RC: Referee comment; **AR**: author's response; **AC**: author's changes in manuscript.

Referee #1 response

RC: The authors of this study firstly extracted the relations between alkalinity and other variable (Salinity, DO, nutrients, depth, temperature and location), and adopted this relationship to generate a monthly climatology of total alkalinity. I am glad to see the a more precise alkalinity climatology dataset. However, the manuscript is poorly structured, inadequate illustrated with a lot of vague expression. There are tons of sloppy description, and the grammar is so poor that I have difficulty to understand the science. There are quite a few things that need clarification from the authors. They are listed from the most to least concerning.

AR: We are pleased to see that you appreciate the more precision of our alkalinity climatology dataset as well as we appreciate all your useful comments. Thank you so much for the thorough revision of the manuscript. We hope to improve it based on all the comments of the 3 referees. In order to maintain a balance between your comment about the poor structure and the vague expression of the manuscript and the comment of the Referee #3: "The manuscript is clearly-structured and well-written", we have reviewed these points which are mainly depicted in the next responses. At the end of the answers is attached the new version of the manuscript for a global view.

RC: I fail to see why "3.3 subsurface layer hypothesis" is included in this manuscript. This section plagued with serious issues: without enough background, it is very difficult for readers to understand the motivation: 1). What is the subsurface layer hypothesis? 2) Whether your finding support or reject the hypothesis? 3) How this part related to the topic of monthly climatology at all? Figure 7 is also not well explained either. I have no idea how to read it. Can you add an autocorrelation figure to show the similarity between surface winter condition and subsurface layer?

AR: The motivation was explained in lines 185-194. In brief, the motivation to test this hypothesis, as Vázquez-Rodríguez et al. (2012) proposed, has been the lack of winter data in some regions. In section 3.3, we demonstrated in the same way as Vázquez-Rodríguez et al. (2012) did, that their hypothesis is verified for other regions with a lack of data in

GLODAPv2. For the non-winter months when samples were taken from the subsurface layer similar relations exist between predictors and AT. This fact avoids a seasonal bias in these areas where the presence of non-winter data could bias the climatology to the months represented in the training dataset. Figure 7 is exactly designed and described as Figure 2 in Vázquez-Rodriguez et al. (2012). The reader is referred for further detail to the cited paper for an in-depth analysis of the similar results that are shown in their Figure 2.

We have added an additional analysis in this section to reinforce the subsurface hypothesis through training a neural network without winter data and testing its ability to fit the independent winter dataset. That is, extract the relations only from non-winter data and test them in winter independent data.

AC: In methodology: "GLODAPv2 contains quality controlled measurements in all ocean basins from the 1970s until 2013 (Olsen et al., 2016). However, winter data are scarce to absent in some high latitude regions because adverse weather conditions prevents field activities in that season (Fig. 3). In surface ocean, this temporal bias can be avoided with the help of the subsurface data from seasons with sufficient samples. Vázquez-Rodríguez et al. (2012) demonstrated how the subsurface ocean layer in the Atlantic Ocean can retain the footprint of the water mass formation from the preceding winter in the following months and, therefore, of the surface conditions. The winter relationship between inputs and AT needed to produce an all-season surface climatology are mostly preserved in this subsurface layer. The validity of this hypothesis was tested in other regions (Fig. 3) following Vázquez-Rodríguez et al. (2012). These areas were chosen based on the non-availability of AT data in two or more consecutive months in the same oceanographic regime as the colored area in Fig. 3.

To reinforce the previous test and to assess the ability of the neural network in overcoming the lack of winter data in other depths, a neural network was trained excluding all winter data in GLODAPv2 (GLODAPv2_nowinter) and tested in the excluded and independent winter dataset (GLODAPv2_winter). The procedure to create and to train the network was the same as described previously."

In results and discussion: "We found that the optimal depth range of the subsurface layer defined by Vázquez-Rodríguez et al. (2012) for the North Atlantic Ocean (100-200 m) must be modified in other regions. In the area analyzed in the Indian Ocean (Fig. 3), the

subsurface layer hypothesis is verified in the same depth range of that study. However, the other areas (Fig. 3) show that the range of the subsurface layer is in the range of 50-100 m. The different strengths of deep mixing and convection in winter could explain this fact.

The properties analyzed in the four areas defined in Fig. 3 show, as expected, a higher monthly variability in the ocean surface than in the subsurface layers. The seasonal variability depicted in Fig. 8 will likely be typical of a larger region within a similar oceanographic regime for each defined area. The surface winter conditions of the analyzed properties are quite similar to those in the subsurface layer during, at least, one of the four consecutive months following winter in all areas (Fig. 8).

The optimal number of neurons in the network trained with GLODAPv2_nowinter dataset to reinforce the subsurface layer hypothesis and to assess the layers below surface ocean was 100. The reduction of the number of neurons compared to the previous networks was because this new dataset contains less data. Thus, maintaining or increasing the number of neurons would produce overfitting. This new network provides statistics in the GLODAPv2_nowinter dataset similar to those of the network used to create the climatology (NNGv2) in GLODAPv2 dataset (Table 1 vs Table 7). But, of greater importance are the statistics resulted from the GLODAPv2_winter dataset (Table 7) which reinforce the subsurface layer hypothesis. The low error reached in this independent winter dataset shows how the network is able to obtain the winter relations in any depth from the function fitted with data from other seasons. Therefore, the lack of winter data in different regions does not automatically mean that the climatology will be biased towards the more sampled seasons."

RC: Lines 315-390, The authors found that climatology of TA is highly dependent on the inputs. However, the logic can be improved. I would suggest re-organizing this session as this: 1). Explain the available TA climatologies, and what is the difference among them; 2) Why did authors choose WOA13 as input at last? 3). Show the monthly climatology of TA caluculate based on WOA13, discuss its variability, and compare yours and others.

I also have a question related to input climatology choosing: the authors showed the difference between "AT NN WOA13"and "AT NNBATS inputs" in Fig. 10. Have the

authors tried to use climatology reported by Lauvset et al 2016 as inputs? And what is the result?

AR: The sequence for the referred lines is: 1) To show the climatology and discuss the patterns and the variability obtained; 2) To analyze if the variability is coherent comparing the climatology with the unique climatological measured data (BATS and HOT time-series); 3) Once we have shown that the climatology is robust, we have compared all the available climatologies and ours. Your point 1) and the last sentence of the 3) is our point 3). We have considered your suggestion but separating the comparisons will break the logic. We have used the climatologies of our input variables given by Lauvset et al. (2016) to assess the difference in both methods and inputs between their climatology and ours. As the climatologies of Lauvset et al. (2016) are not seasonal, the comparison of the Fig. 10 that you suggest cannot be done.

RC: There are a lot of unclear pronoun references across the manuscript, which make sentences very confusing and difficult to understand. There is no way to point out all of them. Please check through the Manuscript.

AR: The whole manuscript has been checked by a native speaker of English.

RC: 1). The caption for tables should be put on the top. It is very confused with the current format.

AR: Changed.

RC: 2). Line 46, increase in temperature and ocean deoxygenation.

AR: Added.

RC: 3). Lines 49-50, It should be five variables if including carbonate saturation state (Ω). I would also suggest adding alkalinity definition before discussing its physical meaning and processes that can impact its distribution.

AR: Added.

AC: Dickson (1981) defined A_T as:

$A_T = [HCO_3^-] + 2[CO_3^{2-}] + [B(OH)_4^-] + [OH^-] + [HPO_4^{2-}] + 2[PO_4^{3-}] + [SiO(OH)_3^-]$

$+ [HS^-] + 2[S^{2-}] + [NH_3] - [H^+] - [HSO_4^-] - [HF] - [H_3PO_4]$

RC: 4). There are a lot of upwelling studies in Californian Upwelling Systems, some references may be needed here.

AR: Added Millero et al. (1998).

RC: 5). Lines 62-63 "For example, phytoplankton blooms (i.e., primary production), and the seasonality in upwelling and river flows" is not a sentence.

AR: Changed

AC: "Phytoplankton blooms (i.e., primary production) and the seasonality in upwelling and river flows are some of the more remarkable processes associated with the time variability of AT."

RC: 6). What do you mean by the "Storage of the anthropogenic CO2"? You mean the TA's seasonal cycle is important to the anthropogenic CO2 storage?

AR: We have added a reference to the paper Renforth and Henderson (2017) about carbon sequestration in the ocean where the seasonality of A_T is mentioned.

Renforth, P., and G. Henderson: Assessing ocean alkalinity for carbon sequestration, Rev. Geophys., 55, 636–674, doi:10.1002/2016RG000533, 2017.

RC: 7). Lines 165-166, It is a very confused sentence. I still cannot figure out how the authors do the training.

AR: Text between lines 151 and 166 has been deleted and new one was added to clarify the training procedure. A figure was also added (Fig. 2).

AC: New text: "The training procedure was carried out in MATLAB. We tested 16, 32, 64, 128 and 264 neurons in the hidden layer based on the results of Velo et al. (2013). For each number of neurons, we trained 10 networks always using the same 90% of GLODAPv2 for training (Fig. 2, Static level). The remaining 10% was used as a static test (Fig. 2, Static level). Both subsets contained samples randomly distributed in the ocean to evaluate the maximum possible relationships between the input variables and AT through all oceanographic regimes, that is, to capture most of the variability in all the variables and not restricting the sets to specific areas. Each of the 10 networks starts the training procedure with random weight and bias values and a random division of the training static dataset into three portions: 70% for training, 15% for testing and 15% for validation (Fig. 2, Dynamic level). These differences make minimization of the cost function different for each network due to the complexity of the weight-error space and, consequently, their different starting points in that space. As each network is different, keeping static sets allows one to determine which network best generalizes in the same test set. The selected network is the one that produces the lowest RMSE in the training data (validation + training dynamic) and in the test data (static + dynamic), considering a non-significant difference between both RMSEs to prevent overfitting. The network derived from this process will be referred as NNGv2.

Once we found an adequate network configuration, we increased the amount of data in the training dynamic set to capture more relations between the inputs and AT. The new percentages of the dynamic sets were: 80% training, 20% validation and 0% testing. The latter set is only necessary to compare different models and is not used during the training. However, the static test set was held to evaluate the generalization of each of the 10 networks to select the best one."

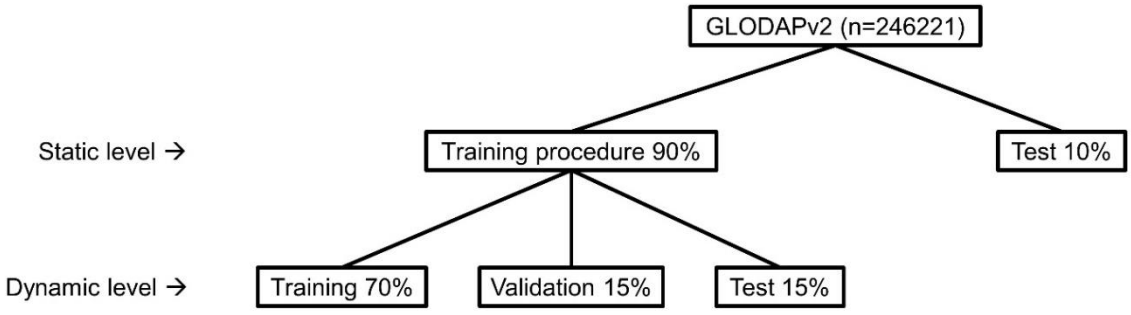


Figure 2. Division of the data for the training of the network. The data in the sets of the static level is the same for all the networks to train. The data in the sets of the dynamic level is randomly selected for each network to train.

RC: 8). Lines 192-193, again, it is a very wordy sentence, and I have no idea how the authors concluded the content after Thus.

AR: Vázquez-Rodríguez et al. (2012) concluded that the relations between some variables and AT in wintertime at surface can be found in other seasons in a subsurface layer. Therefore, we are using their conclusion and adapt it to our study. Furthermore, we test the validity of their hypothesis in other ocean areas.

RC: 9). Line 213. "They make up 6.5% of all the samples in this zone and 85% of them belong to the upper 100m of the water column (Table S2)". What does the "They" represent? How did you get 6.5% based on Table S1? Adding all the %relative over n is 5.35+6.90=12.25%. I would assume the second number should be 1317/(296+1289)=83.09%. By the way, the fifth column number was also miscalculated.

AR: "They" is used to refer to the samples with residuals beyond +-3RMSE in this paragraph because we did not want to repeat "samples with residuals beyond +-3RMSE" all the time.

6.5% is the percentage of samples with residuals beyond +-3RMSE over total GLODAPv2 samples in the indicated area: latitudes greater than 60ºN. From table S2: (296+1289)/(5531+18684)=0.065.

About the second number, 85%, your assumption is correct. We wrote the value obtained with a previous network that we tested. We have changed the number in the text and in Table S2. The percentage in the fifth column was also updated.

AC: The paragraph between lines 210 and 216 has been changed for a better understanding as follows: "Samples with residuals beyond ±3RMSE are 1% of the GLODAPv2 dataset. The spatial distribution of these samples (Fig. S3) show that they are confined to certain areas, mainly in the ocean surface (Fig. 4). Most are in the Northern Hemisphere (Fig. S3 and Fig. 4). Specifically, 64% are from latitudes north of 60ºN (Table S1). In this area, 6.5% of GLODAPv2 samples have residuals beyond ±3RMSE and 83.1% of these samples are from the upper 100m (Table S2). In these depth and latitude ranges, the samples with high residuals make up 14% of the GLODAPv2 samples here and they typically have salinities lower than 34 (Table S3; Fig. S3). A monthly

analysis in the previously indicated ranges shows that the largest number of samples with residuals beyond ±3RMSE are from the summer months. About 15-19% of all the samples from this season in this area have residuals higher than ±3RMSE (Table S4)."

RC: 10). Line 214. "in this layer of this area: : :: : :14% of the total". Please specify what "this layer of this area" represent. And what the "total" here is? The same problem with line 216 "in this area".

AR: Changed.

AC: "in this layer of this area" → "In these depth and latitude ranges"

"total" → "GLODAPv2 samples here"

RC: 11). Lines 219-220. Do not know what author want to say.

AR: Because of the peculiar process in the Arctic, the subsurface layer hypothesis might not be valid here and the climatology could represent only the characteristic AT of the sampled months.

RC: 12). Lines 242-244. This sentence needs to be revised. The current description makes the reader think the Lee et al (2006) have the lowest RMSE comparing to other methods. Also, Line 245. It is not "We", it should be the "NN approach".

AR: This is indeed that we want to say. Compared with other methods on the generation of a monthly climatology, as it is specified in the previous sentence: "previous studies on generation of monthly climatologies. Anyway, we have changed some things in the paragraph for a better understanding, including the suggested one in your comment.

AC: New text: "In the global ocean surface layer, the RMSE obtained with the neural network approach is lower than that obtained by previous studies on generation of monthly climatologies (Table 2 and 3). In the past, relationships between SST and SSS with AT by Lee et al. (2006) have been shown to produce the lowest RMSE (area-weighted RMSE of 8.1 µmol kg-1) in the AT computation to create a monthly climatology. However, applying the relations of that study to GLODAPv2, the obtained

weighted RMSE is higher than the one from the neural network (Table 2). Neural network approach obtained a better fit in all the areas defined in the study of Lee et al. (2006) (Table 2). NN±3RMSE improves the results obtained with the NNGv2 in almost all the regions, being the most remarkable the Equatorial Upwelling Pacific. However, the difference in the weighted RMSE of the two networks is not significant."

RC: 13). Line 251. "The zones defined in the Arctic have higher RMSEs in the two studies" I have no idea what the authors want to say.

AR: Changed.

AC: New text: "The AT computed in the zones defined in the Arctic have higher RMSEs in the two approaches (Takahashi et al. (2014) and this study; Table 3)."

RC: 14). Lines 256-257, is not related to this section. Monthly climatology should be discussed in next section.

AR: We are comparing the accuracy of the available methods designed to create a monthly climatology. Therefore, the monthly climatology is not discussed here, only the fitting techniques.

RC: 15). Lines 257-264, this paragraph should be put after Line 275.

AR: In this paragraph we are showing which of our two networks is selected to create the climatology depending on what is discussed in the previous paragraph. Therefore, it follows a logic sequence and put it after Line 275 would break the sequence of the text.

RC: 16). Line 278, the authors should list the three time-series first.

AR: Added.

RC: The same as Line 360. Have no idea what the "other climatologies" before jumping into figures.

AR: Changed

AC: "Compared to the other climatologies" → "Compared to other climatologies (Lee et al. (2006), Takahashi et al. (2014) and Lauvset et al. (2014))"

RC: Line 279-280, why?

AR: To specify the good performance of the network in datasets different from the previously tested (GLODAPv2), mainly in the time resolution, one of the important features of our study. To sum up, previously we showed the generalization of the network in GLODAPv2 samples located randomly around the world and now we are showing the generalization in specific locations over samples measured monthly and/or seasonally.

RC: 17). Line 284, "We obtained similar values of RMSE of 6 _mol kg-1 and 5.5 _mol kg-1 respectively". At which time series stations? Both values cannot be found in Table 4.

AR: This paragraph has been deleted and we have added a more concise comparison with LIARv2 and CANYON-B.

AC: "The LIARv2 and CANYON-B methods to compute A_T also model the time-series data quite well (Table 6). Significant differences among the three methods are obtained in HOT and ESTOC. In HOT, NNGv2 and CANYON-B reach a better fit of A_T than LIARv2 suggesting that a non-linear technique is more adequately to model A_T in this area (Table 6). In ESTOC, NNGv2 and LIARv2 are the best options to model the A_T variability (Table 6). Here, the A_T computed with LIARv2 with the option of the free equation choice activated results in a greater election of the equations that include nutrients as predictors. This result show how in this area the inclusion of nutrients as predictors contributes to improve the model of A_T. Like NNGv2, both methods have a considerable bias in K2 and KNOT (Table 6) that reinforce the two reasons suggested previously."

Table 6: RMSE and bias between measured A_T and the A_T computed with both LIARv2 and CANYON-B methods. The comparison was done for the same samples evaluated in Table 5.

	LIARv2		CANYON-B	
Time-Series	RMSE (µmol kg^{-1})	bias (µmol kg^{-1})	RMSE (µmol kg^{-1})	bias (µmol kg^{-1})
HOT	**6.6**	**-0.6**	**5.8**	**-0.6**
BATS	**6.3**	**0.1**	**6**	**-0.4**

ESTOC	**3.4**	**0.8**	**4.2**	**3.2**
KNOT	**4.8**	**-6.6**	**4.5**	**-7.2**
K2	**3**	**-3.0**	**3**	**-3.3**

RC: 18). Lines 286-288, Too much repeat.

AR: Deleted in the new version of the manuscript. See previous comment.

RC: 19). The way to mark panel is very confusing in Figure 6. Please assign each panel an ID. By the way, please explain how did you get the AT,residue without measured value in both time series stations.

AR: Figure 6 and its title have been changed. The AT residuals were obtained from the difference between measured and computed AT as is explained in the figure title (Figure 6 panels in the central column). The panels in the right column are the same as those of the central column but applying an interpolation (DIVA) (as it is written in the title of the figure) for visual purposes.

AC:

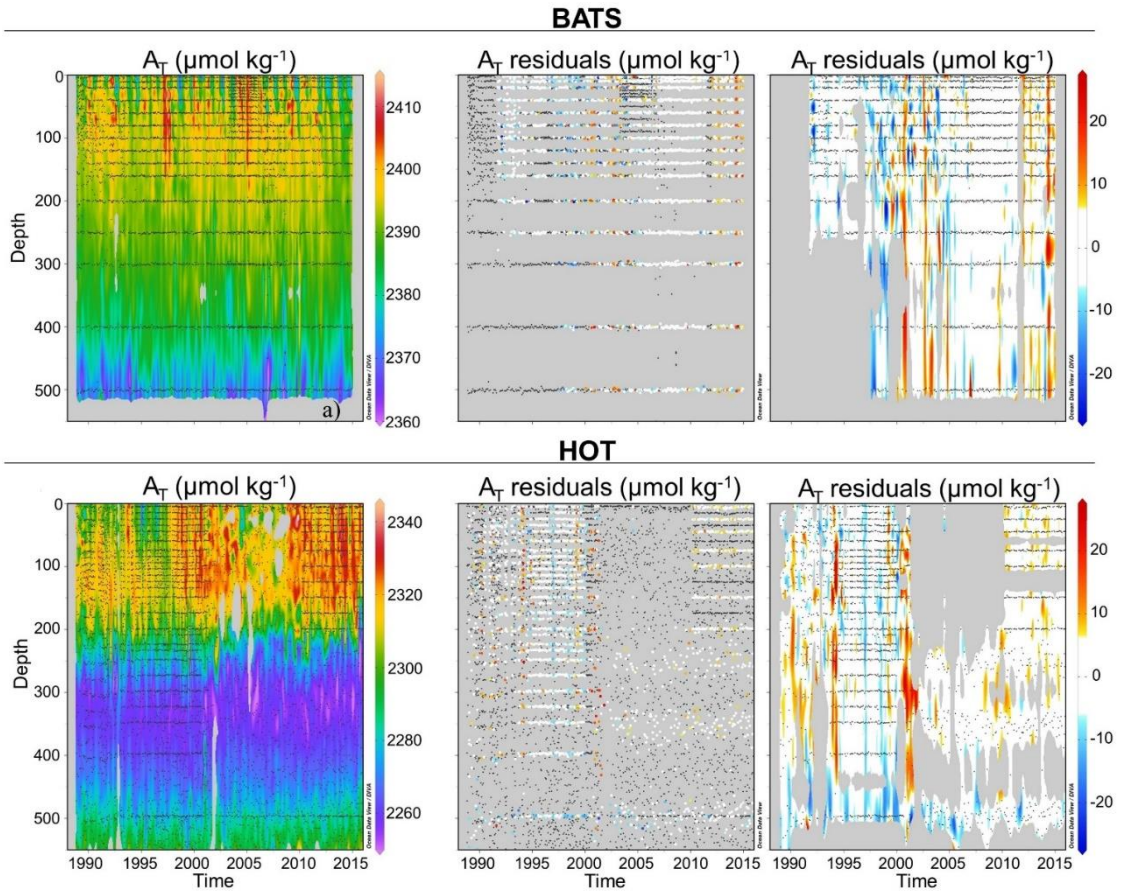


Figure 7: Left column: Computed A_T for the upper 550m of the water column at the BATS and HOT time-series stations. Central column: Difference between measured and computed A_T. Colored dots show samples where A_T was measured. Black dots show samples where A_T was not measured but the network inputs were. Right column: Difference between measured and computed A_T interpolated with Data-Interpolating Variational Analysis (DIVA; Troupin et al., 2010). This figure was made with Ocean Data View (Schlitzer, 2016).

RC: 20). Lines 350-359, the figure across this paragraph should be figure 10!

AR: Changed.

RC: This paragraph (Lines 350-359) and following paragraph is very sloppy written. The authors should re-write it. For example, Line 355 can be simply written as "the comparisons are better (and show how better) when AT was obtained by NN with measured value as inputs".

AR: That sentence has been rewritten following your indication. However, we don't consider including the statistics in the text. The figure is referenced in that sentence and the reader can see the statistics in the figure together with graphs for a better understanding.

AC: "The comparisons are better when A_T is computed by NNGv2 using as inputs the measured values in the time-series (Fig. 11, purple line)"

RC: Line 356. "The differences of the two comparisons show the differences in the input variables". Have no idea what the second "differences" means.

AR: Changed

AC: "The differences of the two comparisons show the differences in the input variables (WOA13 climatological fields vs time-series input data)."

RC: Line 360. Replace "similar" with "close".

AR: Changed.

RC: What is the "one predictor" in Line 362?

AR: It is indicated in the introduction when the description of the study of Takahashi et al. (2014) is written. However, we have repeated it here again.

AC: "…one predictor…" → "…one predictor (salinity)…"

RC: What does DIVA represent?

AR: The acronym is depicted in Figure 6 title, which is previous to this line in the text sequence. However, we have added it here again.

AC: "DIVA (Data-Interpolating Variational Analysis)

RC: Line 363. “Furthermore, the coarser grid in the Takahashi et al. (2014) climatology involves a change of grid for the comparisons which may enhance dissimilarities”. I have no idea what the authors want to say at all! Again, the above questions are only a few examples. the authors have to check through the entire MS and do the corresponding revision.

AR: To compare our climatology with the one of Takahashi et al. (2014), ours was transferred to their grid through an average which could be a contribution to the differences between the climatologies. We have rewritten the sentence.

AC: “Furthermore, the transfer of our climatology to the coarser grid of Takahashi et al. (2014) for the comparisons may enhance dissimilarities.”

RC: 21). Line 370. “The spatial patterns of the differences between in annual mean surface AT between our and the three other climatologies under consideration are not correlated.” Get lost again.

AR: The “in” is a typo.

AC: “the surface spatial patterns of the differences between the annual mean of our AT climatology and the three other ones under consideration are not correlated (Figure S7).”

RC: 22). Lines 374-375. “It shows how the different parametrizations of the AT diverge highly at low salinities.” How do the authors get this conclusion?

AR: The previous sentence is “The largest differences in these two ocean basins are mainly located close to the river mouths.”. Therefore, it is clearly shown that the different approaches to compute A_T compared here (this study vs Takahashi) show the largest differences at low salinities.

RC: 23). Again, what is the “the difference results”?!

AR: Changed.

AC: “An important cause of the differences between the climatologies stems from the use of different inputs to generate them.”

RC: Add Figure 11 at end of Line 384 (… of the WOA13 data).

AR: Added.

RC: Line 387. Do you mean “below 250 m” by “in these layers”?

AR: Yes, we do. We have joined the two sentences for a better understanding.

AC: “The values of the RMSE of the comparisons like those in Fig. 11 but below 250m are in the range of 4 to 6 µmol kg^{-1} and the improvement caused by the inputs usage is reduced to around 1 µmol kg^{-1}”.

RC: 24). Lines 389-390. “to be consistent, it is recommended to use the AT climatology corresponding with the other inputs used in the studies that arise from these products.” Have difficulty to understand it too.

AR: Each climatology depends on the fields used to create them (Lee et al (2006): WOA01; Takahashi et al. (2014): Lauvset et al. (2016): GLODAPv2; this study: WOA13). Therefore and as an example, if an ocean carbon cycle modeling study is using the temperature fields of WOA13, the recommendation that we make is that our AT climatology should then be used too.

RC: Referee comment; **AR**: author's response; **AC**: author's changes in manuscript.

Referee #2 response

RC: Summary: It is clear that hard and good work was done in getting this paper this far and the authors should be congratulated on their progress toward what, to me, looks like 2 papers. However, more must be done before this will be concise, clear, complete, and novel enough to warrant being broadly read by the oceanographic community. This manuscript attempts several things: First, it justifies the need for at TA climatology. Next it produces a new neural network for calculating TA from other seawater measurements. It then assesses the neural network and discusses likely sources of error for the method, paying special attention to riverine influences in the Arctic. Finally, it presents the climatology and compares it to other TA climatologies in the literature. Unfortunately, the paper uses a lot of text to only do half of the job with each of these objectives. I was left confused what use the authors had in mind for the climatology (I don't dispute that uses exist, but rather suggest that the uses were not clearly communicated). The methods used to create the neural network are similarly incomplete, where significant text is devoted to their description but not enough text is devoted to the explanation for it to make sense to people who don't already understand the material. The neural network is created, but there is only an effort to test the optimal number of neurons, and insufficient efforts are made to optimize other aspects of the NN, such as the combination of predictors used to calculate TA. The neural network assessment is incomplete (see below), and insufficient effort is also made toward comparing the new neural network to options in the literature, e.g. the Sauzude et al. CANYON reference, the Carter LIAR et al. reference (updated here: https://aslopubs.onlinelibrary.wiley.com/doi/full/10.1002/lom3.10232), or the recent Bittig et al. CONTENT methods (https://meetingorganizer.copernicus.org/EGU2018/EGU2018-2774.pdf). I attempted to do some of these comparisons on my end, but couldn't get the code to work. Finally, the presentation of the climatology itself is rushed and contains too many vague and general statements. Going forward, consider splitting the paper into two, both halves of which will require more work before being ready for publication. Alternately, shift focus towards or away from algorithm development. If towards, then do a complete job of optimizing parameters and testing the NN against alternatives. If away, then simply omit the results from this new algorithm and use existing ones or present the climatology alongside estimates from alternatives. If split: : : For the first paper, a more complete case must be made as to why the new methods are better or better specifically for generating a climatology than existing methods. Most of this case can be made by showing the new method has decreased (or comparable but independent) errors to alternatives, and this can be shown by improving the validation text with a number of new quantitative comparisons employing the various methods. Randomly selecting testing/validation data is not useful since there are large systematic TA errors along hydrographic sections. This means you will always underestimate error along a section

if you train your routine with data from the same section as the test data. For this reason, CANYON authors reserved entire regions for their test data and LIAR authors omitted entire sections at a time during testing. This may or may not end up making the computed RMSE worse, but it is an important step regardless so readers know what to expect from the algorithm when they deploy it in areas where there weren't nearby dense measurements collected at the time of the estimate. It wasn't clear what omitting data with a >3 umol/kg bias for the 2nd NN training was intended to show. Uncertainty is necessary for these estimates. For the second paper, creating a climatology from an algorithm is fast. The second paper will be complete when the authors have answered the following questions: 0. Why is a TA climatology needed? 1. What does climatological TA distribution look like? 2. What processes make it look that way? 3. What does TA variability look like? 4. What processes make it vary like that? 5. How large are the uncertainties in the climatological values, and how does this uncertainty vary regionally and with depth? And finally, 6. How do we know the answers to these questions? The current paper begins to answer all of these questions, but ultimately falls back on too many qualitative and vague statements. It therefore ends up neither concise nor complete.

AR: Thank you so much for the thorough revision of the manuscript. We hope to improve it based on all the comments of the 3 referees. At the end of the answers is attached the new version of the manuscript for a global view.

We have clarified and added some of the multiple uses of a climatology of AT.

We have modified and included more aspects about the methodology in the creation of the neural network. We have written the key aspects to understand what we are doing but is not the scope of this manuscript to deepen neural networks. Several references about neural networks are given through the text for readers who want to immerse in the neural network world. This methodology section was divided in subsections.

The selection of predictors is based on bibliography to include all the variables related with AT variability. Moreover, with the new comparison with LIARv2 and CANYON-B, that include less predictors, we demonstrate the importance to include all the possible predictors related to AT variability to compute AT with a lower error.

We have added a comparison with the newest methods in AT computation (LIARv2 and CANYON-B). The comparison was done in the GLODAPv2 dataset used in our study, in a GLODAPv2 subset of the samples where AT QC was done and in the time-series data.

We keep the random selection of the sets because we think that is more important to capture more variability in both the training and the test sets to evaluate the complete range of variability in GLODAPv2 AT. A well-trained neural network is able to avoid fitting the error in the possible systematic errors along sections. It can be seen in the new test that we have done in the section on the subsurface layer hypothesis. In this test we remove all winter data, and therefore it is similar to the suggested approach in your comment. The error in this independent set, which includes data of the two hemispheres for three months and therefore complete oceanographic sections, is quite low and of the same magnitude when computed by the two networks (the NN and the new one).

Therefore, there are no significant differences in the results between our approach and the suggested one, but in ours more variability is been captured in the relations created and evaluated in our test. Answering to your comment: "when they deploy it in areas where there weren't nearby dense measurements collected at the time of the estimate", this test also shows the potential of the network in computing AT in areas like those you refer.

AC: These changes are all shown in the new manuscript that we attach at the end of the comments.

RC: Specific comments: L40. "The capacity of the ocean: : :" this sentence doesn't make sense. Are you suggesting the atmospheric pCO2 would today be 520 ppm without the ocean CO2 storage? This estimate is incorrect if so.

AR: No, we are not suggesting that the atmospheric pCO2 would today be 520 ppm. The sentence was poorly written. We wanted to say that the 30% of the anthropogenic emissions were absorbed by the ocean. We have rewritten the sentence.

AC: "The oceanic capacity to dissolve and store atmospheric CO_2, and the subsequent chemical speciation, have resulted in approximately 30% less anthropogenic CO_2 in the atmosphere (Le Quéré et al., 2017) than it would otherwise have."

RC: L51: This definition sounds closer to the Revelle factor definition.

AR: Changed

AC: "AT is a key variable in the framework of ocean acidification because of what it is associated: the oceanic capacity to buffer pH changes"

RC: L53: "Processes that change salinity…" it would be better to name those processes since one can imagine processes that change salinity without changing TA.

AR: They are named in the following two sentences. "… precipitation and evaporation…", "…rivers runoff…"

RC: L61: Hydrothermal TA inputs should perhaps also be mentioned.

AR: Added.

AC: "Finally, hydrothermal vents could modify the concentration of A_T locally (Chen, 2002).

RC: L66: "Therefore, the knowledge of AT variability over the global oceans at monthly timescales is very useful to increase the understanding of the ocean carbon cycle and to make assessments and projections related to ocean acidification with

greater rigor." Build on this. What applications specifically do you have in mind for this climatology? Why use the climatology instead of an algorithm?

AR: The main application and the motivation to design this climatology is its use in modeling studies (for example, coupling a circulation and a biogeochemical model). Therefore, a climatology is needed to have both initial and boundary conditions. To clarify the main application some was added to the end of the referred sentence.

AC: "A monthly A_T climatology that captures most of the spatiotemporal variability can be used as initial and/or boundary conditions in biogeochemical models, in evaluating the $CaCO_3$ pump (e.g., Carter et al., 2014) or computing the ocean inventory of anthropogenic CO_2 (e.g., Steinfieldt et al., 2009)."

RC: L88: Why? Why is it necessary to have a seasonal climatology of subsurface TA? How deep does seasonality affect TA, and how do you know this? See: line 190.

AR: Figure S6 shows the seasonality of AT at different depths. It is reduced with depth mainly because of the low or even absent seasonality in the input variables. However, there are other variables involved in AT variability that can change and are not included directly in our approach (e.g. CaCO3 formation and dissolution) but that could be explicitly represented through other variables that we included (position: latitude, longitude and depth). Analyzing time-series at depth, some seasonality is easily detected even at high depths. Therefore, it is important to have a seasonal climatology at least to certain depths. As there is no constant AT value with time in all locations of the ocean at any depth, it is not a logical to move from a seasonal climatology to an annual from a depth to another the continuity.

About Line 190. It has not been included in the new paragraph.

RC: L103-L122: This is in an unhappy medium of detail… too little detail to make any sense, and too much for the reader to quickly or confidently skip this. Either reference a paper that explains the method or fully explain it. My preference would be to fully explain the method, but move the text to a supplement so it doesn't interrupt the paper with too much detail… more detail in the text or a supplement would be necessary for the first paper if you split the paper into 2.

AR: The referred paragraphs follow the following scheme: 1) Lines 103-107. Description of the main elements of the neural network. 2) Lines 107-109. Operation of a neuron. 3) Lines 110-117. Explanation of the training procedure with enough clarifications to understand how a neural network works. References to know the mathematics of the the Levenberg-Marquardt and the backpropagation algorithms are given here. 4) Lines 118-122. Architecture of the neural network used in this study and why we selected it.

In other studies that use neural networks to fit variables of the seawater CO2 chemistry the description is even shorter (e.g. Velo et al. (2013); Sauzède et al. (2017)). Therefore, we think that we give enough information together with the references provided to

understand what a neural network does. To go deeper, the linear algebra behind neural networks should be explained. However, the reader can find it in the given references, although it is not necessary to understand what we are doing in our work. Therefore, we don't consider splitting the manuscript nor include many more details about neural networks because it would repeat what is already written in other studies. However, we have rearranged the methodology to include some more details about the creation and an operation of a neural network as well as more references. It can be seen in the new manuscript attached at the end of the comments.

RC: L140: Did you include calculated TA? TA where GLODAPv2 did not QC the data?

AR: Yes, we did. If there is some kind of error in these samples the neural network is not going to model it. Evidence of this can be seen in the good generalization of the network as it was shown through the good results in the test set. Other evidence is in the new test we have done comparing LIARv2, CANYON-B and NN in a GLODAPv2 subset excluding the samples where the QC was not done. Here, the statistics of our NN are better than the other two methods, which did not include non-QC samples in their "trainings".

RC: L159: "We kept: : :" I don't know what this sentence means.

AR: This paragraph has been changed for a better understanding.

AC: "The training procedure was carried out in MATLAB. We tested 16, 32, 64, 128 and 264 neurons in the hidden layer based on the results of Velo et al. (2013). For each number of neurons, we trained 10 networks always using the same 90% of GLODAPv2 for training (Fig. 2, Static level). The remaining 10% was used as a static test (Fig. 2, Static level). Both subsets contained samples randomly distributed in the ocean to evaluate the maximum possible relationships between the input variables and AT through all oceanographic regimes, that is, to capture most of the variability in all the variables and not restricting the sets to specific areas. Each of the 10 networks starts the training procedure with random weight and bias values and a random division of the training static dataset into three portions: 70% for training, 15% for testing and 15% for validation (Fig. 2, Dynamic level). These differences make minimization of the cost function different for each network due to the complexity of the weight-error space and, consequently, their different starting points in that space. As each network is different, keeping static sets allows one to determine which network best generalizes in the same test set. The selected network is the one that produces the lowest RMSE in the training data (validation + training dynamic) and in the test data (static + dynamic), considering a non-significant difference between both RMSEs to prevent overfitting. The network derived from this process will be referred as NNGv2."

RC: L162: What is the difference between testing and validation data sets? It's possible I missed the explanatory text, but consider trying to make that a bit clearer.

AR: Validation is used to finish the training process avoiding overfitting. Testing is not used in the training per se and, therefore, it is used as an independent set to compare different models. We have rewritten the paragraph to make it clearer.

AC: "Two different training techniques were tested: the Levenberg-Marquardt method (lm) and the Bayesian Regularization (br) (both detailed in Hagan et al., 2014). In a similar study, Velo et al. (2013) demonstrated that these techniques give the best network performance among those they tested. Except for the number of neurons, the two algorithms were implemented with the default options of the MATLAB functions trainlm and trainbr (detailed in Beale et al., 2017). These two functions prevent overfitting in different ways. The trainlm function usually needs to be fed with the data divided in three sets: a training set to obtain the relationships between variables, a validation set to prevent overfitting and a test set to compare different networks. Here, the training was stopped when the error in the validation set increased during 6 consecutive iterations of the training process to avoid overfitting. This process is known as early stopping (Hagan et al., 2014). The final values of the network weights and biases are those reached before the first of these iterations. The trainbr function adds a regularization parameter to the cost function to make the fit smoother in order to avoid overfitting. The validation set is not present in this technique. The end of the training is based on network convergence through parameter stabilization by an automatic process known as automated Bayesian Regularization (Hagan et al., 2014; Beale et al., 2017). See Beale et al. (2017) and references therein for a detailed description of the two functions tested."

RC: L168: Explain your rationale here. It is unclear why this test would find places where the network is unable to obtain accurate values.

What does it mean to for an individual data point to have a RMSE of >3? By my best guess, this is saying that the version of the NN that includes the data with a >3 absolute offset does better at fitting the data with a "less than 3" absolute offset than the version that only includes the "less than 3" data? The RMSE of the data with a "less than 3" offset, by definition, must be less than 3, and yet it climbs to 5.1 when you omit this data: : : so why bother with this analysis? Why not simply say "if we omit data with large errors our RMSE becomes small." Which do you recommend users adopt? Where and why?

AR: The approach here comes from the hypothesis that not all the processes are captured through the inputs that we used in our network. Therefore, some regions where specific processes occur could have high errors. This is what happens, for example, in the Arctic. So, if you represent in a map the samples with differences between the measured and the computed AT (Fig. 3; or Fig. S2 to see only the samples with the highest errors), you can see that the highest errors are mainly concentrated in specific regions.

The so-called 3σ rule is widely used for outlier detection and this was the main reason for choosing the criterion here (changing the standard deviation for the RMSE). A sample with an error greater than that threshold could mean two main things: 1)

Uncertainties in the measurements of any input and/or in AT; 2) Influence of other variables not included as input in the AT variability. Areas where a high concentration of samples with residuals greater than 3RMSE are included in option 2) (see discussion section: AT inputs from rivers). On the other hand, the scattered samples could fit into option 1).

3RMSE is equal to 24.6 μmol kg^{-1}. Therefore, removing the samples beyond this threshold from the error computation should not give an error of less than 3 μmol kg^{-1} as the reviewer suggests in the comment.

As neural networks try to fit a function with the provided data in the training set, removing the samples that do not fit well to the function created by the network is equal to removing their "bad" influence in this function. That is, in some way the network also tries to model the variability due to both non-included inputs or data with uncertainties. Therefore, the network trained without these samples could reach a more accurate function by not having samples that would need some other variable as input or uncertainties in any of the variables measured.

As you can see in Table 2 and Table 3, the RMSE is lower in almost all areas if NNw3RMSE is used to compute AT. Furthermore, the independent test in the time-series shows the same results (although they are not included in the MS). Therefore, based on these results, we offer the two networks to the scientific community to consider choosing one or the other according to the area of the study. Finally, we chose the NN to build the climatology to include the Arctic with the least possible error.

We have changed this paragraph to clarify this step.

AC: "As a last step we eliminated the data points with a difference between measured and computed A_T with the selected network (residuals) beyond ±3RMSE and then retrained the network as above. This procedure was used to identify regions where the network was unable to obtain accurate values and to improve the network mapping in the other areas omitting in this way data that the network could be trying to model without having the appropriate input variables or because they could be data with high measurement errors. Although a well-trained neural network avoids modeling the error, high errors could slightly modify the derived function in a negative manner. The network derived from this process will be referred as NN±3RMSE."

RC: L171: What does it mean to "illuminate the complexity" of a neural network? Be more specific.

AR: The meaning is explained after the comma: "to determine the contribution of each predictor variable in the output". Furthermore, is quite similar to the title of the referenced paper by Olden and Jackson (2002): "Illumination the black box…".

RC: L208: Random division of the datasets is inadequate for your test. See main points.

AR: Random division is meant to have more relations between input variables and A_T in the training set (greater coverage in the 4 dimensions: latitude, longitude, depth and

time). Therefore, the network is trying to model all the possible variability in the GLODAPv2 data and a random test set probably includes all this variability to evaluate it. In the new test that we have added to the section of the subsurface layer hypothesis deleting all the winter data in all the ocean you can see how the error in this deleted set is even lower than in the initial test set (6.8 vs 8.5 µmol kg^{-1}). This test also contributes to the reviewer's suggestion to leave a complete area without data to test the network there and the error is similar to that of NN. Also see the answer in the main points.

RC: L210: How does "The samples with residuals beyond _3RMSE are 1% of the global dataset: : :" align with "99% of the GLODAPv2 dataset used was modelled by the network with a root-mean-squared error (RMSE) of 5.1 _mol kg-1." I'm guessing this is referring to the 2nd NN, but I was confused for a long time before this statement started to make sense.

AR: "global dataset" is the same as "GLODAPv2 dataset used". The neural network is the first one: NN. It has been changed.

AC: "The samples with residuals beyond ±3RMSE are 1% of the global dataset" → "The samples with residuals beyond ±3RMSE are 1% of the GLODAPv2 dataset"

"99% of the GLODAPv2 dataset used was modelled by the network…" → "99% of the GLODAPv2 dataset used was modelled by the NNGv2…"

RC: L228: This sentence doesn't make sense to me… I suspect "disengage" is an incorrect word, but I'm not sure.

AR: Changed.

AC: "disengage" → "alter"

RC: L235: It's not always riverine TA that is the problem… often it is rivers with little or no TA that dilute seawater TA in a way that is distinct from the mixing patterns in the open ocean.

AR: The dilution of AT because of rivers with little or no AT should not be a problem. The dilution process is the same that in the precipitation process in open ocean and this is not a problem at all to the network.

RC: L235: this paragraph presents a weak argument against removing the region… fortunately, the argument for omitting that region was not made… omit this text, but instead estimate uncertainty regionally more quantitatively.

AR: 77% of the GLODAPv2 samples are well modeled in the Beaufort Sea and 91% in the North Sea. These high percentages suggest including the areas in the climatology is a good consideration. The regional error estimation is in Table 3.

RC: L251: “The zones defined..” this sentence is vague.

AR: In this paragraph we are discussing the results in Table 3, that is, AT computed with Takahashi et al. (2014) relations and AT computed with our neural networks. Therefore, the zones are those defined there. We have changed the sentence for a better understanding.

AC: “The A_T computed in the zones defined in the Arctic have higher RMSEs in the two approaches (Takahashi et al. (2014) and this study; Table 3)”

RC: L252: considerably

AR: Changed.

RC: L261: what is meant by this sentence? Clarify.

AR: Except in the Artic, NN3wRMSE computes AT with a lower error than NN in almost all the areas defined in Table 2 and Table 3. The sentence has been changed for a better understanding.

AC: “Although NN±3RMSE computes AT with lower errors than NNGv2 in the non-Arctic areas, in a global view the improvement is relatively small (Weighted RMSE in Table 2 and Table 3)”

RC: l269: what is meant by this sentence? Clarify.

AR: The formation and degradation of organic matter is reflected in oxygen and nutrients. We have changed the sentence by this one.

AC: “The formation and degradation of organic matter is reflected through both oxygen and nutrients variations.”

RC: L273: S and T collectively provide information regarding interior ocean density and mixing patterns, which are important for predicting TA: : : it is not clear what you are suggesting about the link between T and CaCO3.

AR: We are suggesting that the influence of temperature in both CaCO3 and organic matter cycles could make this variable a good proxy to capture the AT variability because of these processes by the network. We have referred the study by Lee et al. (2006) where this is said.

RC: L278: “The bias is relatively low in the three time-series with the highest number of data. The AT computed by the NN at KNOT and K2 is higher than the measured one.

Summed to the previous test, this independent test with a seasonal time resolution shows the good generalization of the NN." The first two sentences here are difficult to understand and the last one does not make sense to me.

AR: This paragraph describes the statistics in Table 4 to show the good behavior of the network in modeling independent data. The first referred sentence is describing the good performance in most of the data used as test here (HOT, BATS and ESTOC). The second one is to reflect that in both KNOT and K2 the network performance is not as good as in the other time-series. We have changed the text for a better understanding.

AC: "The bias is relatively low in the three time-series with the highest number of data (HOT, BATS and ESTOC). The A_T computed by the NNGv2 at KNOT and K2 is slightly higher than the measured one, probably because of the influence in the A_T variability of some variable not included as an input of the network (although an offset in the measurements of any of the inputs could also give this result). Summed to the previous test, the statistics obtained in this independent test with a good seasonal time resolution shows the good generalization of the NNGv2."

RC: L286: This logic does not follow.

AR: This paragraph has been deleted and a new one was added including the discussion about A_T computation using the code of both LIARv2 and CANYONB. See in the new version of the manuscript attached at the end of the comments.

RC: L302: This sentence does not seem to fit with the rest of the paragraph.

AR: This part has been changed. See in the new version of the manuscript attached at the end of the comments.

RC: L310-L313: I don't understand these sentences.

AR: The properties analyzed in the four areas defined in Fig. 2 show, as expected, a higher monthly variability in the ocean surface than in the subsurface layers. The intra-annual variability depicted in Fig. 7 is likely to be typical also of larger regions within each ocean basin. The surface conditions in winter of the analyzed properties are quite similar to those in the subsurface layer in four consecutive months to winter in all ocean basins (Fig. 7). Thus, using the subsurface layer allows to retrieve winter surface conditions in other seasons. See also in the new version.

RC: L310: intra-annual would be clearer as "seasonal"

AR: Changed.

RC: L340-342: I don't understand the logic.

AR: Changed.

AC: “This result shows the importance of the inclusion of other predictors besides salinity in the network and the non-linearity of the method proposed in this study to explain nearly all the A_T variability”

RC: L354: Isn’t that figure 10?

AR: Thanks, it is. Changed.

RC: L363: Why a change of grid for a time series?

AR: There is no time-series discussion in this section. Here, we are comparing climatologies. The comparison referred to in L363 is between the climatology of Takahashi et al. (2014) and ours. Their climatology is in a 5ºx4º grid and ours in a 1ºx1º one. Therefore, we averaged our climatology to their grid which in some way leads to an error component in the comparison between the climatologies. Finally, the sentence has been changed.

AC: “Furthermore, the transfer of our climatology to the coarser grid of Takahashi et al. (2014) for the comparisons may enhance dissimilarities.”

RC: L380: what is meant by a “continuity” in the differences?

AR: The largest differences in the referred comparison are quite patchy except in the two referred zones. We have changed the sentence for a better understanding.

AC: “there is a large continuity in the differences” → “there is a large spatial continuity in the differences”

RC: L393: I couldn’t get the code to work: : : some info: : : â˘A ´c Tried with the directory with NN as the active directory o As well as just with the directory with NN on the Matlab path â˘A ´c Tried entering within a script and on the command line â˘A ´c Tried with Matlab r2014b with the NN toolbox (also 2018b, but without the NN toolbox) â˘A ´c With inputs as single or double precision numbers â˘A ´c Entered: o AT_values=Neural_network_object(data_inputs); o AT_values=NN(data_inputs); o AT_values=NN_w3RMSE(data_inputs); Invariably got the response: “Undefined function ’NNw3RMSE’/’NN’/’Neural_network_object’ for input arguments of type ’single’/’ double’.” I recommend making the instructions a bit more clear. I also recommend adding an example calculation so users can be sure they are getting the expected answers. You’ll know you are done when the coauthors can use the function reliably without additional guidelines.

AR: The code is only one line. Only the neural network object and the input matrix in the form explained in the readme.txt file are needed. We have attached a matrix of

random inputs from GLODAPv2 and a very simple script. It has been successfully used by other colleagues. We have added this material on the same site as previously.

RC: Figure 7 and elsewhere: Uppercase theta should be reserved for conservative temperature… use lowercase.

AR: Changed.

RC: Figure 8 and elsewhere: The font is too small.

AR: Changed.

RC: Figure 10, right panel: How is the RMSE smaller than the bias? Clarify what is being shown here if there is a good reason.

AR: The measured seasonal variability is well represented by the climatology (low RMSE) but there is an offset between the values. The offset is because of the difference in the data inputs (monthly average of the measured data in the time-series vs. WOA13). Passing the measured data through the network, the bias is lower and positive (0.9 μmol kg^{-1}). This fact demonstrates that the network is not the cause of the relatively high bias of the other mentioned comparison. Therefore, Figure 10 was made for showing the good performance of the network and the differences in the climatological data from WOA13 and those obtained from averaging the measured data at time-series locations (the latter of which are the only locations where the neural network climatology can be properly compared with climatological measured data).

RC: Referee comment; **AR**: author's response; **AC**: author's changes in manuscript.

Referee #3 response

RC: The authors describe a neural network approach to derive an algorithm to estimate AT from concurrent Lat/Lon, depth, T/S, oxygen and nutrient (nitrate, silicate and phosphate) inputs, based on GLODAPv2 data. They use this approach with monthly climatological fields from WOA13 to establish a global, depth-resolved, monthly AT climatology. The manuscript is clearly-structured and well-written.

I see three critical points, (1) the neural network topology selection and the second round of neural network training without control for overfitting, (2) the adequate representation of (surface) seasonality in the training data, by the neural networks, and the derived monthly climatology, and (3) the placement and comparison with other recent work on AT estimation based on GLODAPv2-trained algorithms. I therefore suggest major revisions to the manuscript.

AR: Thank you so much for the thorough revision of the manuscript. We hope to improve it based on all the comments of the 3 referees. At the end of the answers is attached the new version of the manuscript for a global view.

RC: Major points:

(1a) The authors describe training of their neural networks in general terms, however, some important details remain missing.

• The selection of the best performing neural network appears subjective and is not made transparent. This needs to be improved. E.g., l.161: What criterion has been used to assess "best generalization in the initial testing dataset"?; l. 204f: 128 neurons kind of fall from the sky. Figure S1 would probably be more instructive to show RMSE for training and testing set vs. number of neurons, to make the authors' reasoning more transparent.

AR: The selection of the best neural network is based on the RMSE in the test set. We have added it in L.161, although this whole paragraph has been modified also considering the comments of one of the other referees.

The selection of the optimal number of neurons is based on the explanation given in L-142-150. However, we have added the suggested Figure S1.

AC: "The training procedure was carried out in MATLAB. We tested 16, 32, 64, 128 and 264 neurons in the hidden layer based on the results of Velo et al. (2013). For each number of neurons, we trained 10 networks always using the same 90% of GLODAPv2 for training (Fig. 2, Static level). The remaining 10% was used as a static test (Fig. 2, Static level). Both subsets contained samples randomly distributed in the ocean to evaluate the maximum possible relationships between the input variables and AT through all oceanographic regimes, that is, to capture most of the variability in all the

variables and not restricting the sets to specific areas. Each of the 10 networks starts the training procedure with random weight and bias values and a random division of the training static dataset into three portions: 70% for training, 15% for testing and 15% for validation (Fig. 2, Dynamic level). These differences make minimization of the cost function different for each network due to the complexity of the weight-error space and, consequently, their different starting points in that space. As each network is different, keeping static sets allows one to determine which network best generalizes in the same test set. The selected network is the one that produces the lowest RMSE in the training data (validation + training dynamic) and in the test data (static + dynamic), considering a non-significant difference between both RMSEs to prevent overfitting. The network derived from this process will be referred as NNGv2.

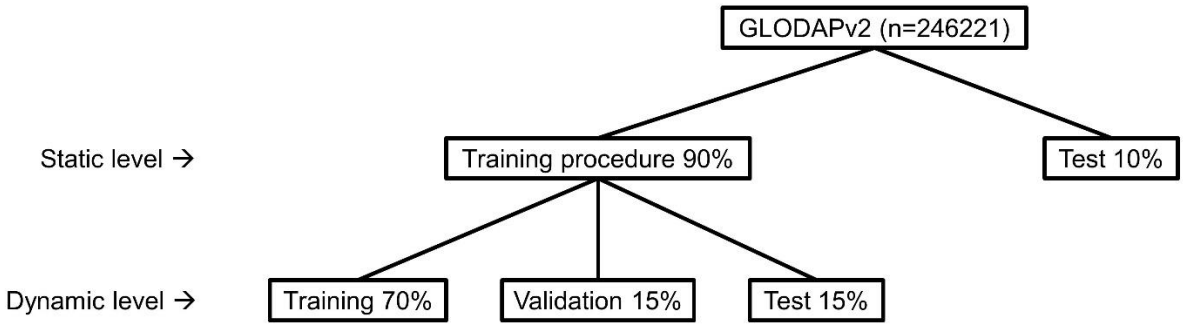


Figure 2. Division of the data for the training of the network. The data in the sets of the static level is the same for all the networks to train. The data in the sets of the dynamic level is randomly selected for each network to train.

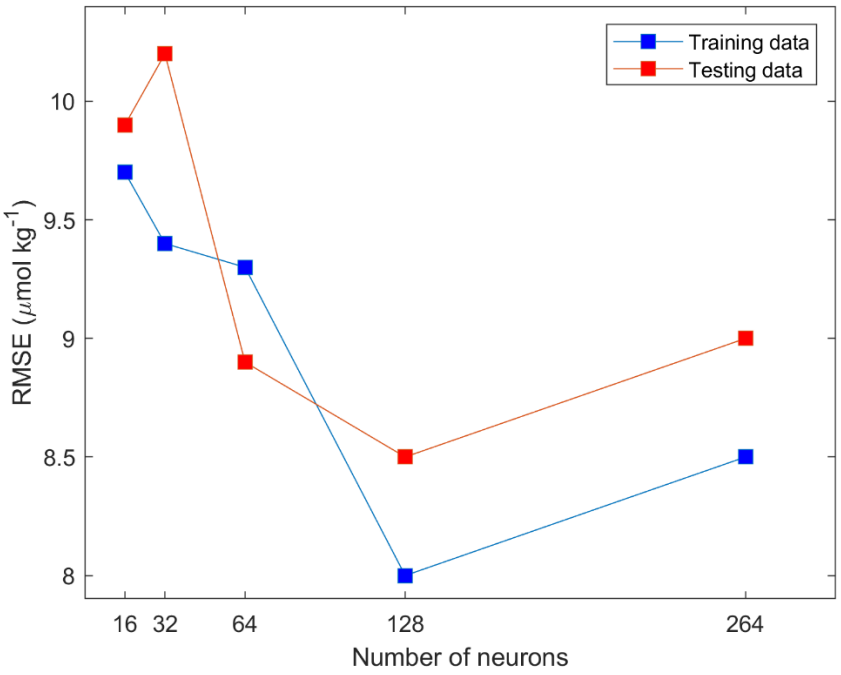


Figure S1. RMSE variation with the number of neurons of the network for the lm algorithm. Training data contain the dynamic training set and the dynamic validation set, that is, 76.5% of the GLODAPv2 dataset used in this study. Testing

data contain both the static and the dynamic test sets, that is, 23.5% of the GLODAPv2 dataset used in this study."

RC: • Do the authors use weight regularization of the network weights? I presume so, at least for the Levenberg-Marquardt variants but probably also for their Bayesian regularization. This should be stated. It should be stated as well how the regularization (hyper-)parameter/weight was chosen (i.e., the balance between data accuracy/loss and weight penalty/loss terms in the cost function; or in other words the balance between accuracy and generalization behavior within the given network topology).

• What exactly is meant by Bayesian Regularization (l. 141 with reference to MacKay, 1992)? Please be more explicit here. If you used a certain, e.g., Matlab implementation/toolbox, make reference to it. MacKay (1992) describes at least three levels of Bayesian treatment, from (I) finding the 'best' (most-probable) set of weight parameters including their regularization (i.e., preserving generalization behavior by avoiding too specialized weight distributions) through (II) finding the 'best' hyperparameter values (i.e., objectively assigning the balance between data loss and complexity/regularization) to (III) model comparison (e.g., quantitatively rank different models or neural network topologies). It seems to me that only (I) has been used here? Please clarify. Also note that, if implemented correctly (!), Bayesian regularization doesn't need cross-validation like, e.g., a backpropagation Levenberg-Marquardt learning scheme.

AR: We have added a more extensive description of the two techniques used in this study and referred to the MATLAB toolbox used, as the reviewer suggested. However, we did not add a full description explaining the math and all the processes behind the different algorithms inside the neural networks. We have added the necessary references for the readers who are looking for more detail. The scope of this manuscript is to explain the basics to create a basis for the readers to understand how we achieve the objectives.

AC: "Two different training techniques were tested: the Levenberg-Marquardt method (lm) and the Bayesian Regularization (br) (both detailed in Hagan et al., 2014). In a similar study, Velo et al. (2013) demonstrated that these techniques give the best network performance among those they tested. Except for the number of neurons, the two algorithms were implemented with the default options of the MATLAB functions trainlm and trainbr (detailed in Beale et al., 2017). These two functions prevent overfitting in different ways. The trainlm function usually needs to be fed with the data divided in three sets: a training set to obtain the relationships between variables, a validation set to prevent overfitting and a test set to compare different networks. Here, the training was stopped when the error in the validation set increased during 6 consecutive iterations of the training process to avoid overfitting. This process is known as early stopping (Hagan et al., 2014). The final values of the network weights and biases are those reached before the first of these iterations. The trainbr function adds a regularization parameter to the cost function to make the fit smoother in order to avoid overfitting. The validation set is not present in this technique. The end of the training is

based on network convergence through parameter stabilization by an automatic process known as automated Bayesian Regularization (Hagan et al., 2014; Beale et al., 2017). See Beale et al. (2017) and references therein for a detailed description of the two functions tested."

RC: (1b) I think the two-step training of the networks with elimination of the testing data must be avoided (with a backpropagation/LM algorithm). Optimization of the network's parameters doesn't stop after training with the 70/15/15 % training/validation/testing data set. It continues well throughout the 80/20/0 % step, where the authors no longer have control over or means to assess overfitting. The authors' conclusion (l.165) is invalid. Given that, e.g., "[the authors] find no improvement by increasing the amount of data points in the training set" (l. 207), I don't see the point in making this questionable second step. Instead, this re-optimization of weights without control for overfitting makes the method vulnerable. It should be removed thus closing this open flank without loss in performance.

AR: It was a mistake. We deleted the set which is used to validate the model (named as test set in MATLAB sets division). We have changed this paragraph. The validation set which is used to prevent the overfitting is maintained. The static test set (named in this way in the new version of the manuscript) is maintained to choose the best network. We deleted the dynamic test set. We make this step to test if to have more data to train a neural network would improve the fitting. The results suggest that no more data is necessary and maybe include other input variables would improve the fitting.

AC: "Once we found an adequate network configuration, we increased the amount of data in the training dynamic set to capture more relations between the inputs and AT. The new percentages of the dynamic sets were: 80% training, 20% validation and 0% testing. The latter set is only necessary to compare different models and is not used during the training. However, the static test set was held to evaluate the generalization of each of the 10 networks to select the best one."

RC: I do see the NNw3RMSE run critical, too. In essence, the authors level out areas of the ocean with higher-than-average variability (>3* global-mean-RMSE *samples* are removed, i.e., only the subset of samples that fit to the mean in these areas is retained). They do this to "improve the network mapping in the other areas" (l. 169). This *spatial* difference/distinction should be captured by the sampling position input (Lat, sLon, cLon, Depth), shouldn't it? I would argue that the (small) improvements they see in certain subregions between the NN and the NNw3RMSE runs is only due to a different local minimum found during neural network training of the one neural network selected for NN vs. the one neural network selected for NNw3RMSE, and not thanks to the omission of data in an at most adjacent or even unrelated ocean region (e.g., Equatorial Upwelling Pacific, while most samples >3*global-mean-RMSE are found at high latitudes/North Sea). Again, given that "the difference in the weighted RMSE of the two networks [NN and NNw3RMSE] is not significant" (l. 247; l.254) and that the

authors consider NN the best candidate for users (l. 263), I'd suggest to drop the NNw3RMSE network.

AR: Most of the data deleted for the retrained network (NNw3RMSE) are from the Beaufort Sea. In Table 3 you can see how big the difference is between the errors computed by the two networks. This result suggests that the NN is obviously trying to fit the data of this area. In this case, it is clear that omitting certain data causes a large difference between the networks. Although the improvements are not as high in many of the areas defined in Table 3, these improvements in almost all of the zones suggest that they are because of this data deletion instead than the different local minimum reached in the error function. The NN training tries to fit these data in some way and it could be fitting a function not good enough to fit the data of the areas that are easier to model as the NNw3RMSE does. However, the suggestion that the reviewer makes could also be a possibility and thus we have included it in the discussion.

We decided to keep the NNw3RMSE based on the suggestion made previously about the reader's ability to choose one or another network regardless of the reason for improvement in certain areas. Moreover, it has been checked in a new application designed for QC (https://github.com/ocean-data-qc/ocean-data-qc) as this network computes better AT in recent data not included in GLODAPv2 than NN or CANYON-B.

AC: "As a last step we eliminated the data points with a difference between measured and computed A_T with the selected network (residuals) beyond ±3RMSE and then retrained the network as above. This procedure was used to identify regions where the network was unable to obtain accurate values and to improve the network mapping in the other areas omitting in this way data that the network could be trying to model without having the appropriate input variables or because they could be data with high measurement errors. Although a well-trained neural network avoids modeling the error, high errors could slightly modify the derived function in a negative manner. The network derived from this process will be referred as NN±3RMSE."

RC: (2) I think it is courageous to derive a monthly-resolved product from GLODAPv2 data, which in many ocean regions is far from being monthly yet seasonally-resolved. This needs further elaboration and the seasonal character needs to be demonstrated clearly.

To tackle the scarcity of winter time observations, the authors state that the lack of surface information during winter can be circumvented by using spring time observations of subsurface waters that retain the winter water signature, illustrated by figure 7. (l. 187-194).

Fair enough, but this information doesn't tell the neural network to learn it that way nor does it imply by any means that the neural network recognizes this connection. Even if winter water properties are similar between spring subsurface and winter surface samples (as in the climatological WOA13 data of figure 7), the vertical sampling location (Depth) is still different, thus ending up in a different area of the neural network input data space - giving potentially very different AT output.

The first step to convince me of this 'seasonal winter gap filling' would be to add the predicted surface and subsurface AT to figure 7 - which should approach each other during winter like the water properties.

A second step would be to give better quantification of the seasonal cycle where possible. This is probably limited to the time series stations and the North Atlantic. If the training data are seasonally well-resolved and the neural network training picks up this seasonality adequately, the seasonal cycle's amplitude from the NN (and measured inputs) should be of the same magnitude as the observed seasonal cycle's amplitude. If the training data do not reflect full seasonality, the NN tend to underestimate the seasonal cycle - with a flat line as the extreme. Such a comparison should complement the for now only qualitative assessments (e.g., figure 5).

Moreover, the (sub-)polar North Atlantic should be added as region for the sub-surface hypothesis due to the high interest in the carbon cycle in this area.

AR: To show the potential of the neural network to obtain accurate values in winter times we have added a new test in the subsurface layer hypothesis section. We trained a new network without the winter data in the same way as the original network. If the network is able to capture the winter relations between inputs and AT from other seasons, the excluded winter dataset from the training should be model by the network. The good statistics show how the network obtains the winter relations for all the ocean areas where the winter data was measured in our GLODAPv2 dataset without the need to be trained with winter samples.

In a very complex error space, with more than 1000 dimensions like the one resulted from our NN, it is very difficult to know how any input variable is related with the others and with the output variable. Therefore, "depth" (and the other input variables) may not be assimilated by the network in a way that we may understand. What is clear with this new test is that the network is able to model winter data even if it does not have such winter data for training.

The subpolar North Atlantic has been evaluated by Vázquez-Rodríguez et al. (2012) as we stated in the manuscript and therefore, we evaluated other zones within the WOA13 data. Nevertheless, this new test has assessed an extensive coverage of the ocean. It can be seen in the new version of the manuscript attached at the end of the answers.

About the North Atlantic time-series that the reviewer refers to (we suppose Irminger and Iceland), we could not include them in the study because no AT was measured there. However, we have evaluated the network performance in these regions with AT computed from pCO2 and DIC (although the temperature at which pCO2 was measured is not clear in the dataset we downloaded from CDIAC) and the statistics are similar to those of the time-series included in this study, with a seasonal cycle amplitude similar to that of the computed AT from measured pCO2 and DIC.

RC: (3) Since the publication of the GLODAPv2 data set, there have been other works that use the data compilation to establish algorithms for AT estimation. Two of them are mentioned (LIARv1, Carter et al. 2016, and CANYON, Sauzede et al. 2017), however,

the manuscript falls short on setting their own work into perspective of the state-of-the art published literature.

(a) Both methods mentioned have received updates (LIRv2, Carter et al. 2018, and CANYON-B, Bittig et al. 2018), to which the comparison of the present work should be made. Both updated algorithms are publicly available as Matlab code and use overlapping (but fewer) inputs as the authors' approach, i.e., there is no obstacle to apply them to any of the authors' data.

(b) The authors already do a decent job in assessing their work with surface-only climatologies (e.g., Lee et al. 2006), but the authors need to demonstrate more clearly how the present work improves / compares with existing, global, depth-resolved algorithms of AT (e.g., see above).

E.g., in terms of accuracy on all their time series data, not just HOT (section 3.2), surface seasonal amplitude (see point 2 above), complexity in terms of input data requirements, etc.

Interestingly, the authors don't use the year day as input either (same as LIR and CANYON-B), and nonetheless get good surface seasonality.

This point (3) is important to improve, since it will give the authors the argument of why use their algorithm (or one of the others) to derive an AT climatology from WOA13 fields, which is the main subject of this work (following the title).

AR: We have added a general comparison with the two updates of the referred methods. We had only focused on the algorithms of Lee et al. (2006) and Takahashi et al. (2014) since they were created to generate a monthly climatology, as in this study.

First, we showed in our GLODAPv2 dataset how our NN fits the AT better than LIARv2 and CANYON-B. Furthermore, we compute the statistics excluding the data were the QC was not made since both referred methods did it to train their algorithms.

AC: In 2. Methodology: "The recent methods to compute A_T proposed by Carter et al. (2018) and Bittig et al. (2018) (LIARv2 and CANYON-B respectively) were also compared to the one proposed here. LIARv2 is based on multilinear regressions (MLRs) including the same predictors used in the present study, excluding phosphate (sample position, salinity (S), potential temperature (θ), nitrate (N), apparent oxygen utilization (AOU) and silicate (Si)). This method is composed of 16 equations with a different combination of the input variables, always maintaining the salinity input in each one. The computations with LIARv2 were obtained by the equation with the lowest uncertainty estimate in each sample that this method determines (Carter et al., 2018). CANYON-B is based on a Bayesian neural network derived from GLODAPv2 data including position, time, salinity, temperature and dissolved oxygen as predictors. The two methods were applied on the GLODAPv2 dataset used here and the on a subset excluding the samples where the quality control (QC) of A_T was not done (QC procedures detailed in Olsen et al. (2016) and references therein)."

In 3.1 Neural network analysis "The newest methods in the AT computation (LIARv2: Carter et al., 2018; CANYON-B: Bittig et al., 2018) model the GLODAPv2 AT with higher errors than the NNGv2 (Table 4). An analysis in a GLODAPv2 subset excluding

the samples where the 2nd (Olsen et al., 2016) QC was not done for AT shows a reduction of the error in these three methods, being CANYON and NNGv2 the lowest (Table 4). All the equations are used to compute AT in the GLODAPv2 dataset when the computation is allowed to be made by the equation with the lowest uncertainty in each sample (Carter et al., 2018). The most used equations are 10 (S, N, Si), 15 (S, AOU) and 14 (S, N), which are used in about 50% of the samples. The equation that used all the input variables (1) is only used to model 3% of the GLODAPv2 samples. Surprisingly, when only this equation is used to compute AT in GLODAPv2 dataset, the error is lower than those obtained with the free election of the equation based on the lowest uncertainty. That result shows the potential of include all possible inputs related with the AT variability, although reasonable results can also be reached with the equations that do not use all the input variables. CANYON-B is an example of using relatively few input variables (position, time, temperature, salinity and oxygen) and getting good results (Table 4). Probably, the non-linear character of the neural networks, like the one used in CANYON-B, gives the high potential to this kind of methods to fit complex functions even with few input variables. However, the NNGv2 designed in the present study is the best option to model more GLODAPv2 data better than the other methods (lower RMSE) and therefore to use the mapped inputs-output relation in order to create the monthly climatology. The availability of all the variables used as inputs of the NNGv2 in WOA13 also contributes to make this method the best choice. Furthermore, methods like CANYON-B which include a predictor that explicitly accounts for the time variation of AT (decimal year in the case of CANYON-B), are not suitable to build a monthly climatology since they generate an unrealistic seasonal amplitude, at least at high depths. This has been checked used WOA13 monthly climatologies (temperature, salinity and dissolved oxygen) as inputs of CANYON-B to compute AT at different depth layers. As an example, in the 3000m depth layer, seasonal amplitudes up to 40 µmol kg-1 were obtained in large areas mainly located between 30 and 60ºS."

In 3.2 Time-series validation "The LIARv2 and CANYON-B methods to compute AT also model the time-series data quite well (Table 6). Significant differences among the three methods are obtained in HOT and ESTOC. In HOT, NNGv2 and CANYON-B reach a better fit of AT than LIARv2 suggesting that a non-linear technique is more adequately to model AT in this area (Table 6). In ESTOC, NNGv2 and LIARv2 are the best options to model the AT variability (Table 6). Here, the AT computed with LIARv2 with the option of the free equation choice activated results in a greater election of the equations that include nutrients as predictors. This result show how in this area the inclusion of nutrients as predictors contributes to improve the model of AT. Like NNGv2, both methods have a considerable bias in K2 and KNOT (Table 6) that reinforce the two reasons suggested previously."

Table 4. RMSE and bias obtained with NN, LIARv2 (Carter et al., 2018) and CANYON-B (Bittig et al., 2018) in both GLODAPv2 dataset and GLODAPv2 dataset without the samples where AT QC was not done.

	RMSE	bias	
Approach	(µmol kg^{-1})	(µmol kg^{-1})	n
NN_GLODAPv2	8.2	0.02	246221

LIARv2_GLODAPv2	11.4	0.08	246221
CANYON-B_GLODAPv2	11.4	-2.8	246221
NN_GLODAPv2_onlyQC	6.6	0.06	215332
LIARv2_GLODAPv2_onlyQC	8.2	0.06	215332
CANYON-B_GLODAPv2_onlyQC	7.8	-3.2	215332

Table 6: RMSE and bias between measured A_T and the A_T computed with both LIARv2 and CANYON-B methods. The comparison was done for the same samples evaluated in Table 5.

	LIARv2		CANYON-B	
Time-Series	RMSE (µmol kg^{-1})	bias (µmol kg^{-1})	RMSE (µmol kg^{-1})	bias (µmol kg^{-1})
HOT	**6.6**	**-0.6**	**5.8**	**-0.6**
BATS	**6.3**	**0.1**	**6**	**-0.4**
ESTOC	**3.4**	**0.8**	**4.2**	**3.2**
KNOT	**4.8**	**-6.6**	**4.5**	**-7.2**
K2	**3**	**-3.0**	**3**	**-3.3**

Minor points:

RC: l. 60: remove oxygen. Nutrient changes contribute to a change in AT, oxygen itself does not contribute to the charge/acid/base balance.

AR: Changed

AC: "(Millero et al., 1998; Fine et al., 2017). Organic matter cycling can also contribute to AT changes. This mechanism can be reflected through the consumption and regeneration of nutrients and oxygen (Brewer and Goldman, 1976; Wolf-Gladrow et al., 2007)."

RC: l. 105: The number of neurons in the output layer is adjustable? It seems to be n=1 for just AT, isn't it?

AR: Changed

AC: "The hidden and output layers are composed of neurons. The number of these elements in the hidden layers is adjustable and in the output layer is dependent on the number of network outputs. The neurons are formed by a series of weights, a bias, a summation, and a transfer function (Russell and Norvig, 2010). They are the connections between the layers."

RC: l. 124: "as previously described" - not yet done, remove.

AR: Removed.

RC: l. 138 and 139: Which spurious oxygen value was removed / Where can it be found? (To allow reproduction by others.); Name the ocean time-series or give their GLODAPv2 cruise IDs.

AR: Added cruise, station and bottle of the sample with spurious oxygen value. Second sentence has been deleted and it has been commented in L. 182 (in the first version of the manuscript).

AC: "From these, we removed one record due to its spurious oxygen value (O2=1026.9 µmol kg^{-1} cruise=102; station=4; bottle=5)."

RC: l. 238: "As an argument ... areas." Unclear.

AR: Changed.

RC: l. 273: Depth is rather associated as vertical sampling position.

AR: Of course, it is. But, together with temperature and salinity it can help to model the variability of AT because of CaCO3 cycle.

RC: l. 279: Any ideas why there is such a bias? Should be commented.

AR: Added.

AC: "The AT computed by the NNGv2 at KNOT and K2 is slightly higher than the measured one, probably because of the influence in the AT variability of some variable not included as an input of the network (although an offset in the measurements of any of the inputs could also give this result)."

"… Like NNGv2, both methods have a considerable bias in K2 and KNOT (Table 6) that reinforce the two reasons suggested previously."

A global monthly climatology of total alkalinity: a neural network approach

Daniel Broullón[1], Fiz F. Pérez[1], Antón Velo[1], Mario Hoppema[2], Are Olsen[3], Taro Takahashi[4], Robert M. Key[5], Toste Tanhua[6], Melchor González-Dávila[7], Emil Jeansson[8], Alex Kozyr[9] and Steven M.A.C. van Heuven[10]

[1]Instituto de Investigaciones Marinas, CSIC, Eduardo Cabello 6, 36208 Vigo, Spain

[2]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Postfach 120161, 27515 Bremerhaven, Germany

[3]Geophysical Institute, University of Bergen and Bjerknes Centre for Climate Research, Allègaten 70, 5007 Bergen, Norway

[4]Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY 10964, USA

[5]Atmospheric and Oceanic Sciences, Princeton University, 300 Forrestal Road, Sayre Hall, Princeton, NJ 08544, USA

[6]GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20D-24105 Kiel, Germany

[7]Instituto de Oceanografía y Cambio Global, IOCAG, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

[8]Uni Research Climate, Bjerknes Centre for Climate Research, Jahnebakken 5, 5007 Bergen, Norway

[9]NOAA National Centers for Environmental Information, 1315 East-West Hwy Silver Spring, MD 20910 USA

[10]Faculty of Science and Engineering, Isotope Research – Energy and Sustainability Research Institute Groningen, University of Groningen, Nijenborgh 6, 9747 AG Groningen, The Netherlands

Correspondence to: Daniel Broullón (dbroullon@iim.csic.es)

Abstract. Global climatologies of the seawater CO_2 chemistry variables are necessary to assess the marine carbon cycle in depth. The climatologies should adequately capture seasonal variability to properly address ocean acidification and similar issues related to the carbon cycle. Total alkalinity (A_T) is one variable of the seawater CO_2 chemistry system involved in ocean acidification and frequently measured. We used the Global Ocean Data Analysis Project version 2 (GLODAPv2) to extract relationships among the drivers of the A_T variability and A_T concentration using a neural network (NNGv2) to generate a monthly climatology. 99% of the GLODAPv2 dataset used was modelled by the NNGv2 with a root-mean-squared error

(RMSE) of 5.1 µmol kg^{-1}. Validation tests with independent datasets revealed the good generalization of the network. Data from five ocean time-series stations showed an acceptable RMSE range of 3.1-6.2 µmol kg^{-1}. Successful modeling of the monthly A_T variability in the time-series suggests that the NNGv2 is a good candidate to generate a monthly climatology. The monthly climatological fields of AT were obtained passing World Ocean Atlas 2013 (WOA13) monthly climatologies through the NNGv2. The spatiotemporal resolution is set by WOA13: 1ºx1º in the horizontal, 102 depth levels (0-5500m) in the vertical, and monthly temporal resolution. The product is distributed through the data repository of the Spanish National Research Council (CSIC; doi: http://dx.doi.org/10.20350/digitalCSIC/8564).

1 Introduction

Because of its interaction with the atmospheric carbon dioxide, the marine carbon cycle has fundamental significance for the Earth's climate (Tanhua et al., 2013). The oceanic capacity to dissolve and store atmospheric CO_2, and the subsequent chemical speciation, have resulted in approximately 30% less anthropogenic CO_2 in the atmosphere (Le Quéré et al., 2017) than it would otherwise have. One unfortunate byproduct of this process is ocean acidification (Doney et al., 2009). As the ocean absorbs anthropogenic CO_2, the seawater pH decreases being the main change in the ocean chemistry which defines ocean acidification. Combined with other climate change effects (e.g., temperature increase and deoxygenation), this process could have severe consequences for marine ecosystems (Orr et al., 2005; Fabry et al., 2008; Hoegh-Guldberg and Bruno, 2010; Kroeker et al., 2013) and, consequently, for life on our planet.

Detailed spatiotemporal knowledge about the marine carbon cycle is necessary to understand and evaluate the consequences of climate change. There are 4 variables of the seawater CO_2 chemistry more frequently measured in carbon chemistry campaigns: total alkalinity (A_T), total dissolved inorganic carbon (TCO_2, also known as DIC), partial pressure of CO_2 (pCO_2) and pH. A_T is a key variable in the framework of ocean acidification because of what it is associated: the oceanic capacity to buffer pH changes. Dickson (1981) defined A_T as:

$$A_T = [HCO_3^-] + 2[CO_3^{2-}] + [B(OH)_4^-] + [OH^-] + [HPO_4^{2-}] + 2[PO_4^{3-}] + [SiO(OH)_3^-] + [HS^-] + 2[S^{2-}] + [NH_3] - [H^+] - [HSO_4^-] - [HF] - [H_3PO_4] \quad (1)$$

The global A_T distribution is a result of physical and biogeochemical processes that change the concentration of species in Eq. (1) (Wolf-Gladrow et al., 2007). Processes that change salinity are the most influential. The strong linear correlation between salinity and A_T is well documented (e.g. Millero et al., 1998; Friis et al., 2013; Takahashi et al., 2014). In the surface layer precipitation and evaporation are the primary processes that control the A_T distribution. Rivers and submarine groundwater discharge can affect marine A_T locally, with the degree controlled by runoff and the riverine A_T (Hoppema, 1990; Anderson, 2004; Schneider et al., 2007; Cooper et al., 2008). The formation and dissolution of carbonate minerals also contribute to A_T variability (Fry et al., 2015). Upwelling areas that overlie zones of relatively shallow subsurface carbonate dissolution can also

have elevated surface A_T (Millero et al., 1998; Fine et al., 2017). Organic matter cycling can also contribute to A_T changes. This mechanism can be reflected through the consumption and regeneration of nutrients and oxygen (Brewer and Goldman, 1976; Wolf-Gladrow et al., 2007). Finally, hydrothermal vents could modify the concentration of A_T locally (Chen, 2002).

In addition to the spatial variability, most of the drivers mentioned above generate seasonal A_T variability. Phytoplankton blooms (i.e., primary production) and the seasonality in upwelling and river flows are some of the more remarkable processes associated with the time variability of A_T. Even though A_T is the variable of the seawater CO_2 chemistry system with the least seasonal variability (Lee et al. (2006) estimated a range from near 0 up to 80 μmol kg^{-1}), it is important to account for such changes because of the strong connection of A_T oceanic anthropogenic carbon storage (Renforth and Henderson, 2017) and to buffer seawater pH changes. A monthly A_T climatology that captures most of the spatiotemporal variability can be used as initial and/or boundary conditions in biogeochemical models, in evaluating the $CaCO_3$ pump (e.g., Carter et al., 2014) or computing the ocean inventory of anthropogenic CO_2 (e.g., Steinfieldt et al., 2009).

High-quality data is a crucial first requirement to address the problem. Ocean time-series data represent excellent records to study the seasonality of the ocean carbon cycle as well as its inter-annual trends (e.g., Bates et al., 2014). Unfortunately, there are only a few time-series that include sufficiently precise measurements of the seawater CO_2 chemistry at seasonal resolution. Alternately, various global data products have been released for public usage in recent years. The main ones for the surface ocean are the Surface Ocean CO_2 Atlas (SOCAT; Bakker et al., 2016) and the Lamont-Doherty Earth Observatory database (LDEO; Takahashi et al., 2016). These two are complementary, offer annual updates and include tens of millions of pCO_2 measurements in the global ocean. For the interior ocean, a comprehensive and global database and data product was recently made public: Global Ocean Data Analysis Project version 2 (GLODAPv2) (Key et al., 2015; Olsen et al., 2016). This quality-controlled collection contains thousands of measured seawater data, including CO_2 chemistry variables, over the full water column from more than 700 globally distributed cruises over the past four decades.

The logical next step is to generate a globally consistent climatology for the different variables that captures seasonal variability. Different approaches have been used to fill spatial and temporal gaps in A_T observations to generate a global seasonal climatology (Lee et al., 2006; Takahashi et al., 2014). These studies only cover the surface ocean. However, a robust climatology of the entire water column is necessary to assess more than surface ocean.

In this study, we present a global monthly climatology for A_T in a 1ºx1º grid in the upper 102 standard depth levels (between 0 and 5500m) of the World Ocean Atlas 2013 (WOA13) designed using a neural network approach. Other studies have demonstrated the capacity of these techniques to reconstruct global pCO_2 variability at monthly resolution over the last few decades (e.g., Landschützer et al., 2013, 2014). Our A_T climatology uses available high-quality measurements and the neural network ability to capture natural variability. We were able to reduce the errors obtained by the previous efforts to build a seasonal A_T climatology (Lee et al., 2006; Takahashi et al., 2014) and to extend the climatology through the water column.

2 Methodology

2.1 Neural network design

A feed-forward neural network was configured to compute A_T globally at monthly resolution. It was selected based on the ability to learn the relationships between A_T and the variables related to its spatiotemporal variability as shown in Velo et al. (2013).

Feed-forward neural networks are composed of layers: the input layer, a variable number of hidden layers and the output layer (Fig. 1). The input layer is a matrix representing the entry to the network of the data from which the outputs will be obtained. The hidden and output layers are composed of neurons. The number of these elements in the hidden layers is adjustable and in the output layer is dependent on the number of network outputs. The neurons are formed by a series of weights, a bias, a summation, and a transfer function (Russell and Norvig, 2010). They are the connections between the layers. A neuron receives all outputs from the previous layer and multiplies them by a matrix of weights. These results are summed and a bias is added. Finally, the transfer function is applied over the sum and an output is obtained from each neuron.

The ability of the network to produce a reasonable output stems from a training process. Given a set of inputs and their targets, the network is trained to learn the relationships between both sets. The training process is possible due to a backpropagation training algorithm (Rumelhart et al., 1986). Generally, the network is initialized with random values of weights and biases and an output is obtained. This output is compared with the target through a cost function, that typically is the mean squared error. Then, the algorithm "backpropagates" this error through the network and iteratively adjusts the weights and biases to minimize the cost function. The minimization is commonly based on the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963). Once the network is trained, output values can be obtained from a set of inputs with unknown targets. The more accurate and generalized the training data, the more accurate the output values.

The feed-forward neural network used in this study has a two-layer architecture. The first layer has a sigmoid transfer function and the second layer a linear transfer function (Fig. 1). This choice of functions allows both the linear and non-linear relationships between A_T and its predictors to be represented. This network configuration can approximate most functions arbitrarily well (Hagan et al., 2014). In the Atlantic Ocean, this arrangement has been shown to accurately estimate A_T from diverse predictors (Velo et al., 2013).

The GLODAPv2 discrete data were used to train the network. Input variables (left hand in Fig. 1) were selected based on their potential influence on A_T following Velo et al. (2013). They include the sampling position (coordinates and depth), potential temperature, salinity, nutrients (phosphate, nitrate and silicate) and dissolved oxygen. Position was included to help the network learn characteristic patterns associated with this input when the other variables cannot fully explain the A_T variability Takahashi et al. (2014) and Lee et al. (2006) showed how the relations between A_T and the predictor variables used in these

studies are different depending on the ocean area. The periodicity of the input longitude was represented by the equations used by Zeng et al. (2014):

$$clongitude = \cos\left(\frac{\pi}{180} \cdot longitude\right) \quad (2)$$

$$slongitude = \sin\left(\frac{\pi}{180} \cdot longitude\right) \quad (3)$$

Our approach only uses measured inputs from GLODAPv2, that is, those input data derived from the same Rosette sample bottle as the A_T value. Other studies with a similar approach take the inputs from reanalysis products or satellite data (e.g., Landschützer et al. 2013), that are inherently less accurate than direct measurements. The relations created by the network in the training procedure are likely to be more realistic using in situ measured values for the input variables.

The samples where all input variables and A_T were measured were selected from GLODAPv2 (https://www.nodc.noaa.gov/ocads/oceans/GLODAPv2/). From these, we removed one record due to its spurious oxygen value (O2=1026.9 μmol kg^{-1} cruise=102; station=4; bottle=5). The final dataset contained 246,221 samples. "GLODAPv2" hereinafter refers to the subset used in this study unless otherwise indicated.

Two different training techniques were tested: the Levenberg-Marquardt method (lm) and the Bayesian Regularization (br) (both detailed in Hagan et al., 2014). In a similar study, Velo et al. (2013) demonstrated that these techniques give the best network performance among those they tested. Except for the number of neurons, the two algorithms were implemented with the default options of the MATLAB functions *trainlm* and *trainbr* (detailed in Beale et al., 2017). These two functions prevent overfitting in different ways. The *trainlm* function usually needs to be fed with the data divided in three sets: a training set to obtain the relationships between variables, a validation set to prevent overfitting and a test set to compare different networks. Here, the training was stopped when the error in the validation set increased during 6 consecutive iterations of the training process to avoid overfitting. This process is known as early stopping (Hagan et al., 2014). The final values of the network weights and biases are those reached before the first of these iterations. The *trainbr* function adds a regularization parameter to the cost function to make the fit smoother in order to avoid overfitting. The validation set is not present in this technique. The end of the training is based on network convergence through parameter stabilization by an automatic process known as automated Bayesian Regularization (Hagan et al., 2014; Beale et al., 2017). See Beale et al. (2017) and references therein for a detailed description of the two functions tested.

The number of network neurons is problem dependent with no fixed criterion for establishment. It is related to the complexity of the input-output mapping, the amount of training data available and their noise (Gardner and Dorling, 1998). Using too few neurons will not enable to learn complex relations. Using too many neurons could overfit the data, that is, the network might

model the uncertainty of the data used in the training. We determined the optimal number of neurons through a trade-off between the root-mean-squared error (RMSE) of the computed values and the generalization of the network. This last concept refers to network performance when a set of unused inputs is passed through the network to obtain an output. If the RMSE in this set is of the same order of magnitude as the RMSE in the training set, there is no substantial overfitting and the network generalizes well.

The training procedure was carried out in MATLAB. We tested 16, 32, 64, 128 and 264 neurons in the hidden layer based on the results of Velo et al. (2013). For each number of neurons, we trained 10 networks always using the same 90% of GLODAPv2 for training (Fig. 2, Static level). The remaining 10% was used as a static test (Fig. 2, Static level). Both subsets contained samples randomly distributed in the ocean to evaluate the maximum possible relationships between the input variables and A_T through all oceanographic regimes, that is, to capture most of the variability in all the variables and not restricting the sets to specific areas. Each of the 10 networks starts the training procedure with random weight and bias values and a random division of the training static dataset into three portions: 70% for training, 15% for testing and 15% for validation (Fig. 2, Dynamic level). These differences make minimization of the cost function different for each network due to the complexity of the weight-error space and, consequently, their different starting points in that space. As each network is different, keeping static sets allows one to determine which network best generalizes in the same test set. The selected network is the one that produces the lowest RMSE in the training data (validation + training dynamic) and in the test data (static + dynamic), considering a non-significant difference between both RMSEs to prevent overfitting. The network derived from this process will be referred as NNGv2.

Once we found an adequate network configuration, we increased the amount of data in the training dynamic set to capture more relations between the inputs and A_T. The new percentages of the dynamic sets were: 80% training, 20% validation and 0% testing. The latter set is only necessary to compare different models and is not used during the training. However, the static test set was held to evaluate the generalization of each of the 10 networks to select the best one.

As a last step we eliminated the data points with a difference between measured and computed A_T with the selected network (residuals) beyond ±3RMSE and then retrained the network as above. This procedure was used to identify regions where the network was unable to obtain accurate values and to improve the network mapping in the other areas omitting in this way data that the network could be trying to model without having the appropriate input variables or because they could be data with high measurement errors. Although a well-trained neural network avoids modeling the error, high errors could slightly modify the derived function in a negative manner. The network derived from this process will be referred as NN±3RMSE.

2.2 Comparison of methods

The relations proposed by Lee et al. (2006) and Takahashi et al. (2014) to generate a monthly surface climatology of A_T from different predictors were applied over GLODAPv2. Lee et al. (2006) grouped A_T data (< 20-30 m depth) into 5 oceanographic regimes and obtained a best fit to a quadratic function of sea surface temperature (SST) and sea surface salinity (SSS) in each basin. Takahashi et al. (2014) divided the global ocean into 33 hydrographic provinces and expressed the potential alkalinity (PALK = $A_T + NO_3^-$, < 50 m depth) as a linear regression of salinity in 27 of them. PALK was used instead of A_T for the purpose of eliminating seasonal biological effects, and the inter-province variation reflected differences in $CaCO_3$ production in the mixed layer as well as the contributions of lateral and vertical mixing of waters. The analysis was carried out in the areas defined in the two studies.

The recent methods to compute A_T proposed by Carter et al. (2018) and Bittig et al. (2018) (LIARv2 and CANYON-B respectively) were also compared to the one proposed here. LIARv2 is based on multilinear regressions (MLRs) including the same predictors used in the present study, excluding phosphate (sample position, salinity (S), potential temperature (θ), nitrate (N), apparent oxygen utilization (AOU) and silicate (Si)). This method is composed of 16 equations with a different combination of the input variables, always maintaining the salinity input in each one. The computations with LIARv2 were obtained by the equation with the lowest uncertainty estimate in each sample that this method determines (Carter et al., 2018). CANYON-B is based on a Bayesian neural network derived from GLODAPv2 data including position, time, salinity, temperature and dissolved oxygen as predictors. The two methods were applied on the GLODAPv2 dataset used here and the on a subset excluding the samples where the quality control (QC) of A_T was not done (QC procedures detailed in Olsen et al. (2016) and references therein).

2.3 Validation

To illuminate the complexity of neural networks, several methods to determine the contribution of each predictor variable in the output were proposed in different studies (see Gevrey et al. (2003) and Olden et al. (2004)). We used the Connection Weight Approach (Olden and Jackson, 2002) to evaluate if the network properly associates the A_T variability with the predictor variables. This method was proposed to be the most accurate (Olden et al., 2004). It uses the weights obtained in the training stage to extract the influence of each predictor variable in fitting the A_T values. The expression followed was:

$$C_i = \sum_{k=1}^{H} w_{ik} \cdot w_k \quad (4)$$

where C_i is the relative importance of the predictor variable i, H is the number of neurons in the hidden layer, w_{ik} is the weight of the connection between the variable i and the neuron k of the hidden layer and w_k is the weight of the connection between the neuron k of the hidden layer and the final output, that is, the computed A_T. Finally, the absolute value of C_i was expressed as a percentage of the sum of all C_i.

In addition to the test in the GLODAPv2 independent set, the network potential was tested on five ocean time-series in different oceanographic regimes that were not included in GLODAPv2: Hawaii Ocean Time-Series (HOT), Bermuda Atlantic Time-Series Study (BATS), European Station for Time-Series in the Ocean at the Canary Islands (ESTOC), Kyodo North Pacific Ocean Time-Series (KNOT) and K2.

GLODAPv2 contains quality controlled measurements in all ocean basins from the 1970s until 2013 (Olsen et al., 2016). However, winter data are scarce to absent in some high latitude regions because adverse weather conditions prevents field activities in that season (Fig. 3). In surface ocean, this temporal bias can be avoided with the help of the subsurface data from seasons with sufficient samples. Vázquez-Rodríguez et al. (2012) demonstrated how the subsurface ocean layer in the Atlantic Ocean can retain the footprint of the water mass formation from the preceding winter in the following months and, therefore, of the surface conditions. The winter relationship between inputs and A_T needed to produce an all-season surface climatology are mostly preserved in this subsurface layer. The validity of this hypothesis was tested in other regions (Fig. 3) following Vázquez-Rodríguez et al. (2012). These areas were chosen based on the non-availability of A_T data in two or more consecutive months in the same oceanographic regime as the colored area in Fig. 3.

To reinforce the previous test and to assess the ability of the neural network in overcoming the lack of winter data in other depths, a neural network was trained excluding all winter data in GLODAPv2 (GLODAPv2_nowinter) and tested in the excluded and independent winter dataset (GLODAPv2_winter). The procedure to create and to train the network was the same as described previously.

2.4 Climatology

Finally, we generated a 1ºx1º global monthly climatology of A_T on 102 depth levels from the objectively analyzed climatological fields of WOA13 (Locarini et al., 2013; Zweng et al., 2013; Garcia et al., 2014a; Garcia et al., 2014b). From this database, the same input variables as in the training stage were selected to estimate A_T from the relationships learned by the network. This final product was compared with the monthly sea surface climatologies of A_T of Lee et al. (2006) and Takahashi et al. (2014). Furthermore, the annual mean was compared with the annual mapped climatology by Lauvset et al. (2016). The availability in Lauvset et al. (2016) of the climatologies of the variables used as inputs in the network were used to test how the network represents their climatology of A_T and to evaluate the sources of the possible differences.

3 Results and discussion

3.1 Neural network analysis

The lowest RMSE was reached in the training and in the test sets when 128 neurons were used (Fig. S1). Similar RMSE values for both sets (training: 8 µmol kg^{-1} vs test: 8.5 µmol kg^{-1}; Fig. S1 and Fig. S2) showed that no overfitting occurred, and that

the network generalizes well. The two training techniques did not show significant differences (Table 1). The Levenberg-Marquardt algorithm was selected for its higher computing speed. We also found no improvement by increasing the number of data points in the dynamic training set. The main reason is perhaps the random division of the datasets. All possible relations the network can learn could be represented using only 70% of the static training set, that is, 63% of the GLODAPv2. This result suggests the necessity to include other input variables rather than more data to improve the network mapping.

Samples with residuals beyond ±3RMSE are 1% of the GLODAPv2 dataset. The spatial distribution of these samples (Fig. S3) show that they are confined to certain areas, mainly in the ocean surface (Fig. 4). Most are in the Northern Hemisphere (Fig. S3 and Fig. 4). Specifically, 64% are from latitudes north of 60ºN (Table S1). In this area, 6.5% of GLODAPv2 samples have residuals beyond ±3RMSE and 83.1% of these samples are from the upper 100m (Table S2). In these depth and latitude ranges, the samples with high residuals make up 14% of the GLODAPv2 samples here and they typically have salinities lower than 34 (Table S3; Fig. S3). A monthly analysis in the previously indicated ranges shows that the largest number of samples with residuals beyond ±3RMSE are from the summer months. About 15-19% of all the samples from this season in this area have residuals higher than ±3RMSE (Table S4).

The previous results show that the Arctic Ocean is the region with the largest RMSE, although the network computes well most of the measured A_T in this area. However, the low availability of winter data, the ice-sea dynamics and the transport of A_T by the rivers (Fig. S4) could alter the presence of the surface winter conditions in the summer subsurface layer shown by Vázquez-Rodríguez et al. (2012) in other areas and generate a temporal bias in the climatology. The high discharge of high A_T waters by the rivers in the summer (Cooper et al., 2008; Shiklomanov et al., 2018; Fig. S5) generates the greatest errors and shows how the network fails to model riverine A_T.

In further detail, many of the samples with residuals beyond ±3RMSE are located in the Beaufort Sea (66°N - 80°N, 140°W-180°W). Here, Takahashi et al. (2014) also found the largest RMSE (60.5 µmol kg^{-1}; 57.6 µmol kg^{-1} applying their regression on GLODAPv2) of their SSS-PALK relations in the upper 50m of the water column. This area is specifically complex for the model surface A_T because of significant river runoff having high and possibly variable A_T concentrations (Fig. S4 and S5; Anderson et al. 2004; Cooper et al. 2008). Therefore, in spite of the good reproduction of A_T for the most samples, one should be cautious with the results in this zone and for the entire Arctic Ocean.

The North Sea also contains many samples with large residuals. Those samples shallower than 100m and close to the coasts surrounding this sea do not have an accurately computed A_T (Fig. S3 and Fig. S4). Some studies have shown the complexity of the processes occurring in this shallow sea where the high river runoff also has elevated levels of A_T (Fig. S4; e.g., Hoppema, 1990; Artioli et al. 2012). Hence, the same caveats as for the Arctic Ocean should be made.

In general, the network mainly fails to compute A_T in some samples of areas with rivers carrying significant amounts of A_T to the ocean. The samples beyond ±3RMSE represent 23% and 9.4% of the total above 100m for the Beaufort Sea and the North Sea respectively. The inclusion of predictors related to riverine A_T (and probably to ice melt) could improve the computation in these areas. Although one should be cautious, these zones still should be taken into account and be represented in the climatology since most of the samples have a well-computed A_T.

In the global ocean surface layer, the RMSE obtained with the neural network approach is lower than that obtained by previous studies on generation of monthly climatologies (Table 2 and 3). In the past, relationships between SST and SSS with A_T by Lee et al. (2006) have been shown to produce the lowest RMSE (area-weighted RMSE of 8.1 μmol kg^{-1}) in the A_T computation to create a monthly climatology. However, applying the relations of that study to GLODAPv2, the obtained weighted RMSE is higher than the one from the neural network (Table 2). Neural network approach obtained a better fit in all the areas defined in the study of Lee et al. (2006) (Table 2). NN±3RMSE improves the results obtained with the NNGv2 in almost all the regions, being the most remarkable the Equatorial Upwelling Pacific. However, the difference in the weighted RMSE of the two networks is not significant.

Similar to the previous case, the analysis of the error in the areas defined in Takahashi et al. (2014) also shows a better fit of the neural network (Table 3). Except for the zone with the lowest number of samples (Red Sea), the other 26 areas have a lower RMSE when the A_T is computed by a neural network. The NN±3RMSE improves the fitting of the NNGv2 in the non-Arctic areas. The A_T computed in the zones defined in the Arctic have higher RMSEs in the two approaches (Takahashi et al. (2014) and this study; Table 3). As discussed before, the Beaufort Sea is the zone with the highest RMSE. The inclusion of this area in calculating a global RMSE raises its value considerably. The NN±3RMSE has a higher global weighted RMSE because of the exclusion of most of the samples in this area to train this network. However, the weighted RMSE calculated excluding this area shows again a non-significant difference between the two networks (Table 3).

The results of the two networks clearly show how this fitting technique computes A_T more accurately than the other methods used in studies on the generation of monthly climatologies. The non-linear nature of the neural networks used in this study and the inclusion of multiple predictor variables related to the A_T variability are the main reasons for a good fit. Furthermore, we only used one neural network for the entire ocean. This has the advantage of obtaining the computed A_T anywhere in the ocean in only one step. No "patches" or smoothing are needed between different zones in the climatology as there are in previous studies. Finally, the NNGv2 has been chosen to generate the climatology. Although NN±3RMSE computes A_T with lower errors than NNGv2 in the non-Arctic areas, in a global view the improvement is relatively small (Weighted RMSE in Table 2 and Table 3). In order to include the Arctic in the climatology, the better fit in this area with the NNGv2 approach makes it the best candidate. In any case, the NN±3RMSE is also offered to the users who want to obtain a climatology or A_T computations in a specific area where this network computes A_T better than NNGv2 (e.g., Equatorial Upwelling Pacific, Southern Ocean, etc.).

The newest methods in the A_T computation (LIARv2: Carter et al., 2018; CANYON-B: Bittig et al., 2018) model the GLODAPv2 A_T with higher errors than the NNGv2 (Table 4). An analysis in a GLODAPv2 subset excluding the samples where the 2nd (Olsen et al., 2016) QC was not done for A_T shows a reduction of the error in these three methods, being CANYON and NNGv2 the lowest (Table 4). All the equations are used to compute A_T in the GLODAPv2 dataset when the computation is allowed to be made by the equation with the lowest uncertainty in each sample (Carter et al., 2018). The most used equations are 10 (S, N, Si), 15 (S, AOU) and 14 (S, N), which are used in about 50% of the samples. The equation that used all the input variables (1) is only used to model 3% of the GLODAPv2 samples. Surprisingly, when only this equation is used to compute A_T in GLODAPv2 dataset, the error is lower than those obtained with the free election of the equation based on the lowest uncertainty. That result shows the potential of include all possible inputs related with the A_T variability, although reasonable results can also be reached with the equations that do not use all the input variables. CANYON-B is an example of using relatively few input variables (position, time, temperature, salinity and oxygen) and getting good results (Table 4). Probably, the non-linear character of the neural networks, like the one used in CANYON-B, gives the high potential to this kind of methods to fit complex functions even with few input variables. However, the NNGv2 designed in the present study is the best option to model more GLODAPv2 data better than the other methods (lower RMSE) and therefore to use the mapped inputs-output relation in order to create the monthly climatology. The availability of all the variables used as inputs of the NNGv2 in WOA13 also contributes to make this method the best choice. Furthermore, methods like CANYON-B which include a predictor that explicitly accounts for the time variation of A_T (decimal year in the case of CANYON-B), are not suitable to build a monthly climatology since they generate an unrealistic seasonal amplitude, at least at high depths. This has been checked used WOA13 monthly climatologies (temperature, salinity and dissolved oxygen) as inputs of CANYON-B to compute A_T at different depth layers. As an example, in the 3000m depth layer, seasonal amplitudes up to 40 µmol kg^{-1} were obtained in large areas mainly located between 30 and 60ºS.

The NNGv2 seems to associate the A_T variability to the predictor variables in coherence with the processes that contribute to it. The relative importance of these variables depicted in Fig. 5 shows that salinity is the most influential variable, followed by dissolved oxygen and nutrients. In the surface layer, where A_T variability is the largest, different studies showed how changes in salinity are highly correlated with this variability (Millero et al., 1998; Takahashi et al., 2014). The organic matter cycle also has a significant component in the A_T variability (Kim and Lee, 2009). The formation and degradation of organic matter is reflected through both oxygen and nutrients variations. The network seems to capture the A_T variability because of the organic matter cycle giving a second place in importance to these variables. The third group of variables in the ranking of importance is comprised by depth and temperature. The former variable could be associated to the A_T variability accounting for the variation produced by the $CaCO_3$ cycle and the processes acting through the global ocean circulation. The latter has also been associated to the A_T variability as a proxy of both the $CaCO_3$ and the organic matter cycles (Lee et al., 2006). Finally, the minor contribution of the variables of horizontal sampling position could help to separate the different relations shown by previous studies in different ocean areas (Lee et al., 2006; Takahashi et al., 2014).

3.2 Time-series validation

The network can compute A_T well at 5 different ocean time-series stations. Low RMSEs and high coefficients of determination (r^2) were obtained (Table 5). The bias is relatively low in the three time-series with the highest number of data (HOT, BATS and ESTOC). The A_T computed by the NNGv2 at KNOT and K2 is slightly higher than the measured one, probably because of the influence in the A_T variability of some variable not included as an input of the network (although an offset in the measurements of any of the inputs could also give this result). Summed to the previous test, the statistics obtained in this independent test with a good seasonal time resolution shows the good generalization of the NNGv2.

The ability of NNGv2 to capture surface A_T variability is exemplified in Fig. 6. The other largest time-series also show a good agreement between the computed and the measured seasonal A_T in this surface layer (RMSE HOT: 5.3 µmol kg^{-1}; RMSE ESTOC: 4.2 µmol kg^{-1}). In general, A_T measured in each month of the year are well modeled by NNGv2 (inner charts in Fig. 6). The same holds for other depth layers (Fig. 7, panels in left column). Only some extreme values are not fully captured but almost all the trends between months are well represented. The differences may be caused by bias in measured A_T or some of the input variables; they may also be due to an under/overestimation of the network. Furthermore, the time-series areas are not fully represented in all months in GLODAPv2 so that NNGv2 might not represent seasonality well. However, the network computes A_T in any month with a very low error. This shows again the potential of the generalization of a well-designed neural network.

The NNGv2 also has the capacity to increase the number of A_T data in the time-series. In many samples, A_T was not measured but the other input variables needed for the NNGv2 are available. Therefore, the computed A_T has a higher temporal and spatial resolution than observations only. This enables the computation of more reliable trends than with the less frequently measured A_T and allows the identification of possible high frequency changes. The improvement in resolution is especially visible in the longer time-series: HOT and BATS (Fig. 7). In the former we increased the number of A_T data from 3852 to 14089 and in the latter from 3033 to 11342 (Fig. 7, panels in central column).

The LIARv2 and CANYON-B methods to compute A_T also model the time-series data quite well (Table 6). Significant differences among the three methods are obtained in HOT and ESTOC. In HOT, NNGv2 and CANYON-B reach a better fit of A_T than LIARv2 suggesting that a non-linear technique is more adequately to model A_T in this area (Table 6). In ESTOC, NNGv2 and LIARv2 are the best options to model the A_T variability (Table 6). Here, the A_T computed with LIARv2 with the option of the free equation choice activated results in a greater election of the equations that include nutrients as predictors. This result show how in this area the inclusion of nutrients as predictors contributes to improve the model of A_T. Like NNGv2, both methods have a considerable bias in K2 and KNOT (Table 6) that reinforce the two reasons suggested previously.

3.3 Subsurface Layer Hypothesis

We found that the optimal depth range of the subsurface layer defined by Vázquez-Rodríguez et al. (2012) for the North Atlantic Ocean (100-200 m) must be modified in other regions. In the area analyzed in the Indian Ocean (Fig. 3), the subsurface layer hypothesis is verified in the same depth range of that study. However, the other areas (Fig. 3) show that the range of the subsurface layer is in the range of 50-100 m. The different strengths of deep mixing and convection in winter could explain this fact.

The properties analyzed in the four areas defined in Fig. 3 show, as expected, a higher monthly variability in the ocean surface than in the subsurface layers. The seasonal variability depicted in Fig. 8 will likely be typical of a larger region within a similar oceanographic regime for each defined area. The surface winter conditions of the analyzed properties are quite similar to those in the subsurface layer during, at least, one of the four consecutive months following winter in all areas (Fig. 8).

The optimal number of neurons in the network trained with GLODAPv2_nowinter dataset to reinforce the subsurface layer hypothesis and to assess the layers below surface ocean was 100. The reduction of the number of neurons compared to the previous networks was because this new dataset contains less data. Thus, maintaining or increasing the number of neurons would produce overfitting. This new network provides statistics in the GLODAPv2_nowinter dataset similar to those of the network used to create the climatology (NNGv2) in GLODAPv2 dataset (Table 1 vs Table 7). But, of greater importance are the statistics resulted from the GLODAPv2_winter dataset (Table 7) which reinforce the subsurface layer hypothesis. The low error reached in this independent winter dataset shows how the network is able to obtain the winter relations in any depth from the function fitted with data from other seasons. Therefore, the lack of winter data in different regions does not automatically mean that the climatology will be biased towards the more sampled seasons.

3.4 Climatology

The monthly climatology of A_T is based on the relations obtained in the training procedure of the neural network applied to the WOA13 monthly climatological fields. We have demonstrated that the A_T computed by the two offered neural networks agrees reasonable with the measured A_T when the inputs associated to it are passed through the networks, i.e. the relations obtained from GLODAPv2 in the training stage are robust. Therefore, the A_T patterns in the climatology are forced by the patterns of the WOA13 variables used as inputs. The monthly climatology can be found in a netCDF file at the data repository of the Spanish National Research Council (CSIC; doi: http://dx.doi.org/10.20350/digitalCSIC/8564) together with a video of the monthly variation at the surface and in three longitudinal sections of the three main oceans.

The distribution of the surface annual mean A_T (Fig. 9) is similar to that shown in previous climatologies (e.g., Lee et al. 2006; Takahashi et al. 2014; Lauvset et al. 2016). Not surprisingly, there is a high correlation with the salinity distribution and, consequently, with the evaporation-precipitation patterns. The largest values in the surface layer occur in the Mediterranean Sea, Red Sea, and in the subtropical gyres of the Atlantic and South Pacific Oceans, all of them prevailing throughout the year

in the monthly climatology. At depth, these maxima are all present at least up to 150m (Fig. 9). Below 700m, the Pacific and Indian Oceans show higher A_T concentrations than the younger waters of the Atlantic (Fig. 9). Furthermore, features such as the high-A_T Mediterranean Water entering the Atlantic Ocean are captured in the climatology (Fig. 9, 1000m chart, black circle). In general, the patterns agree with the main ocean processes responsible for the A_T variability as explained previously.

The seasonal amplitude of sea surface A_T (Fig. 10) is generally in agreement with that obtained by Lee et al. (2006). The highest amplitudes are in the north equatorial zone, in the Arctic Ocean and in coastal zones, i.e., at locations where there are rivers with a large water discharge (like the Amazonas, Congo, La Plata or Arctic rivers). The seasonal amplitude of the surface salinity (Fig. S6) can explain most of the variability in the seasonal amplitude of A_T. In areas with a large seasonal amplitude of salinity (more than 1 unit; mainly the Arctic Ocean and coastal zones near rivers with high discharge), this variable linearly explains 76% of the seasonal amplitude A_T variability. However, the seasonal amplitude in the Arctic Ocean should be taken with caution due to the difficulty to accurately model this complex zone, as discussed previously. Despite the presence of high levels of A_T in some river mouths in the melting months, the A_T carried by the rivers could be not represented in the climatology and this can enhance the seasonal cycle due to an underestimated value in low salinity waters with high riverine A_T. On the other hand, in areas with a low seasonal amplitude of salinity (less than 1 unit; mainly oceanic areas and coastal regions without rivers with high discharge) about 61% of variability is linearly explained. This result shows the importance of the inclusion of other predictors besides salinity in the network and the non-linearity of the method proposed in this study to explain nearly all the A_T variability.

The seasonal amplitude of A_T is progressively reduced at depth (Fig. S7). The changes in the variables which influence the changes in A_T are smaller than in the surface layer or null causing this reduction. The seasonality disappears almost completely below 500m depth; not surprising due to the lack of seasonal resolution in the climatologies of nutrients in WOA13 below this level. Some patches of variability are present likely because of a conjunction of the error of the network and the monthly changes in the other WOA13 input variables. In addition, they could also come from the learning stage since the training data present monthly variations of up to ~10 µmol kg^{-1} for the same area, even at depths greater than 1000m.

Although it was shown that the neural network can accurately compute A_T in both GLODAPv2 and time-series datasets, the quality of WOA13 data also determines the robustness of the climatology is. Unfortunately, WOA13 does not offer uncertainty fields associated to the objectively analyzed climatologies to compute a coherent estimation of the uncertainty in the A_T climatology. Therefore, the climatological values offered in this study should be evaluated by comparing them with observations in a monthly average over many years. This can only be done at the locations of time-series with representative amounts of data; Fig. 11 shows this analysis at surface. At both the BATS and HOT time-series, the differences between the averaged measured A_T (Fig. 11, red line) and the climatology (Fig. 11, yellow line) are quite low. The comparisons are better when A_T is computed by NNGv2 using as inputs the measured values in the time-series (Fig. 11, purple line). The differences of the two comparisons show the differences in the input variables (WOA13 climatological fields vs time-series input data).

The previous results hold true also for other depth layers. A comparison of monthly profiles up to about 500m between the A_T climatology obtained from WOA13 and the one from the averaging of the time-series data shows low differences. In BATS, the RMSE of this comparison ranges between 1.1 and 2.8 µmol kg^{-1} (mean RMSE of 2 µmol kg^{-1}) and the bias between 0 and 4.7 µmol kg^{-1} for all months. In HOT, the RMSE of this comparison ranges between 5 and 10.5 µmol kg^{-1} (mean RMSE of 6.4 µmol kg^{-1}) and the bias between -0.3 and 6.3 µmol kg^{-1} for all months. The climatological measured data are for the periods between 1991 and 2015 (BATS) and 1989 and 2016 (HOT) and WOA13 data are supposed to cover a larger range. Despite this time difference, the A_T climatology represents quite accurately the measured values averaged in each month.

Compared to the other climatologies, the surface annual mean A_T of this study is closer to that of Lee et al. (2006) (Table 8). This is likely because temperature and salinity are included as non-linear predictors of A_T. In Takahashi et al. (2014), A_T derives from the linear regression between PALK and one predictor (salinity) and in the Lauvset et al. (2016) study, DIVA (Data-Interpolating Variational Analysis; Troupin et al., 2010) was used. Furthermore, the transfer of our climatology to the coarser grid of Takahashi et al. (2014) for the comparisons may enhance dissimilarities.

The comparison of the monthly values of our climatology and the other climatologies available at the same time frequency (Table 9) shows the greatest similarity of ours and that of Lee et al. (2006). The reasons given above may also hold here. In addition, part of the differences between the comparisons may originate from the different versions of the WOA used in each study (Lee et al., 2006: temperature and salinity from WOA01; Takahashi et al., 2014: salinity from WOA09 and nitrate from WOA94; this study: all inputs from WOA13).

In general, the surface spatial patterns of the differences between the annual mean of our A_T climatology and the three other ones under consideration are not correlated (Figure S8). Compared to Takahashi et al. (2014), the largest differences are in the Beaufort Sea and in three zonal bands: 54-60º S, 8-28º N and 40-60º N (Fig. S8a). The Pacific Ocean has the highest dissimilarities in these three bands. In general, the Atlantic Ocean and the Indian Ocean have the smallest differences. The largest differences in these two ocean basins are mainly located close to the river mouths. It shows how the different parametrizations of the A_T diverge highly at low salinities. On the other hand, the major differences with Lee et al. (2006) (Fig. S8b) are surrounding North America's Pacific coast, the area of influence of the Amazon river, the zone between both the Niger and the Congo rivers and the North Sea. In the open ocean there are some wide areas where the differences are remarkably high. They are mainly in the South Pacific. It should also be noted that the transition zone between the 1 ((sub)tropics) and 2 (equatorial upwelling Pacific) areas defined in the study of Lee et al. (2006) generates a discontinuity in the difference map. Finally, the largest differences with Lauvset et al. (2016) (Fig. S8c) are less localized. The Arctic Ocean and the Pacific sector of the Southern Ocean are the areas where there is a large spatial continuity in the differences.

An important cause of the differences between the climatologies stems from the use of different inputs to generate them. As an example, this can be seen when the climatologies of Lauvset et al. (2016) are used as input variables to compute A_T with

the neural network instead of the WOA13 data (Fig. 12). In the surface layer, a considerable reduction of the RMSE (15.7 to 12.3 μmol kg^{-1}) and an increase of the r^2 from 0.91 to 0.95 are obtained (Fig. 12). In the deeper layers, the differences are progressively decreasing. The values of the RMSE of the comparisons like those in Fig. 12 but below 250m are in the range of 4 to 6 μmol kg^{-1} and the improvement caused by the inputs usage is reduced to around 1 μmol kg^{-1}. This last result shows an increasing similarity between WOA13 climatologies and Lauvset et al. (2016) climatologies with increasing depth. However, and to be consistent, it is recommended to use the A_T climatology corresponding with the other inputs used in the studies that arise from these products.

4 Data availability

The climatology and the two neural networks designed in this study are available at the data repository of the Spanish National Research Council (CSIC; doi: http://dx.doi.org/10.20350/digitalCSIC/8564).

5 Conclusions

A neural network to compute A_T anywhere in the ocean has been presented. As evaluated by the RMSE between the measured and the computed data, the neural network approach presented in this study offers increased precision compared to most of the approaches in previous studies. Furthermore, the global relationship between A_T and input variables was obtained from a higher number of quality-controlled data than before in the generation of a monthly climatology, with a greater temporal and spatial resolution. We have demonstrated how one single global algorithm is able to compute A_T satisfactorily for the entire global ocean. This has enabled us to generate a monthly climatology without the need to use smoothing techniques between different oceanic areas. Furthermore, the seasonal variability in depth is more realistic than the one computed by other methods that overestimate it.

The validation using different independent datasets demonstrates the good network generalization. In addition, the spatiotemporal A_T variability is well captured by the network as shown in time-series validation. Therefore, the obtained climatology using WOA13 inputs should reflect this variability due to the good network performance to new independent data.

We offer this global monthly climatology of A_T to the scientific community for advancing the understanding of the ocean carbon cycle. Our new climatology may particularly be useful as input to modeling efforts. It is worthwhile mentioning that the networks offered here are also useful to obtain A_T values for samples where the inputs for the neural network are present.

6 Author contributions

DB, FFP and AV designed the study. The manuscript was written by DB and revised and discussed by all the authors. The dataset of the climatology and the neural networks were created by DB.

7 Competing interests

The authors declare that they have no conflict of interest.

8 Acknowledgements

This research was supported by Ministerio de Educación, Cultura y Deporte (FPU grant FPU15/06026), Ministerio de Economía y Competitividad through the ARIOS (CTM2016-76146-C3-1-R) project co-funded by the Fondo Europeo de Desarrollo Regional 2014-2020 (FEDER) and EU Horizon 2020 through the AtlantOS project (grant agreement 633211). The authors want to thank the comments of Siv K. Lauvset to improve the manuscript.

9 References

Anderson, L. G., Jutterström, S., Kaltin, S., Jones, E. P. and Björk, G.: Variability in river runoff distribution in the Eurasian Basin of the Arctic Ocean, J. Geophys. Res., 109(C1), 1–8, doi:10.1029/2003JC001773, 2004.

Artioli, Y., Blackford, J. C., Butenschön, M., Holt, J. T., Wakelin, S. L., Thomas, H., Borges, A. V and Allen, I.: The carbonate system in the North Sea: sensitivity and model validation, J. Mar. Syst., 102–104, 1–13, doi:10.1016/j.jmarsys.2012.04.006, 2012.

Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S. D., Nakaoka, S. I., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T., Tilbrook, B., Wada, C., Wanninkhof, R., Alin, S. R., Balestrini, C. F., Barbero, L., Bates, N. R., Bianchi, A. A., Bonou, F., Boutin, J., Bozec, Y., Burger, E. F., Cai, W. J., Castle, R. D., Chen, L., Chierici, M., Currie, K., Evans, W., Featherstone, C., Feely, R. A., Fransson, A., Goyet, C., Greenwood, N., Gregor, L., Hankin, S., Hardman-Mountford, N. J., Harlay, J., Hauck, J., Hoppema, M., Humphreys, M. P., Hunt, C. W., Huss, B., Ibánhez, J. S. P., Johannessen, T., Keeling, R., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lo Monaco, C., Manke, A., Mathis, J. T., Merlivat, L., Millero, F. J., Monteiro, P. M. S., Munro, D. R., Murata, A., Newberger, T., Omar, A. M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L. L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K. F., Sutherland, S. C., Sutton, A. J., Tadokoro, K., Telszewski, M., Tuma, M., Van Heuven, S. M. A. C., Vandemark, D., Ward, B., Watson, A. J. and Xu, S.: A multi-decade record of high-quality fCO2 data in version 3 of the Surface Ocean CO2 Atlas (SOCAT), Earth Syst. Sci. Data, 8(2), 383–413, doi:10.5194/essd-8-383-2016, 2016.

Bates, N., Astor, Y., Church, M., Currie, K., Dore, J., Gonaález-Dávila, M., Lorenzoni, L., Muller-Karger, F., Olafsson, J. and Santa-Casiano, M.: A Time-Series View of Changing Ocean Chemistry Due to Ocean Uptake of Anthropogenic CO2 and Ocean Acidification, Oceanography, 27(1), 126–141, doi:10.5670/oceanog.2014.16, 2014.

Beale, M. H., Hagan, T. M. and Demuth, H. B.: Deep Learning Toolbox™. User's Guide. Release 2018a, The MathWorks, Inc., Natick, Massachusetts, United States. Available at: https://es.mathworks.com/help/pdf_doc/deeplearning/nnet_ug.pdf Last access: 20 august 2018. 2018

Bittig, H. C., Steinhoff, T., Claustre, H., Fiedler, B., Williams, N. L., Sauzède, R., Körtzinger, A. and Gattuso, J.-P.: An alternative to static climatologies: robust estimation of open ocean CO2 variables and nutrient concentrations from T, S, and O2 data using Bayesian neural networks, Front. Mar. Sci., 5, 328, doi:10.3389/fmars.2018.00328, 2018.

Brewer, P. G. and Goldman, J. C.: Alkalinity changes generated by phytoplankton, Limnol. Oceanogr., 21(1), 108–117, doi:10.4319/lo.1976.21.1.0108, 1976.

Broecker, W. S.: "NO", a conservative water-mass tracer, Earth Planet. Sci. Lett., 23(1), 100–107, doi:10.1016/0012-821X(74)90036-3, 1974.

Carter, B. R., Toggweiler, J. R., Key, R. M. and Sarmiento, J. L.: Processes determining the marine alkalinity and calcium carbonate saturation state distributions, Biogeosciences, 11(24), 7349–7362, doi:10.5194/bg-11-7349-2014, 2014.

Carter, B. R., Feely, R. A., Williams, N. L., Dickson, A. G., Fong, M. B. and Takeshita, Y.: Updated methods for global locally interpolated estimation of alkalinity, pH, and nitrate, Limnol. Oceanogr. Methods, 16(2), 119–131, doi:10.1002/lom3.10232, 2018.

Chen, C.-T. A.: Shelf-vs. dissolution-generated alkalinity above the chemical lysocline, Deep Sea Res. Part II Top. Stud. Oceanogr., 49(24), 5365–5375, doi:https://doi.org/10.1016/S0967-0645(02)00196-0, 2002.

Cooper, L. W., McClelland, J. W., Holmes, R. M., Raymond, P. A., Gibson, J. J., Guay, C. K. and Peterson, B. J.: Flow-weighted values of runoff tracers (δ18O, DOC, Ba, alkalinity) from the six largest Arctic rivers, Geophys. Res. Lett., 35(18), 3–7, doi:10.1029/2008GL035007, 2008.

Dickson, A. G.: An exact definition of total alkalinity and a procedure for the estimation of alkalinity and total inorganic carbon from titration data, Deep Sea Res. Part A. Oceanogr. Res. Pap., 28(6), 609–623, doi:10.1016/0198-0149(81)90121-7, 1981.

Doney, S. C., Fabry, V. J., Feely, R. A. and Kleypas, J. A.: Ocean Acidification: The Other CO $_2$ Problem, Ann. Rev. Mar. Sci., 1(1), 169–192, doi:10.1146/annurev.marine.010908.163834, 2009.

Fabry, V. J., Seibel, B. A., Feely, R. A., Fabry, J. C. O. and Fabry, V. J.: Impacts of ocean acidification on marine fauna and ecosystem processes, ICES J. Mar. Sci., 65(December), 414–432, doi:10.1093/icesjms/fsn048, 2008.

Fine, R. A., Willey, D. A. and Millero, F. J.: Alkalinity from Aquarius satellite data, Geophys. Res. Lett., 44, 261–267, doi:10.1002/2016GL071712, 2017.

Friis, K., Körtzinger, A. and Wallace, D. W. R.: The salinity normalization of marine inorganic carbon chemistry data, Geophys. Res. Lett., 30(2), doi:10.1029/2002GL015898, 2003.

Fry, C. H., Tyrrell, T., Hain, M. P., Bates, N. R. and Achterberg, E. P.: Analysis of global surface ocean alkalinity to determine controlling processes, Mar. Chem., 174, 46–57, doi:10.1016/j.marchem.2015.05.003, 2015.

Garcia, H. E., R. A. Locarnini, T. P. Boyer, J. I. Antonov, O.K. Baranova, M.M. Zweng, J.R. Reagan, D.R. Johnson.: World Ocean Atlas 2013, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 75, 27 pp., 2014a.

Garcia, H. E., R. A. Locarnini, T. P. Boyer, J. I. Antonov, O.K. Baranova, M.M. Zweng, J.R. Reagan, D.R. Johnson.: World Ocean Atlas 2013, Volume 4: Dissolved Inorganic Nutrients (phosphate, nitrate, silicate). S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 76, 25 pp., 2014b.

Gardner, M. . and Dorling, S. .: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmos. Environ., 32(14–15), 2627–2636, doi:10.1016/S1352-2310(97)00447-0, 1998.

Gevrey, M., Dimopoulos, I. and Lek, S.: Review and comparison of methods to study the contribution of v ariables in artificial neural network models, Ecol. Modell., 160, 249–264, 2003.

Hagan, M. T., Demuth, H. B., Beale, M. H. and De Jesús, O.: Neural network design. ISBN 978-0971732117, 2014.

Hoegh-Guldberg, O. and Bruno, J. F.: The Impact of Climate Change on the World's Marine Ecosystems, Science (80-.)., 328(5985), 1523 LP-1528 [online] Available from: http://science.sciencemag.org/content/328/5985/1523.abstract, 2010.

Hoppema, M.: The distribution and seasonal variation of alkalinity in the Southern Bight of the North Sea and in the Western Wadden Sea, Netherlands J. Sea Res., 26(1), 11–23, doi:10.1016/0077-7579(90)90053-J, 1990.

Key, R. M., Olsen, A., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., Jutterström, S., Steinfeldt, R., Jeansson, E., Ishi, M., Perez, F. F. and Suzuki, T.: Global Ocean Data Analysis Project, Version 2 (GLODAPv2), ORNL/CDIAC-162, NDP-093, doi:10.3334/CDIAC/OTG.NDP093_GLODAPv2, 2015.

Kim, H. and Lee, K.: Significant contribution of dissolved organic matter to seawater alkalinity, Geophys. Res. Lett., 36(September), 1–5, doi:10.1029/2009GL040271, 2009.

Kroeker, K. J., Kordas, R. L., Crim, R., Hendriks, I. E., Ramajo, L., Singh, G. S., Duarte, C. M. and Gattuso, J.-P.: Impacts of ocean acidification on marine organisms: quantifying sensitivities and interaction with warming, Glob. Chang. Biol., 19(6), 1884–1896, doi:10.1111/gcb.12179, 2013.

Landschützer, P., Gruber, N., Bakker, D. C. E., Schuster, U., Nakaoka, S., Payne, M. R., Sasse, T. P. and Zeng, J.: A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink, Biogeosciences, 10(11), 7793–7815, doi:10.5194/bg-10-7793-2013, 2013.

Landschützer, P., Gruber, N., Bakker, D. C. E. and Schuster, U.: Recent variability of the global ocean carbon sink, Global Biogeochem. Cycles, 28(9), 927–949, doi:10.1002/2014GB004853, 2014.

Lauvset, S. K., Key, R. M., Olsen, A., Van Heuven, S., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., Jutterström, S., Steinfeldt, R., Jeansson, E., Ishii, M., Perez, F. F., Suzuki, T. and Watelet, S.: A new global interior ocean mapped climatology: The 1° × 1° GLODAP version 2, Earth Syst. Sci. Data, 8(2), 325–340, doi:10.5194/essd-8-325-2016, 2016.

Lee, K., Tong, L. T., Millero, F. J., Sabine, C. L., Dickson, A. G., Goyet, C., Park, G. H., Wanninkhof, R., Feely, R. A. and Key, R. M.: Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans, Geophys. Res. Lett., 33(19), 1–5, doi:10.1029/2006GL027207, 2006.

Levenberg, K.: A Method for the solution of certain non-linear probles in least squares., Q. Appl. Math., II(2), 164–168, 1944.

Marquardt, D.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters, J. Soc. Ind. Appl. Math., 11(2), 1963.

Millero, F. J., Lee, K. and Roche, M.: Distribution of alkalinity in the surface waters of the major oceans, Mar. Chem., 60(1–2), 111–130, doi:10.1016/S0304-4203(97)00084-4, 1998.

Locarnini, R. A., A. V. Mishonov, J. I. Antonov, T. P. Boyer, H. E. Garcia, O. K. Baranova, M. M. Zweng, C. R. Paver, J. R. Reagan, D. R. Johnson, M. Hamilton, and D. Seidov.: World Ocean Atlas 2013, Volume 1: Temperature. S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 73, 40 pp., 2013.

Olden, J. D. and Jackson, D. A.: Illuminating the '" black box "': a randomization approach for understanding variable contributions in artificial neural networks, Ecol. Modell., 154, 135–150, doi:10.1016/S0304-3800(02)00064-9, 2002.

Olden, J. D., Joy, M. K. and Death, R. G.: An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data, Ecol. Modell., 178, 389–397, doi:10.1016/j.ecolmodel.2004.03.013, 2004.

Olsen, A., Key, R. M., Van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., Jutterström, S., Steinfeldt, R., Jeansson, E., Ishii, M., Pérez, F. F. and Suzuki, T.: The global ocean data analysis project version 2 (GLODAPv2) - An internally consistent data product for the world ocean, Earth Syst. Sci. Data, 8(2), 297–323, doi:10.5194/essd-8-297-2016, 2016.

Orr, J. C., Fabry, V. J., Aumont, O., Bopp, L., Doney, S. C., Feely, R. A., Gnanadesikan, A., Gruber, N., Ishida, A., Joos, F., Key, R. M., Lindsay, K., Maier-Reimer, E., Matear, R., Monfray, P., Mouchet, A., Najjar, R. G., Plattner, G. K., Rodgers, K. B., Sabine, C. L., Sarmiento, J. L., Schlitzer, R., Slater, R. D., Totterdell, I. J., Weirig, M. F., Yamanaka, Y. and Yool, A.: Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms, Nature, 437(7059), 681–686, 2005.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Pongratz, J., Manning, A. C., Korsbakken, J. I., Peters, G. P., Canadell, J. G., Jackson, R. B., Boden, T. A., Tans, P. P., Andrews, O. D., Arora, V. K., Bakker, D. C. E., Barbero, L., Becker, M., Betts, R. A., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Cosca, C. E., Cross, J., Currie, K., Gasser, T., Harris, I., Hauck, J., Haverd, V., Houghton, R. A., Hunt, C. W., Hurtt, G., Ilyina, T., Jain, A. K., Kato, E., Kautz, M., Keeling, R. F., Klein Goldewijk, K., Körtzinger, A., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Lima, I., Lombardozzi, D., Metzl, N., Millero, F., Monteiro, P. M. S., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S., Nojiri, Y., Padín, X. A., Peregon, A., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Reimer, J., Rödenbeck, C., Schwinger, J., Séférian, R., Skjelvan, I., Stocker, B. D., Tian, H., Tilbrook, B., van der Laan-Luijkx, I. T., van der Werf, G. R., van Heuven, S., Viovy, N., Vuichard, N., Walker, A. P., Watson, A. J., Wiltshire, A. J., Zaehle, S. and Zhu, D.: Global Carbon Budget 2017, Earth Syst. Sci. Data Discuss., 1–79, doi:10.5194/essd-2017-123, 2017.

Renforth, P. and Henderson, G.: Assessing ocean alkalinity for carbon sequestration, Rev. Geophys., 55(3), 636–674, doi:10.1002/2016RG000533, 2017.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors, Nature, 323(6088), 533–536, doi:10.1038/323533a0, 1986.

Russell, S. J. and Norvig, P.: Artificial intelligence: a modern approach, Prentice Hall., 2010.

Schlitzer, R., Ocean Data View, http://odv.awi.de, 2016.

Schneider, A., Wallace, D. W. R. and Körtzinger, A.: Alkalinity of the Mediterranean Sea, Geophys. Res. Lett., 34(15), doi:10.1029/2006GL028842, 2007.

Shiklomanov, A.I., R.M. Holmes, J.W. McClelland, S.E. Tank, and R.G.M. Spencer.: Arctic Great Rivers Observatory. Discharge Dataset, Version 20180724. https://www.arcticrivers.org/data , 2018

Steinfeldt, R., Rhein, M., Bullister, J. L. and Tanhua, T.: Inventory changes in anthropogenic carbon from 1997-2003 in the Atlantic Ocean between 20°S and 65°N, Global Biogeochem. Cycles, 23(3), n/a-n/a, doi:10.1029/2008GB003311, 2009.

Takahashi, T., Sutherland, S. C., Chipman, D. W., Goddard, J. G. and Ho, C.: Climatological distributions of pH, pCO2, total CO2, alkalinity, and CaCO3 saturation in the global surface ocean, and temporal changes at selected locations, Mar. Chem., 164, 95–125, doi:10.1016/j.marchem.2014.06.004, 2014.

Takahashi, T., Sutherland S.C. and Kozyr, A.: Global Ocean Surface Water Partial Pressure of CO2 Database: Measurements Performed During 1957-2015 (Version 2015). ORNL/CDIAC-161, NDP-088(V2015). Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, doi: 10.3334/CDIAC/OTG.NDP088(V2015), 2016

Tanhua, T., Bates, N. R. and Körtzinger, A.: The marine carbon cycle and ocean carbon inventories, Int. Geophys., 103, 787–815, doi:10.1016/B978-0-12-391851-2.00030-1, 2013.

Troupin, C., Machín, F., Ouberdous, M., Sirjacobs, D., Barth, A. and Beckers, J.-M.: High-resolution climatology of the northeast Atlantic using Data-Interpolating Variational Analysis (Diva), J. Geophys. Res. Ocean., 115(C8), 1–20, doi:10.1029/2009JC005512, 2010.

Vázquez-Rodríguez, M., Padin, X. A., Pardo, P. C., Ríos, A. F. and Pérez, F. F.: The subsurface layer reference to calculate preformed alkalinity and air-sea CO 2 disequilibrium in the Atlantic Ocean, J. Mar. Syst., 94, 52–63, doi:10.1016/j.jmarsys.2011.10.008, 2012.

Velo, A., Pérez, F. F., Tanhua, T., Gilcoto, M., Ríos, A. F. and Key, R. M.: Total alkalinity estimation using MLR and neural network techniques, J. Mar. Syst., 111–112, 11–18, doi:10.1016/j.jmarsys.2012.09.002, 2013.

Wolf-Gladrow, D. A., Zeebe, R. E., Klaas, C., Körtzinger, A. and Dickson, A. G.: Total alkalinity: The explicit conservative expression and its application to biogeochemical processes, Mar. Chem., 106(1–2 SPEC. ISS.), 287–300, doi:10.1016/j.marchem.2007.01.006, 2007.

Zeng, J., Nojiri, Y., Landschützer, P., Telszewski, M. and Nakaoka, S.: A global surface ocean fCO2 climatology based on a feed-forward neural network, J. Atmos. Ocean. Technol., 31(8), 1838–1849, doi:10.1175/JTECH-D-13-00137.1, 2014.

Zweng, M.M, J.R. Reagan, J.I. Antonov, R.A. Locarnini, A.V. Mishonov, T.P. Boyer, H.E. Garcia, O.K. Baranova, D.R. Johnson, D. Seidov, M.M. Biddle.: World Ocean Atlas 2013, Volume 2: Salinity. S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 74, 39 pp., 2013.

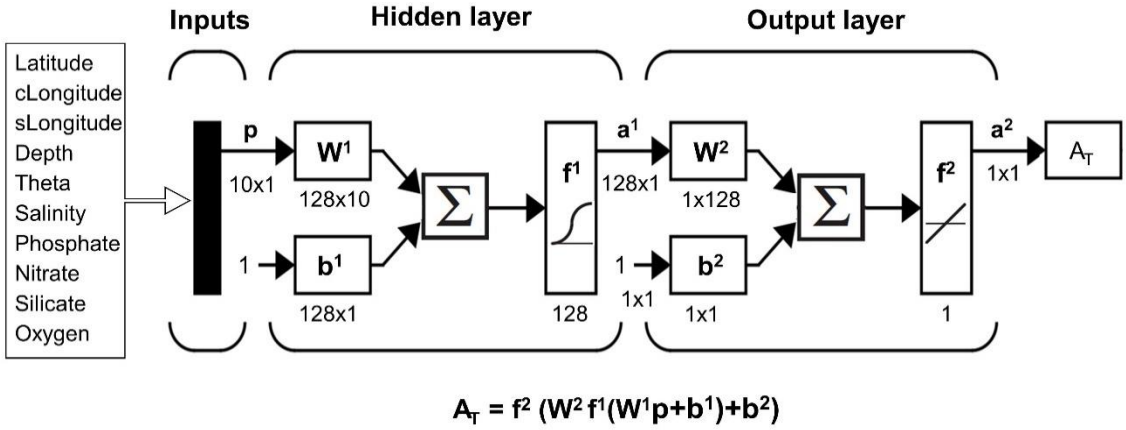


Figure 1: Neural network configuration. The notation is in agreement with Hagan et al. (2014). Theta: potential temperature; p: input vectors; W: weight matrix; b: bias matrix; ∑: sum; f: transfer function; a: output matrix. The superscripts indicate the number of the layer. The c and s preceding month and longitude variables represent cosine and sine (See equations below)). The dimensions of the matrices are for an individual sample. Modified from Hagan et al. (2014).

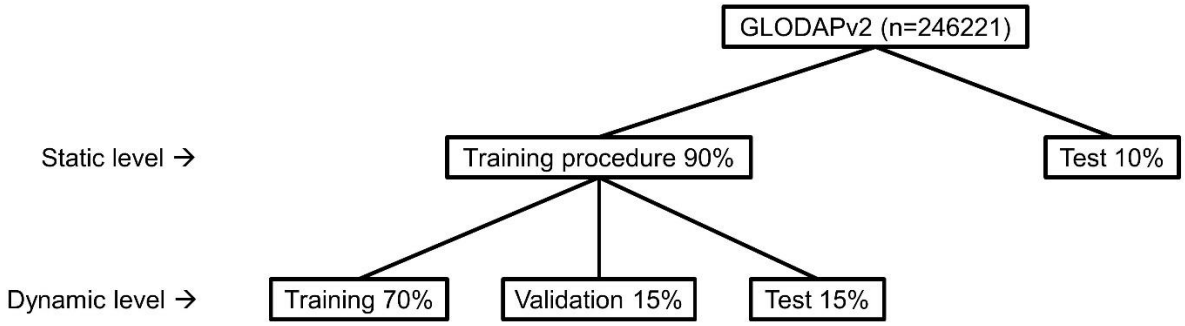


Figure 2. Division of the data for the training of the network. The data in the sets of the static level is the same for all the networks to train. The data in the sets of the dynamic level is randomly selected for each network to train.

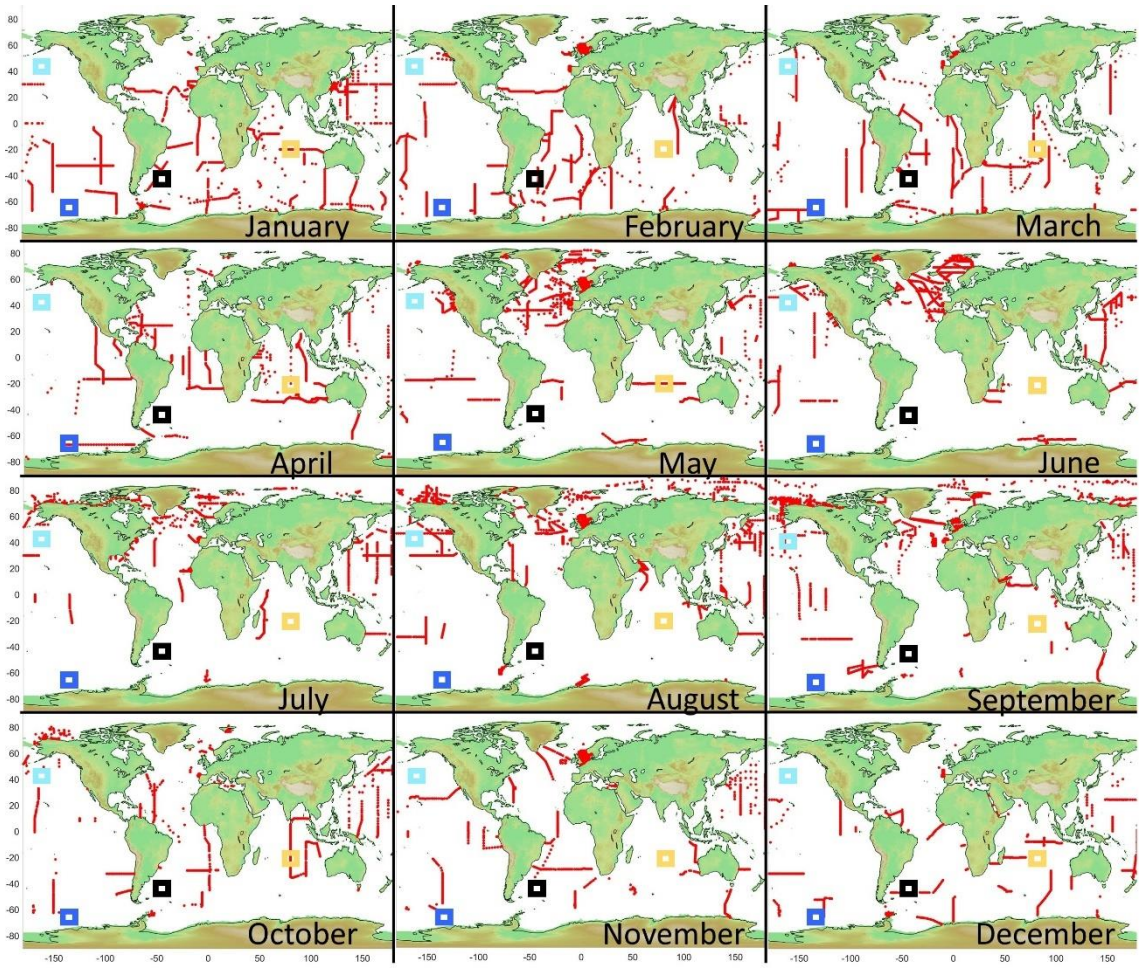


Figure 3: Locations of GLODAPv2 data used in this study presented by month of observation (red dots). Areas where subsurface layer hypothesis was evaluated are shown as colored rectangles.

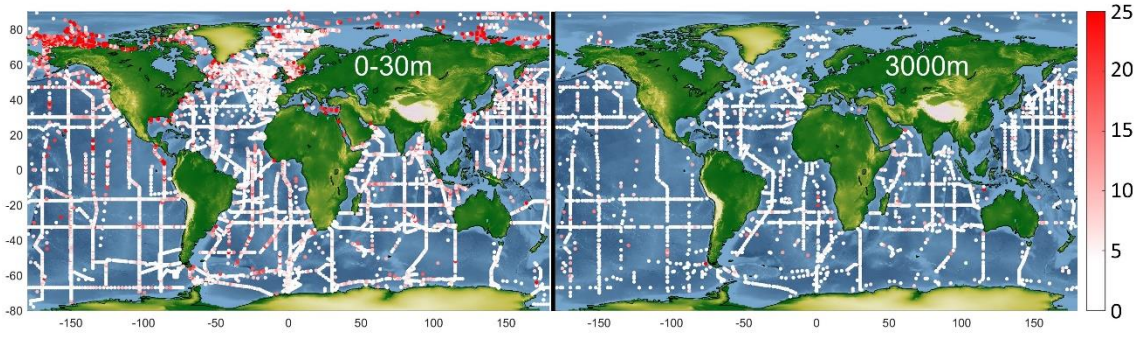


Figure 4: The absolute differences between GLODAPv2 A_T and NNGv2 A_T. Left: samples in the layer 0-30m. Right: samples in the layer 2950-3050m.

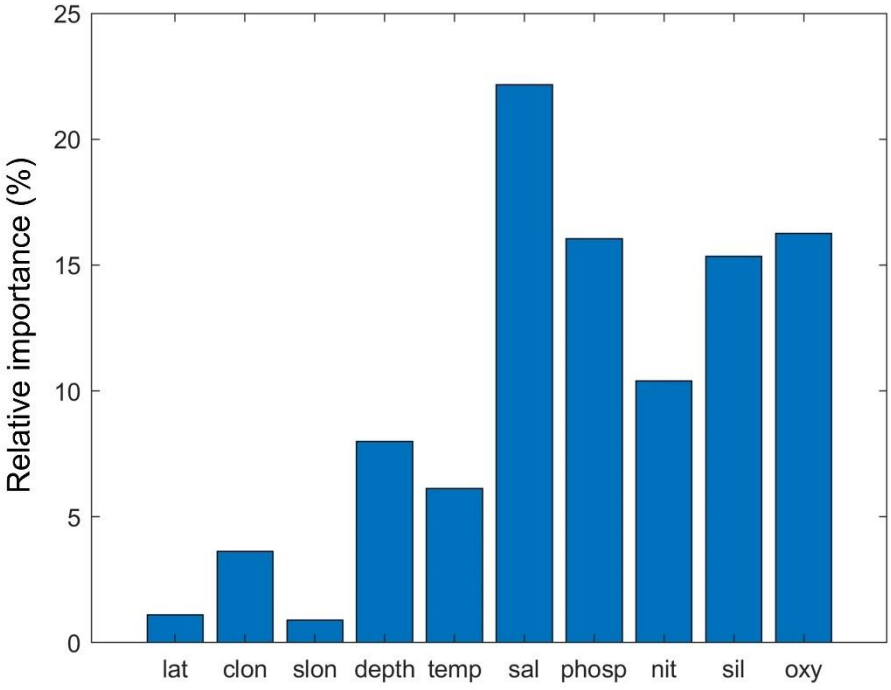
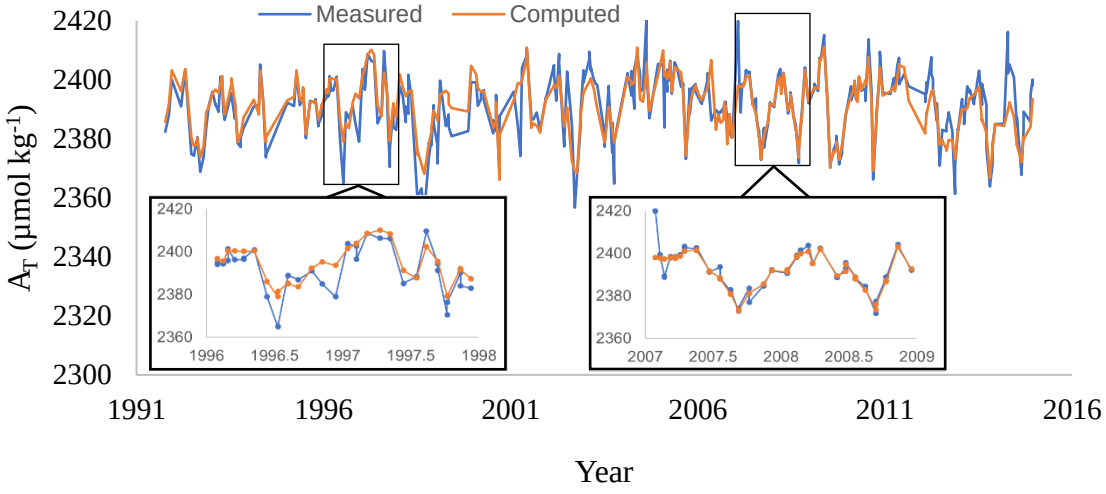
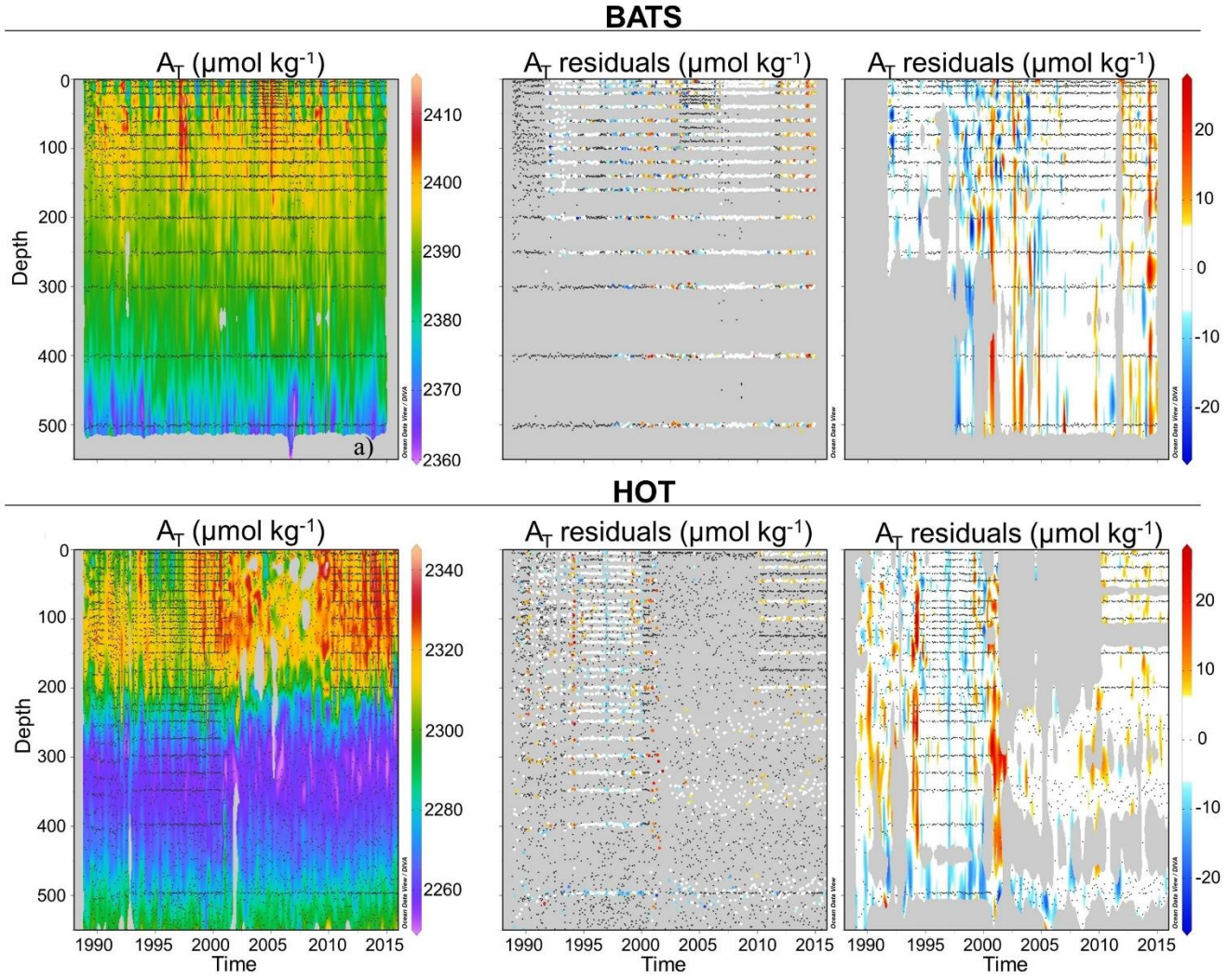


Figure 5: The relative importance of the predictor variables for the NN. lat: latitude; clon: Eq. (3) slon: Eq. (4); temp: temperature; sal: salinity; phosp: phosphate; nit: nitrate; sil: silicate; oxy: oxygen.



660

Figure 6: Comparison of measured and computed A_T for the depth range 0-10 m at time-series station BATS. The RMSE in that depth range for the whole time-period is 5.7 µmol kg^{-1}. The years 1996-1997 and 2007-2008 are amplified to show the monthly variations because they are the years with A_T measurements in all the months.



665 **Figure 7: Left column: Computed A_T for the upper 550m of the water column at the BATS and HOT time-series stations. Central column: Difference between measured and computed A_T. Colored dots show samples where A_T was measured. Black dots show samples where A_T was not measured but the network inputs were. Right column: Difference between measured and computed A_T interpolated with Data-Interpolating Variational Analysis (DIVA; Troupin et al., 2010). This figure was made with Ocean Data View (Schlitzer, 2016).**

670

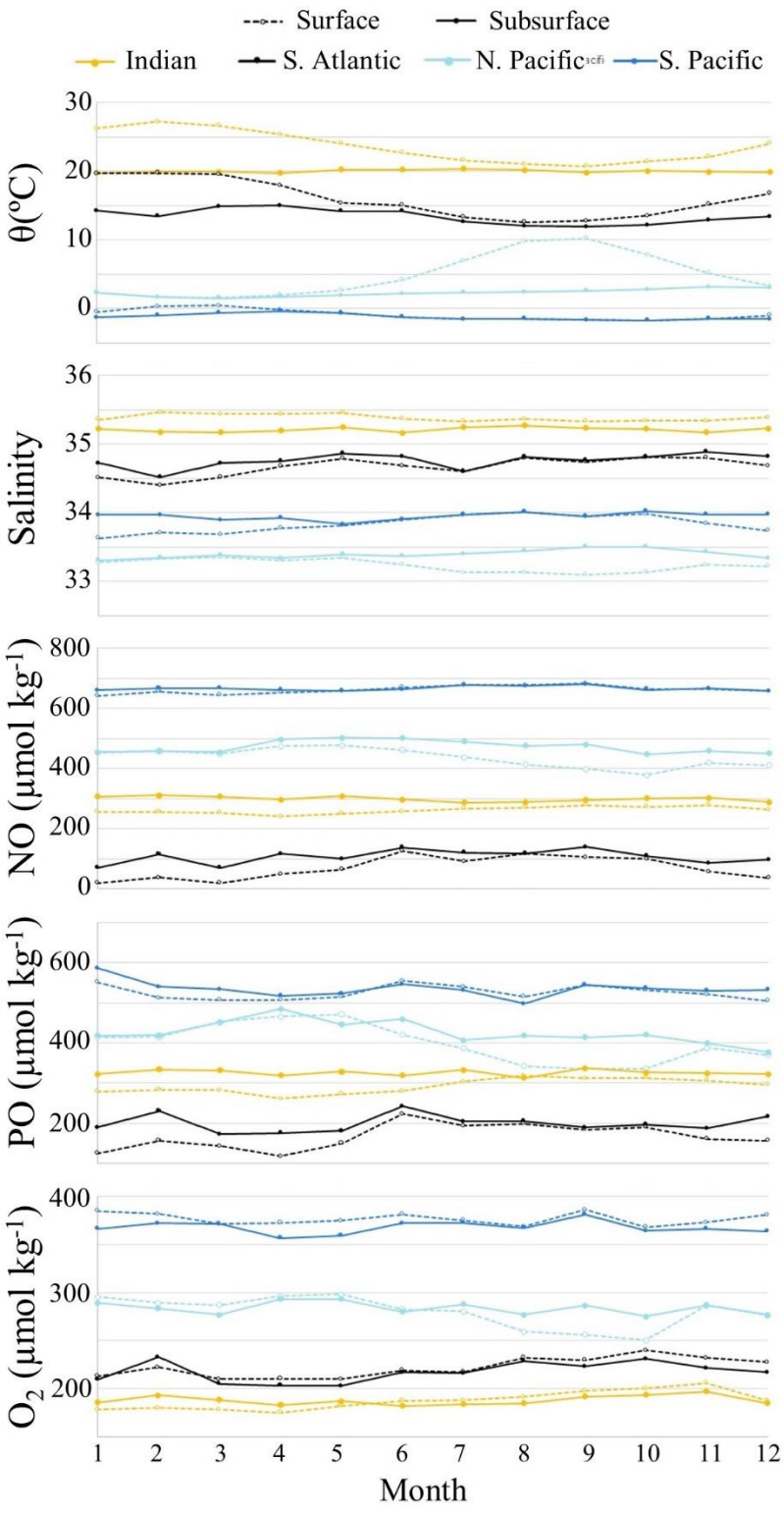

Surface
Subsurface
Indian
S. Atlantic
N. Pacific
S. Pacific
θ(ºC)
Salinity
NO (µmol kg-1)
PO (µmol kg-1)
O2 (µmol kg-1)
Month

Figure 8: Monthly variability of θ (potential temperature), salinity, NO = 9*NO_3 + O_2 and PO = 135*PO_4 + O_2 (defined according to Broecker, 1974) for different ocean basins. Data from WOA13 objectively analyzed monthly climatologies were averaged for each area defined in Figure 2. Each zone is displaced in each graph for a certain constant quantity of the variable for a better visualization, that is, the data shown are not the real values. Indian Ocean: 100-200m; South Atlantic, South Pacific and North Pacific: 50-100m.

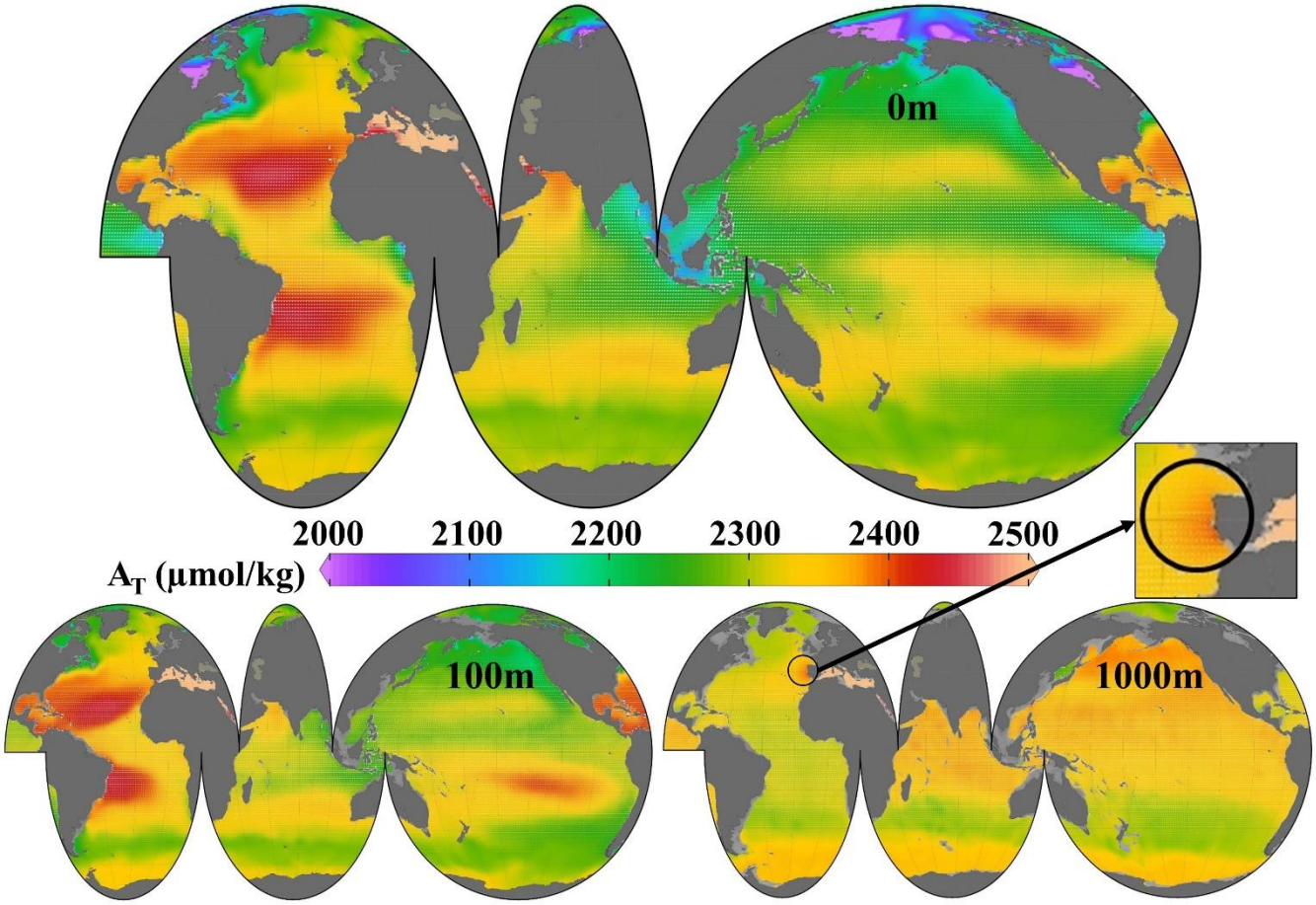


Figure 9: Annual mean climatology of A_T at 3 depths. Black circle in 1000m panel points out the area of influence of the Mediterranean Water in the Atlantic Ocean. This figure was made with Ocean Data View (Schlitzer, 2016).

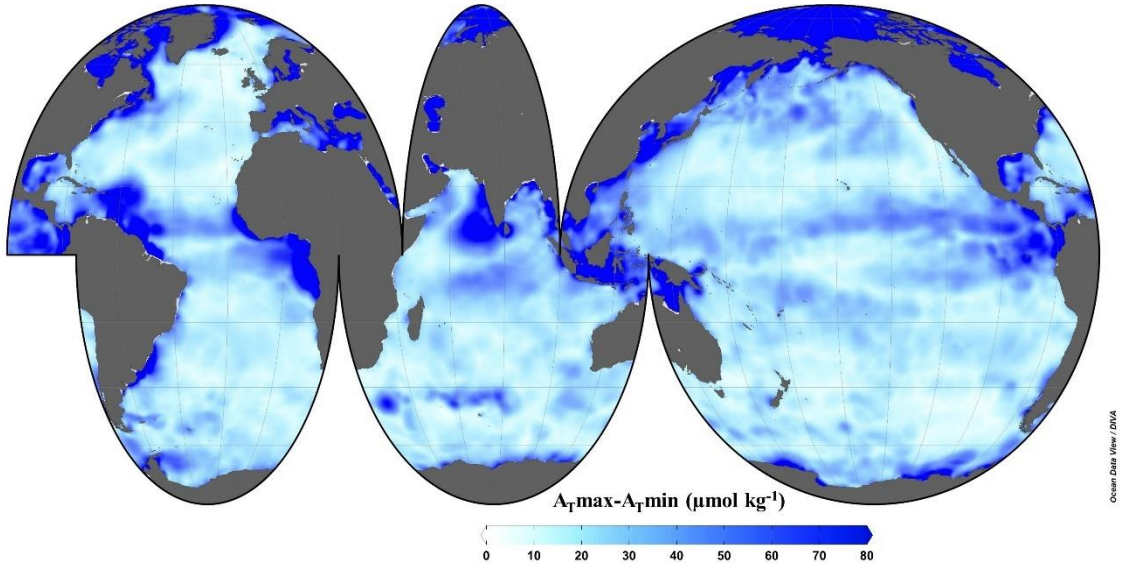


Figure 10: Seasonal amplitude of sea surface A_T. This figure was made with Ocean Data View (Schlitzer, 2016).

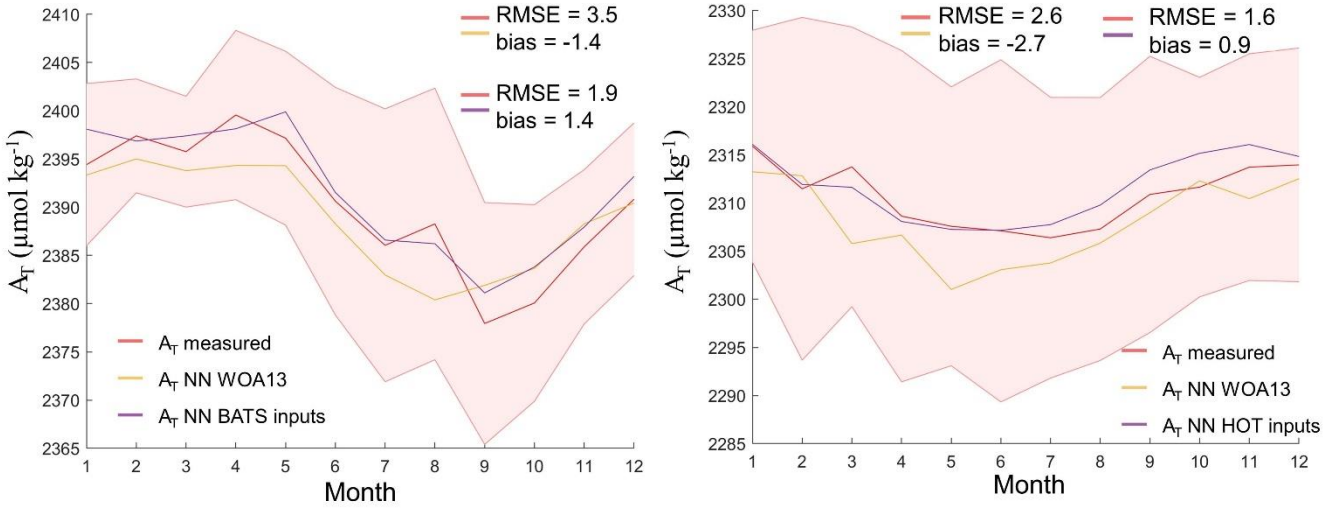


Figure 11: Climatology of A_T from measured data, from NNGv2 using measured data as inputs and from NNGv2 using WOA13 data as inputs at BATS (0-5 m; left panel) and HOT (0-30 m; right panel) time-series location. The shading represents the standard deviation of the average of the measured data. Units of RMSE and bias are µmol kg^{-1}

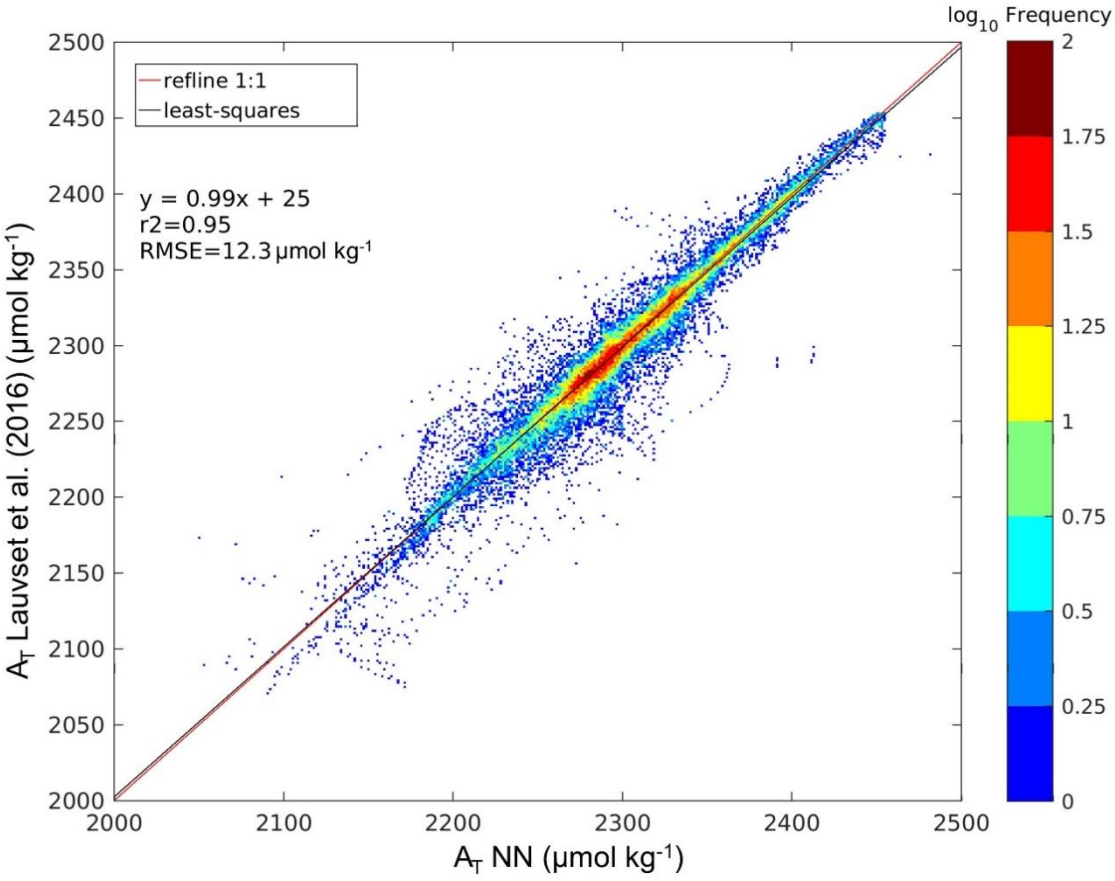


Figure 12: Regression between A_T computed with NNGv2 applied on the climatologies of Lauvset et at. (2016) and A_T from Lauvset et at. (2016) at 0m. The graph is divided in pixels. The color of each pixel is determined by the number of points inside it. Note the logarithmic scale to account for the large amount of data.

690 **Table 1: RMSE and bias between GLODAPv2 A_T and A_T computed by the neural network. n: number of samples. NNGv2: neural network trained with the initial dataset. NN±3RMSE: neural network trained with the dataset without samples with residuals beyond ±3RMSE. GLODAPv2: initial dataset. GLODAPv2w3RMSE: dataset without samples with residuals beyond ±3RMSE. lm: Levenberg-Marquardt. br: Bayesian Regularization**

Approach	RMSE (µmol kg^{-1})	bias (µmol kg^{-1})	n
NNGv2_GLODAPv2 (lm)	8.2	0.02	246221
NNGv2_GLODAPv2 (br)	8	0.03	246221
NNGv2_GLODAPv2w3RMSE (lm)	5.1	-0.002	243754
NN±3RMSE_GLODAPv2w3RMSE (lm)	4.8	-0.006	243754

Table 2: RMSE obtained by the relations of Lee et al. (2006), NNGv2 and NN±3RMSE over GLODAPv2. In bold the lowest RMSE in each area defined in Lee et al. (2006). To be consistent with the surface layer defined in Lee et al. (2006) the samples evaluated here are from above 20m (subtropics) and 30m (the rest).

	RMSE			
Areas defined in Lee et al. (2006)	Lee et al. (2006)	NNGv2	NN±3RMSE	n
North Atlantic	13.5	12.3	**12.1**	2765
North Pacific	16.8	10.5	**9.6**	2087
Equatorial Upwelling Pacific	7.8	9.5	**5.7**	481
Subtropics	20.8	**15.1**	15.2	4309
Southern Ocean	10.1	5.9	**5.3**	3610
Weighted RMSE	15.3	11.1	**10.6**	13252

Table 3: RMSE obtained by the relations of Takahashi et al. (2014), NNGv2 and NN±3RMSE over GLODAPv2. In bold, the lowest RMSE in each area defined in Takahashi et al. (2014). To be consistent with the surface layer defined in Takahashi et al. (2014) the samples evaluated here are from above 50m.

	RMSE (µmol kg-1)			
Areas defined in Takahashi et al. (2014)	Takahashi et al. (2014)	NNGv2	NN±3RMSE	n
West GIN Seas	29.2	**10.4**	11.8	623
East GIN Seas	11.6	9.5	**9.0**	990
High Arctic	24.9	**15.2**	16.2	594
Beaufort Sea	57.6	**46.9**	79.3	2086
Labrador Sea	27.7	22.7	**22.1**	736
Subarctic Atlantic	15.6	**10.4**	11.3	1041
North Atlantic Drift	7.7	7.9	**7.2**	1403
Central Atlantic	23.1	**19.9**	20.1	3276
South Atlantic Transition Zone	6.7	6.8	**5.9**	291
Antarctic (Atlantic)	7.5	5.8	**5.2**	727
Kuroshio-Alaska Gyre	16.2	10.8	**9.8**	1412
North Central Pacific	13.2	10.0	**9.4**	1224
Okhotsk Sea	**5.4**	7.8	**5.4**	20
Central Tropical North Pacific	9.3	7.3	**7.0**	1328
Tropical East North Pacific	30.9	11.2	**10.3**	308
Panama Basin	8.1	13.4	**7.4**	58
Central South Pacific	9.7	6.4	**5.8**	2834
East Central South Pacific	11.6	9.3	**8.8**	249
Subpolar South Pacific	8.2	5.2	**4.6**	431
Antarctic (Pacific)	4.9	4.3	**3.0**	524
Main North Indian	7.0	6.2	**4.6**	493
Red Sea	**6.3**	9.3	9.2	19

Bengal Basin	8.9	7.8	**6.3**	96
Main South Indian	8.8	7.1	**6.8**	2536
South Indian Transition	7.9	5.4	**3.8**	330
Antarctic (Indian)	8.1	5.0	**4.0**	865
Circumpolar Southern Ocean	10.1	5.9	**5.3**	1970
Weighted RMSE	17.0	**12.8**	15.0	26464
Weighted RMSE without Beaufort Sea	13.5	9.9	**9.5**	24378

Table 4. RMSE and bias obtained with NNGv2, LIARv2 (Carter et al., 2018) and CANYON-B (Bittig et al., 2018) in both GLODAPv2 dataset and GLODAPv2 dataset without the samples where AT QC was not done.

Approach	RMSE ($\mu mol\ kg^{-1}$)	bias ($\mu mol\ kg^{-1}$)	n
NNGv2_GLODAPv2	8.2	0.02	246221
LIARv2_GLODAPv2	11.4	0.08	246221
CANYON-B_GLODAPv2	10.2	0.1	246221
NNGv2_GLODAPv2_onlyQC	6.6	0.06	215332
LIARv2_GLODAPv2_onlyQC	8.2	0.06	215332
CANYON-B_GLODAPv2_onlyQC	6.8	-0.04	215332

705

Table 5: RMSE and bias between measured A_T and neural network computed A_T. r^2 from the regression between measured A_T vs computed A_T. The comparison was done for all the samples where the input variables and the A_T were measured in the same water sample.

Time-Series	Location	RMSE ($\mu mol\ kg^{-1}$)	bias ($\mu mol\ kg^{-1}$)	r^2	n
HOT	22º45'N, 158º00'W	**5.8**	**-0.8**	**0.99**	4010
BATS	31º40'N, 64º10'W	**6.2**	**-0.2**	**0.77**	3033
ESTOC	29º10'N, 15º30'W	**3.3**	**0.6**	**0.99**	1700
KNOT	44ºN, 155ºE	**4.7**	**-6.4**	**0.996**	1234
K2	47ºN, 160E	**3.1**	**-3.0**	**0.998**	561

710 **Table 6: RMSE and bias between measured A_T and the A_T computed with both LIARv2 and CANYON-B methods. The comparison was done for the same samples evaluated in Table 5.**

	LIARv2		CANYON-B	
Time-Series	RMSE ($\mu mol\ kg^{-1}$)	bias	RMSE ($\mu mol\ kg^{-1}$)	bias

		(µmol kg^{-1})		(µmol kg^{-1})
HOT	**6.6**	**-0.6**	**5.8**	**-0.6**
BATS	**6.3**	**0.1**	**6**	**-0.4**
ESTOC	**3.4**	**0.8**	**4.2**	**3.2**
KNOT	**4.8**	**-6.6**	**4.5**	**-7.2**
K2	**3**	**-3.0**	**3**	**-3.3**

Table 7: RMSE and bias obtained with the neural network trained without winter data in both GLODAPv2 dataset without winter data and GLODAPv2 dataset only containing winter data.

Dataset	RMSE (µmol kg^{-1})	bias (µmol kg^{-1})	n
GLODAPv2_nowinter	8.7	-0.3	225189
GLODAPv2_winter	6.8	-0.4	21032

715

Table 8: Comparison of four annual mean surface climatologies of A_T. *The Arctic Ocean and the Baltic Sea are not included in the comparisons for coherency reasons.

RMSE (µmol kg^{-1})\r^2	NNGv2	Lauvset et al. 2016*	Takahashi et al. 2014	Lee et al. 2006
NN		0.91	0.92	0.97
Lauvset et al. 2016*	15.7		0.90	0.92
Takahashi et al. 2014	15.3	17.8		0.93
Lee et al. 2006	8.0	14.6	12.4	

Table 9: Comparison between the three monthly climatologies of A_T.

Month	Lee et al. (2006) vs NNGv2		Takahashi et al. (2014) vs NNGv2		Lee et al. (2006) vs Takahashi et al. (2014)	
	RMSE (µmol kg^{-1})	r2	RMSE (µmol kg^{-1})	r2	RMSE (µmol kg^{-1})	r2
January	12.6	0.93	18.5	0.89	14.2	0.92
February	12.2	0.94	24.2	0.82	14.7	0.91
March	12.1	0.94	19.5	0.87	14.3	0.91
April	12.1	0.94	18.4	0.88	15.0	0.91
May	12.4	0.93	19.0	0.86	13.8	0.92
June	12.7	0.93	17.7	0.89	14.3	0.91
July	12.3	0.93	24.9	0.84	14.8	0.91
August	12.9	0.93	19.5	0.89	14.8	0.91

September	12.5	0.93	17.9	0.91	14.9	0.91
October	11.9	0.94	20.8	0.88	13.1	0.93
November	12.0	0.94	27.9	0.80	12.8	0.93
December	11.7	0.94	18.9	0.89	13.9	0.92

720