Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2018-111-RC3, 2018

Open Access Earth System Science Data Discussions

Interactive comment on "A global monthly climatology of total alkalinity: a neural network approach" *by* Daniel Broullón et al.

Anonymous Referee #3

Received and published: 19 December 2018

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.

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.
- 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.

(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.

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.

(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.

(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).

Minor points:

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

l. 105: The number of neurons in the output layer is adjustable? It seems to be n=1 for just AT, isn't it?
l. 124: "as previously described" - not yet done, remove.
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.
l. 238: "As an argument ... areas." Unclear.
l. 273: Depth is rather associated as vertical sampling position.
l. 279: Any ideas why there is such a bias? Should be commented.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2018-111, 2018.