Dear Gloria Pietropolli, Luca Manzoni, and Gianpiero Cossarini,

I would like to thank the authors for their efforts to improve the manuscript. It now reads much more smoothly and the consistent use of terminology to describe items/aspects make it much easier understandable.
The Med Sea focus is still a bit hidden, only to appear in the last sentence of the abstract. I would prefer this info to be presented more up-front, but I guess this is at the author's discretion.

However, there is one aspect where I dissent with what the authors claim. They write:

> l. 8f. [1]: "However, MLPs lack awareness of the typical shape of biogeochemical variable profiles they aim to infer" ('claim *a*')

and continue:

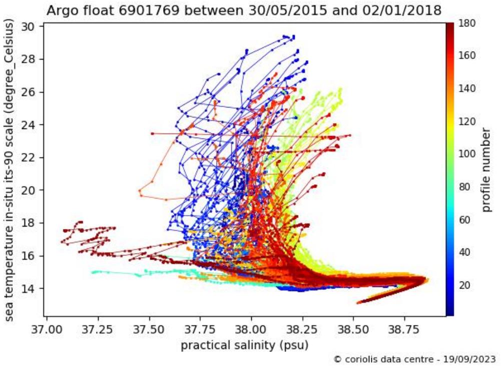
> l. 9f.: "resulting in irregularities such as jumps and gaps when used for the prediction of vertical profiles.". ('claim *b*')

As written in my previous comment, (1.) I don't think either of the two statements is true, and (2.) I don't see evidence presented by the authors to convince me otherwise (more on 2. later).
I see and appreciate in the remainder of the manuscript that these two claims have been toned down, compared to the initial version. However, I don't think they are justifiable.

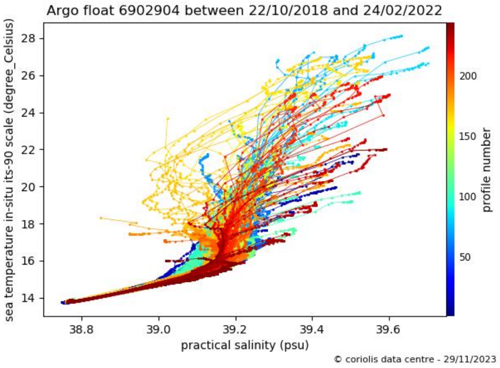
To **claim *a***: The authors are correct in that MLPs act and are trained on point-wise data, whereas CNNs take strides of data (e.g., profiles) and consider both their value and their arrangement (e.g., profile shape). To use CNNs to predict some profile data from other profile data, taking benefit of the shape of that other profile data, is a (promising) step forward compared to MLPs predicting some profile data.
However, that does not mean that MLPs for Ocean prediction are agnostic to their neighbouring data points. Instead of an explicit shape awareness like CNNs, MLPs have an implicit awareness of the typical shape of profiles they want to infer. Why?

Because the Ocean is smooth.

The parameter space in the Ocean is continuous and smooth (with a large thanks to mixing). Just for illustration: Below two T-S diagrams for two of the floats used. (Quality-controlled) Ocean data are smooth to start with.


Argo float 6901769 between 30/05/2015 and 02/01/2018

© coriolis data centre - 19/09/2023


Argo float 6902904 between 22/10/2018 and 24/02/2022

© coriolis data centre - 29/11/2023

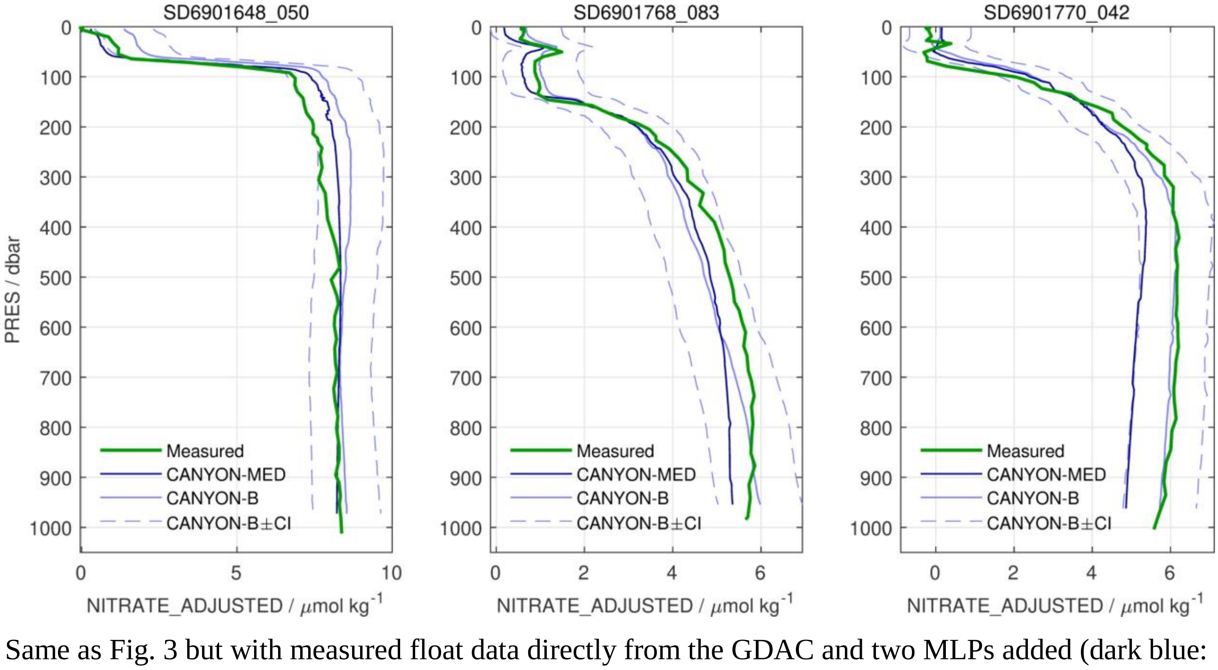
Going along a profile step-by-step (both in parameter space or against depth) gives only small, step-wise modifications of the variable to be predicted and the variables used as predictors alike. Which means that, at any given point in parameter space, you have a certain idea of how your environment will look like: Probably not an awful lot different, but just a little. I.e., even if only given point-wise knowledge at a time, MLPs do have an awareness of how their profile will look like nearby (i.e., not a

[1] Line numbers refer to the tracked-changes manuscript version "egusphere-2023-1876-ATC2.pdf".

lot different). I.e., there is (some) spatial awareness of profile shape in MLPs, too, due to point-wise proximity in parameter space and due to the Ocean's parameter space being smooth.

To **claim *b***: Starting off from (*i*) a smooth parameter space (previous point), and (*ii*) using a neural network (MLP) architecture complexity suited for the size of the training data as well as (*iii*) ensuring that there is no overfitting by properly selected regularization, then the (MLP) neural network outcome must be an approximation of the parameter space trained on. If properly regularized, the neural network is by definition smoother than the training data from (*i*). If not overly complex, then the network's weights and parameters are sufficiently constrained by the training data from (*i*), so that the (MLP) network represents a continuous function (no poles, irregularities, gaps). If the training data are smooth to start with, then that cannot cause irregularities or gaps either.

So I am left with claims where my arguments run against them. The two claims are brought up again in l. 68-73 but without illustration. So when coming to Figure 3 (nitrate profiles for selected floats), I wondered why the measurements and PPCon are shown but the Pietropolli et al. 2023a MLP comparison (as well as other MLPs for comparison like CANYON-MED, CANYON-B) were dropped?
Thank you for adding the WMO and profile numbers, which allowed me to go search for the data and do those comparison plots on my side. Here's what I get for the measurements as well as the two MLPs where I had access to the code for prediction: CANYON-MED (specifically trained on the Med Sea with a dataset extended beyond GLODAPv2) and CANYON-B (trained with very little Med Sea data from GLODAPv2, so no great performance to expect; but it provides confidence intervals, which I find instructive):



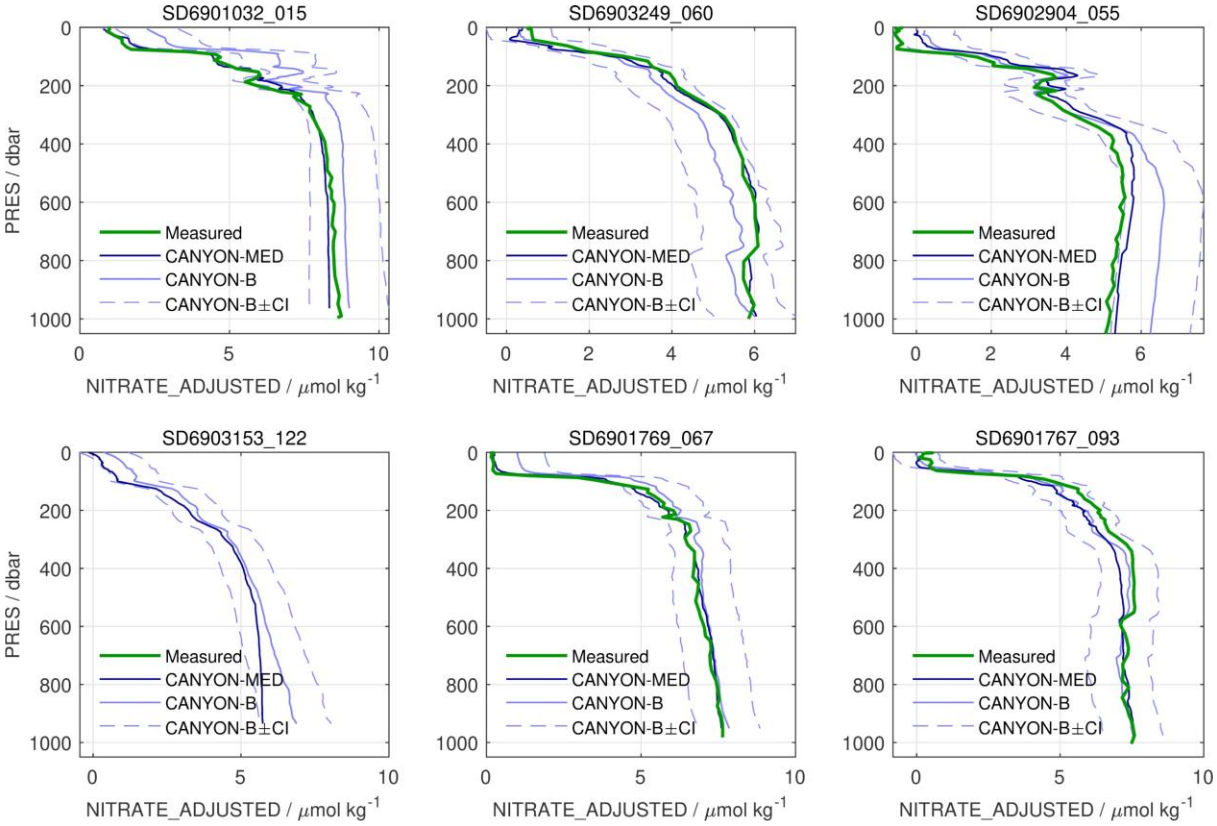
Same as Fig. 3 but with measured float data directly from the GDAC and two MLPs added (dark blue: CANYON-MED, light blue: CANYON-B).

From the comparison, I find it rather comforting and confirming my line of argument that the MLPs give a similarly smooth profile shape as the CNN (in the manuscript Fig. 3) and no irregularities or jumps in the profile – unless governed by the water mass properties (e.g., middle profile just shallower than 50 dbar) and seen in the measured nitrate, too. (As said, I could not confirm/falsify whether the EMODNET-based Pietropolli et al. 2023a MLP shows irregularities or jumps in the profile.)

What I find a bit discomforting is that the measured data in my case looks not the same as shown in Figure 3 of the manuscript: All profiles in Fig. 3 are much smoother (e.g., base of the mixed layer is significantly eroded; interleaved water mass around ~50 dbar in middle panel fully absent) and also the bottom portion of the middle profile has a different shape (and value). I have confirmed the profile's locations and dates that I used – they match to the floats and cycle numbers (Thank you for Table 5)!

The two claims are then taken up again in the results (l. 317-318; no illustration) and later on in the discussion (l. 411-412; no illustration) and the reader is referred to Appendix B (l. 392; l. 414).

This Appendix B (comparison between reconstructions by PPCon and MLP architectures) is very promising. And after experience with Fig. 3 I tried to redo the figure B1 on my end, too:



Same as Fig. B1 but with measured float data directly from the GDAC and two MLPs added (dark blue: CANYON-MED, light blue: CANYON-B).

From my perspective, both MLPs do an excellent job in reproducing the measured profile (even CANYON-B, for which no great things should be expected in the Med Sea). Both MLPs even include proper reproduction of variability caused by interleaved water masses along the vertical profile.

However, I see a couple of discrepancies to the manuscript Fig. B1:

- The measured data is again of different shape, value, and smoothness – for all profiles shown. None of the fine scale features are visible in the manuscript (from what I can eye-ball).
- The fine scale features are pretty much mirrored in CANYON-MED; and CANYON-MED seems to match between my figure and manuscript Fig. B1, unlike the measured data. The CANYON-MED MLP is also spot-on on most of the measured profile (in my figure; not in the manuscript Fig. B1).
- The float 6903153 in the ION (lower left panel) does not carry a nitrate sensor at all.

These aspects are quite discomforting and raise an eyebrow on whether there are similar discrepancies in other parts of the data, and on which data were compared with what measurements with respect to performance (RMSE) of the various ANNs, both PPCon and others.

There are a few instances throughout the manuscript, where WMO numbers got a bit scrambled, so it might be as simple as incorrect WMOs and cycles labelled. Or, there may be a more profound issue with the data used, or a mess-up in measured profiles compared to actually different MLP/PPCon predicted profiles. This would be dramatic. In any case, these issues need to be addressed before a publication.

(a) Where do the discrepancies in the measured data come from? How large is the extent of the discrepancies? What's the impact on the method and its evaluation?

(b) Both from the line of argument presented in this comment, as well as from the illustrations (reproduced Figs. 3 and B1), I cannot recognize that either of claim *a* or *b* can stand. (Rather, MLPs can do a surprisingly good job in reproducing profile shape thanks to (point-wise) water mass characteristics.)

   I would therefore ask you to remove those claims entirely, or to substantiate them. (Again, my perspective neglects the Pietropolli et al. 2023a MLP, where I don't have access to the code. It may be that some of your claims *a* or *b* may apply to and be true for the Pietropolli et al. 2023a MLP. But then it cannot take hostage of all MLPs, given that it doesn't apply to other MLP models or architectures.)

(c) Could it be that fine scale characteristics of the vertical profile, like interleaved water masses, cannot be well reproduced by the CNN-based PPCon model/architecture? Because it puts more emphasis on the entire/large-scale profile shape and 'neglects' the information from the water mass properties, which is the only information available to MLPs?

(d) Why are the measured data in the manuscript without fine scale and with mostly eroded mixed layer base?

Further points:

- Figure 1: I cannot find float 6901767, the second nitrate float used later on (Fig. 7 and Tab. 8). Or is it a WMO mix-up there (and it's 6901648??).

- l. 133: "Table 2" referred to is missing.

- l. 134: From the description of Amadio et al. 2023 (and the above experience) it reads that the float data were treated with the "bit.sea python package (Bolzon et al., 2023)", including "a smoothing task". For a manuscript that focuses and puts emphasis on profile shapes, this is important information that belongs into the description of the dataset (and cannot be hidden somewhere inside some reference). And which warrants some (brief) discussion, because it seems that your figure's mixed layer bases are much more eroded than in the original float data. Which may have implications if PPCon were to be used for augmenting a float dataset.

- section 3.1: The question remains on why such a complex architecture was chosen and what should be gained from 4 separate MLPs per singular input (over replication of input data into a 200x1 vector each) to be fed into the CNN. But I won't insist and just note that this is still weakly motivated.

- l. 182: "adding zero padding to the borders of the input tensor": This requires the input to be normalized, i.e., mean data subtracted (and ideally divided by the standard deviation). A batch normalization is mentioned only for the output tensors in the text, not for the input tensor. From table 2 I understand that batch normalization (BN) is done for every layer's input, is it? So maybe follow the logic as it is outlined in the table also in this paragraph: Layers consist of BN, SELU and have Dropout. To compress/decompress information we have kernels/strides, which require padding with ...

- l. 314: "For the nitrate variable, the reconstruction performed by the MLP model is also reported (Pietropolli et al., 2023a)." No, it's not – but it should! :-)

For Figure 3 and also Figure B1, I'd also suggest to use a bit wider spectrum of colour than different shades of blue, to be able to better distinguish the different models.

- l. 317f: "...than the previous MLP architecture by Pietropolli et al. (2023a), but similar to MLPs by Fourrier et al. or Bittig et al."? But my general recommendation would be to drop the second part of the sentence.

- l. 472: That predictions tend to better in deep waters compared to surface waters is true for of all approaches; it's not unique to PPCon. It's a feature due to the different variability of the Ocean's parameter space at depth vs. at the surface, nothing of a particular model's architecture.

- Fig. B1 caption: 6903153 in ION has no nitrate sensor.

Minor points:

- l. 42: replace "frequency" by something more fitting? Maybe

- l. 133: "Table 2" referred to is missing.

- l. 145: sample date?

- l. 303: "in the ION[+, SWM, and TYR,] with RMSE values below 0.5 [+unit]"?

- l. 306f.: check "which are the highest ..." – Does this still apply?

- l. 395: "DT" was this abbreviation introduced? Why not spell it out

Typo's:

- quite a few instances: "BCG" instead of "BGC"

- l. 41: closing ")"

- l. 263: α_s ?

- l. 377: closing ")"

- l. 403: understand

- l. 439; BGC-Argo network

- Fig. B1 caption: 6902904

- l. 102/l. 509: Cross-check the Bittig et al. 2019 reference. That's the one you intended? (technical documentation vs. a published manuscript?)