bajon et al. (2026) showed that the seasonal variability reconstructed by the NNs exhibits a physically plausible behavior, providing confidence that the NNs are able to reproduce the main seasonal processes affecting C_{ant} variability (see section 2.4 in the revised manuscript).

However, we acknowledge that the neural networks are not explicitly constrained by observations from different seasons and are instead biased toward summer observations. This limitation is relevant in the Irminger Sea, where winter convection can strongly modify the vertical distribution of biogeochemical properties. The GLODAPv2 observations used for training provide limited winter coverage in this region, and therefore the neural network is less constrained under wintertime conditions. We have clarified this limitation in the revised manuscript.

Importantly, our analysis does not rely exclusively on winter observations: we construct monthly C_{ant} estimates from Argo observations and compare the resulting temporal evolution with independent ship-based observations and with an independently scaled C_{ant} inventory. The consistency between these approaches provides some support for the robustness of the observed multi-annual evolution. Although we acknowledge that the limited representation of winter conditions in the training dataset prevents us from fully quantifying the uncertainty associated with winter convection.

Minor comments:

1. 43 ‘interpretation from ship-based measurements’

Changed

2. 58: 'to the ocean bottom'

The plain summary language has been removed because Biogeosciences doesn't require one.

3. 82: what is meant by 'intensified ocean acidification rates by 7-10% (Curbelo-Hernandez et al, 2024)? Was the ocean acidification rate estimated both in 2009 and 2019 (and over which period each time), or is it an average 2009-2019 compared to estimates from earlier periods (in this case, which one?)

Yes the ocean acidification rate was estimated both in 2009 and 2019, and during this period the acidification rates increased by 7-10%. We rephrased.

4. 126: 'this study investigates...'

Changed

5. 165: '... with a few additional profiles in the Gulf of Mexico and in the Mediterranean Sea'

Changed

6. 190-193: uncertainties seem here to be based only on the ones on the T, S, O₂ profile data base, whereas the uncertainties estimated in the CANYON-B + CONTENT tools are not taken into account (I realize that these might not be directly compatible, in the way the Bayesian approach is done...). Thus, it seems to me that the uncertainty reported is only part of the total uncertainty. On the other hand, this random uncertainty might be the most relevant here. L. 193 reports Supporting information, but which I did not find when downloading the preprint?

We refer the reviewer to the revised manuscript to have access to the supporting information. Here below we briefly summarize how the uncertainties are estimated :

The accuracies associated with the four Argo-O₂ input variables (P, T, S, and O₂) are taken as uncertainties to estimate the overall C_{ant} uncertainty. For each data point, 100 realizations were generated using a Monte Carlo approach. For each input variable X, the perturbed value X_{MC} was randomly sampled according to its associated uncertainty σ_X :

$$X - \sigma_X \leq X_{MC} \leq X + \sigma_X$$

where X_{MC} is the perturbed input variable, X is the original value, and σ_X is its associated uncertainty. The resulting perturbed P, T, S, and O₂ values were then propagated through the neural networks and the C_{ant} back-calculation method. This procedure produced 100 independent realizations of each C_{ant} estimate. The standard deviation of these 100 realizations was then calculated for each data point and used as the overall uncertainty of the C_{ant} estimate.

Because we went through the whole methodological chain (neural networks + back-calculation method), to estimate the overall uncertainty, this uncertainty takes into account the uncertainties estimated in the CANYON-B + CONTENT tools.

Note that the revised version of the manuscript also includes a new Section 2.4 on the *Evaluation of neural network predictions against observations*, where we assess in more detail the uncertainties associated solely with the neural networks.

7. 216: again, the use of the Monte Carlo method with no precise information on how it is implemented (how are the uncertainties correlated in the vertical? Maybe this is in the Supp Mat that I have not read?)

We refer the reviewer to the answer above and to the revised manuscript to have more information on how the Monte Carlo method was implemented.

The uncertainties are the smallest in the ocean surface (0-100 dbar) and the occasional maximum uncertainty is around 1000-1500 dbar. This information is given in the revised manuscript.

Section 2.6 l.238-254. I appreciate the use of the scaling method, as a way to weigh non-local effects, changes of circulation, convection, etc... compared to this very simple local accumulation assumption. I would add that the assumption of the exponential change is not only related to the 10 years investigated here, but has to be valid further in time, due to the average age of the different deep and intermediate water masses (but usually not that old in the Irminger Sea).

We agree with the reviewer and added a sentence in the revised manuscript: “More generally, the validity of the exponential atmospheric perturbation assumption must extend over timescales comparable to the mean age of the water masses considered, particularly for intermediate and deep waters.”

Fig. 3: I don't understand the difference with OVIDE data in Panel c, which I would have expected to be the sum of the differences in panels a and b (for example, for the 2016 Ovide data: by eye I would have thought that the sum of a and b would yield a difference with blue curve on the order of 2-3 mol/m², much less than what is plotted on c and is close to 10). Is something else done?

We would like to thank the reviewer for this comment. We found an inconsistency in the selection of profiles for the different depth ranges (0-2000 dbar, 2000 dbar to bottom or full water column). This has now been corrected, which yielded to slightly different OVIDE-based Cant values that do not change our conclusions.

Also, on Fig. 3, is there an explanation for the large difference in early 2015 with the Norwegian cruise data (I am aware that on top panel it is the TDD method used and not the f0CT one, but the lower panel for integrated value show a similar change, and this does use the TDD method?). Is it an issue of spatial locations of the April 2015 stations emphasizing the convection event, and not so much a wider average, or is it that the neural network does not well capture the increase in Cant during convection (and in which case, does one have good confidence in the increase in early 2016 with the neural network?). Altogether, although the difference is mentioned, it is not always clear when and where the comparison is carried (separating the product averaged along the Ovide line, but then which stations are kept from the early 2015 cruise, or whether it is basin-averaged).

For the top panel of Fig. 3, our Argo estimates show inventories 5-6 mol/m² lower than Fröb et al. (2016). This might come from the use of the TTD method by Fröb et al. (2016).

For the lower panel of Fig.3, the reviewer is right, Fröb et al. (2018) used the ϕC_T^O method. Unfortunately, no OVIDE observations are available for 2015, preventing a direct comparison between the OVIDE-based inventory and the 2015 estimate of Fröb et al. (2018). However, 2015 was characterized by an intense deep convection in the Irminger Sea (Fröb et al., 2018; Piron et al., 2017), which could explain the high C_{ant} inventory reported by Fröb et al. (2018) for that year.

We add these explanations in the revised manuscript.