Review of 1st manuscript revision of "A global monthly climatology of total alkalinity: a neural network approach" by D. Broullón for ESSD

April 5, 2019

Thank you for revising and trying to improve your manuscript. The authors better explain their network training procedure and what dataset they use for which statistics. They also make better comparisons of their method to other state of the art approaches (though there is still room to make them adequate, see below).

However, the manuscript did not improve sufficiently. There are still significant shortcomings, which require major revisions.

Based on the initial submission, their approach showed good potential for a fully seasonal, monthly climatology, but the authors fail to clear the concerns raised by all three reviewers in this revision. This encompasses both the seasonality of their NNGv2 network based on GLODAPv2 training data (I am afraid their 'reinforcement' of the subsurface layer hypothesis does not reinforce anything in its present form.), as well as the seasonality of the produced A_T climatology from WOA13. It encompasses as well the need for a more robust uncertainty assessment of the produced A_T climatology.

Find below my comments ordered from critical to major to minor.

1 Critical points

1.1 Seasonality at depth

There is a lack of transparency as to what depth the climatology of A_T is seasonal or not.

l. 406 states that "seasonality disappears almost completely below 500 m depth; not surprising due to the lack of seasonal resolution in the climatologies of nutrients in WOA13 below this level." but otherwise, the authors claim to provide a "global monthly climatology of A_T on 102 depth levels" (l. 223 and abstract l. 31), i.e., to full depth?

WOA13 input data have monthly resolution down to 500 m for the nutrients, and down to 1500 m for T, S, and O_2. WOA13 input data have quarterly-resolved files down to 500 m for the nutrients, and down to full 5500 m depth for T, S, and O_2. Finally, WOA13 has annual mean files down to full 5500 m depth for nutrients, T, S, and O_2.

The quarterly-resolved fields of T, S, and O_2 show in some parts strong differences at high depths, e.g., at 3000 m (see attached figure 1, left column). The authors appear to interprete that as seasonality ("The seasonal amplitude of A_T is progressively reduced at depth" l. 405f/Figure S7, and l. 310-315)?

For perspective, WOA13 quarterly variations at 3000 m depth have (e.g., in the Southern Ocean) a range up to 0.4 °C (on a total range of variability of ca. 40 °C), 0.05 psu (on a total range of ca. 2 psu), and beyond 40 μmol/kg (on a total range of ca. 400 μmol/kg). This certainly exceeds any expectation for a seasonal cycle, e.g., of oxygen, and demonstrates rather a data coverage and/or mapping issue. (Please consult / check the data coverage fields 'x_dd'!)

This has three consequences:

1. The authors must decide until what depth they use which monthly-/ quarterly-/ annually-resolved WOA13 input fields, determining until what depth they can claim to provide a monthly-/ quarterly-/ annually-resolved A_T climatology.

   This must be 500 m, if they decide that their seasonality is caused by the organic matter cycle, reflected through both oxygen and nutrients variations (l. 319/321) (summed 57 % relative importance). It may be 1500 m if they decide that *monthly*-resolved oxygen (16 % relative importance), together with *annual means* of the nutrients (summed 41 % relative importance), may provide a sufficient driver for A_T seasonality to the NNGv2 network, but this already must be clearly justified. It would be quite a stretch for any reasonable seasonality below 1500 m, and I would suggest to revert to annual mean WOA13 fields, rather than the quarterly ones, to built the A_T climatology. Please consult the 'x_dd' data coverage fields, too, to assess if sufficient data went into the monthly / quarterly fields for a robust seasonality – or a robust assessment at all.

2. The seasonality and its limits must be made transparent through the author's work/manuscript. It should not be the task of the reviewer / user to check.

3. The CANYON-B/WOA13 comparison (l. 310-315) must be moved to the Climatology section 3.4 rather than the Neural network analysis 3.1 and discussed accordingly. A computation method "using relatively few input variables (position, time, temperature, salinity and oxygen)" (l. 305) is more prone to bad input data in one variable than a method that uses all the variables as inputs (l. 309). Particularly, (1) if the only biogeochemical predictor may be biased (oxygen; CANYON-B) rather than just one out of a total of three (LIARv2) or four (NNGv2), of which the three nutrients nitrate, silicate, and phosphate just have a WOA13 *mean annual* field below 500 m (!); (2) if exactly this predictor has a "seasonal" O_2 amplitude

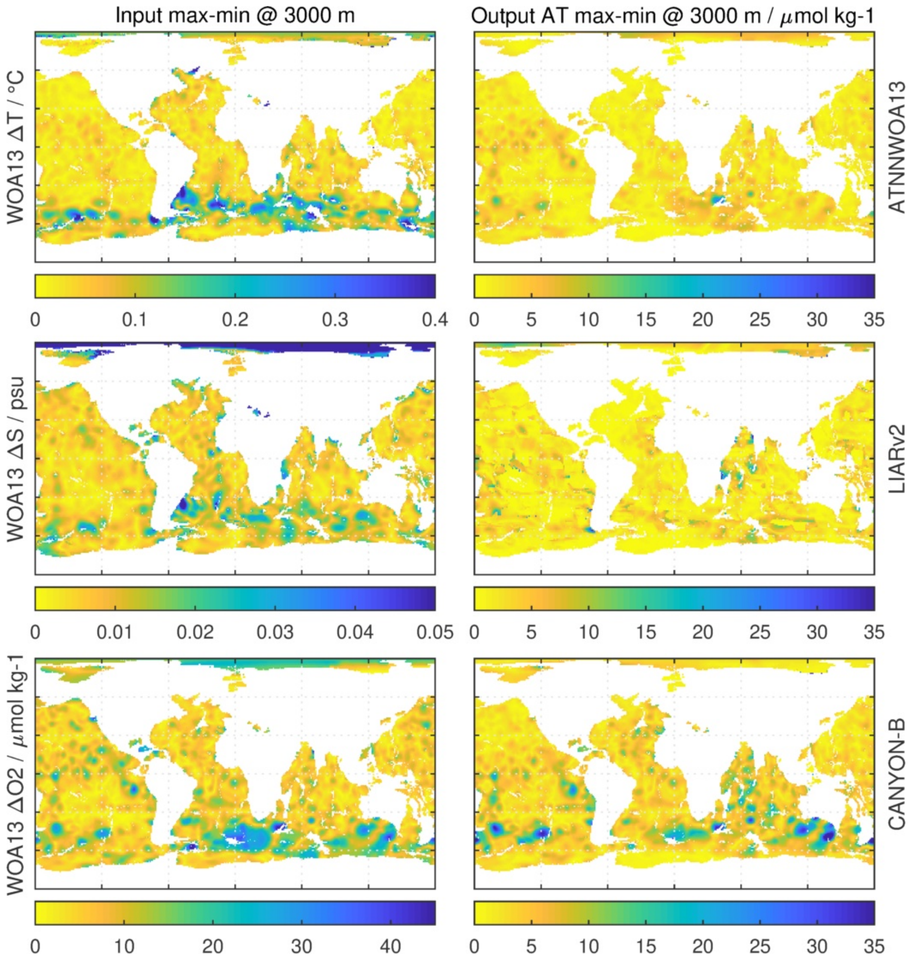

Figure 1: Left column: WOA13 "seasonal" amplitude at 3000 m of temperature (top), salinity (middle), and oxygen fields (bottom) from WOA13 quarterly-resolved files, which are (probably?) used as input to the A_T climatology and for CANYON-B 'comparison' (l. 310-315). Right column: A_T "seasonal" amplitude at 3000 m for the author's *monthly* climatology (probably?) based on WOA13 *quarterly* fields of T, S, O_2 and *annual mean* fields of nutrients as input (top), LIARv2 using WOA13 *quarterly* fields of T, S, O_2 and *annual mean* fields of nutrients as input (middle), and CANYON-B using just WOA13 *quarterly* fields of T, S, and O_2 (bottom). Note the correspondance of elevated patches of "seasonal" amplitude between WOA13 quarterly O_2 and CANYON-B A_T in the Southern Ocean (a.k.a. 'garbage in, garbage out').

> at 3000 m of up to 40 μmol/kg (versus the time / decimal year with a pretty modest variability of ca. 0.5 years on a total range of 40 years!); and (3) considering the strong correspondence at 3000 m between "seasonal" oxygen and "seasonal" A_T amplitude (attached figure 1 right column).
>
> To claim / blame the time for these variations in A_T is pretty bold and wrong.[1]
>
> Please correct the text, if you decide to keep it, and please make an effort for a balanced comparison of methods!

Please also correct the conclusion (l. 471/472), which neglects the impact of the quality of the WOA13 seasonality.

1.2 Uncertainty of climatology

Why is the monthly A_T climatology not compared to the measured GLODAPv2 A_T data? This would give a much more robust assessment than just two time series sites (HOT and BATS), at which the WOA13 input data arguably should be on the better side of the spectrum of possibilities, i.e., underestimating the climatology's uncertainty.

I don't see why such a comparison should be limited to *locations* with repeated sampling (l. 415) and not extended to *times/months* with repeated sampling (read: basin-crossing cruises as in GLODAPv2).

At the end of the day, the A_T climatology should represent both temporal and spatial variability within its resolution – What better dataset to assess temporal and spatial variability than the largest available one, GLODAPv2? This should also include spatial / regional differences in the uncertainty.

At least some assessment of A_T climatology uncertainty must be given before the dataset is acceptable for publication.

1.3 Subsurface layer hypothesis

Quoting from the text, the "winter relationship between inputs and A_T needed to produce an all-season surface climatology are mostly preserved in [the] subsurface layer." (l. 214).

However, the authors try to reinforce the hypothesis (1) using the nowinter network on all depths rather than just the surface, and (2) the NNGv2 network is never evaluated on the same data as the nowinter network. In fact, the numbers in table 7 might be nice to show, but don't give any indication about the validity of the subsurface layer hypothesis.

[1] In addition, year-to-year A_T variability, i.e., decimal year +/-1.0, is almost negligible in CANYON-B A_T with WOA13 input. If the "seasonal" A_T variations were to be caused by the decimal year variable, the representation inside the neural network would have to be strongly oscillatory, which contradicts the principle of early stopping / regularization to produce smooth network representations (e.g., l. 134-140; Hagan et al., 2014).

The lower winter RMSE may just be related to less variability (at all depths?) during this one season compared to the three nowinter seasons. Many other reasons are plausible, too.

What the authors should do: (1) As the authors suggest and do in l. 370, control that the nowinter network is comparable to NNGv2 on the domain it is trained on by providing statistics for NNGv2 and nowinter network on the GLODAPv2_nowinter dataset (full depth and surface; can be moved to the supplementary if desired); (2) Provide statistics for NNGv2 and nowinter network on the GLODAPv2_winter dataset using ***only surface*** data (above the subsurface layer defined in lines 358-362). Only if they are comparable, or at least not exceedingly higher for nowinter over NNGv2, or not exceedingly higher than surface RMSEs in other seasons (e.g., GLODAPv2_nowinter dataset surface only), this would reinforce the subsurface layer hypothesis. (That is, the exclusion of the (scarce) winter data did not degrade the winter surface predictions (← nowinter network and NNGv2 network on GLODAPv2_winter surface data) thanks to the still present signature in the spring subsurface layer (← GLODAPv2_nowinter training data and full GLODAPv2 training data).)

Other than that, the subsurface layer hypothesis remains a hypothesis, which I'd doubt the NNGv2 to recognize and would suggest to remove.

Figure 8: Same question as before: Why are calculated NNGv2 A_T values not shown? They should.

2 Major points

- Table 2/3: Why are LIARv2 and CANYON-B not added here? They should be added!
- l.295. "The newest methods in the A_T computation (...) model the GLODAPv2 A_T with higher errors than the NNGv2 (Table 4)." **This is because both LIARv2 and CANYON-B used only the GLODAPv2 A_T subset for training where the 2nd QC was done, whereas our GLODAPv2 A_T data for training included samples, too, where the 2nd QC was not done.** "An analysis in a GLODAPv2 subset excluding the samples where the 2nd QC was not done for A_T shows a reduction of the error [...]".

  NNGv2 results are not comparing to independent data in the GLODAPv2_no_secondary_QC subset because of correlations within cruises and the random splitting of cruises between testing/training, whereas LIARv2 / CANYON-B truly haven't seen any of these data.
- Table 5 and 6 should be merged, such as table 4.

- I still find it hard to justify the NNGv2_3RMSE network. The only clue to this one is that you remove an area (Arctic Ocean) with higher-than-average variability and, naturally, get better statistics. If you remove the same area from the NNGv2 assessment, you get the same, better statistics, too (table 1)!

  There is still a lack of justification for the NNGv2_3RMSE, and no, I don't think that a few decimal places better RMSEs in a few out of the regions in table 3 justify the 3RMSE removal – you can get the same better performance here or there by having a closer look at the 10 NNGv2 networks you trained! Also, the conclusions from the authors response are not supporting their argument and are not substantial. ("[...] In this case, it is clear that omitting certain data causes a large difference between the networks." I don't see a large difference. If you insist, please use an appropriate test to verify significance; the "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." That's only what the authors want to see, I'd still see a different local minimum as more plausible. And no further evidence is given that this may not be the case.)

  Please improve the NNGv2 vs. NNGv2_3RMSE network aspect or remove either one of the two.

- What about the seasonal amplitude of A_T at the time series sites of measured A_T vs. NNGv2-measured inputs-based vs. NNGv2-WOA13-based?

- To be complete, the subpolar North Atlantic should still be added to the current manuscript as test region for the current methods, even if it was the object of a previous work (Vázquez-Rodríguez et al., 2012).

3 Minor points

- l. 316: "The NNGv2 seems to associate the A_T variability to the predictor variables *in coherence with the processes that contribute* to it."

  So, does it? Please give evidence or remove/rephrase.

- Table S1: What does 'HS' mean?

- Table S2/S3: Column headings 'relative ... lat>60° N' should probably correspond?

- l. 30: missing subscript A_T