Anonymous Referee #1
Summary:

This paper presents an approach for predicting the vertical profiles of temperature and salinity over the top 1000 m from satellite surface observations by training an empirical machine learning model using in-situ profiles in the western North Atlantic Ocean. The paper emphasises the treatment of the mixed layer depth and, specifically, a procedure to remove negative stratification from the profiles.

Overall, I found the paper to be an interesting contribution with sufficient novelty to be valuable. The ultimate impact of the work remains to be seen, but I think the paper will be worthy of publication after revision.

We thank reviewer #1 for his review. It allowed to clarify the text in several places. We changed the RMSE to a normalized RMSE that is indeed better suited. We also tried to compare OSnet to other climatologies (Roemmich and Gilson 2009, ISAS) but did not find it very instructive since there is a significant mismatch in temporal and spatial resolutions in all these products.

My major concerns are as follows:

I find it surprising and confusing that the paper does not carefully separate capability to model the 4-D climatological annual-cycle from capability to model 4-D anomalies from this climatological annual cycle. Perhaps such an approach is superior, and the methods are fine as they are, but the evaluation should clearly separate errors in the climatology from errors in the anomalies therefrom. I think the paper would be stronger if it included more explicit and quantitative evaluation of model performance on anomalies from the climatology (Nonetheless, I like the illustrative examples).

We think section 4.5 explicitly shows that OSnet captures well the surface seasonal and inter-annual variability (Fig. 12 and 13). Figure 12 is especially showing the anomalies from the climatological annual cycle (without the small frequency noise too). We only plotted the surface trends because it is comparable with SSS and SST.
Regarding the spatial anomaly relatively to the climatology : we note that in the Gulf Stream the time mean climatological field is not a relevant physical state and is never observed. A “steady” Gulf Stream is intrinsically unstable and hence never shows a laminar straight path but rather meanders and detached eddies. So the mathematical decomposition in annual cycle and anomalies is not relevant here. The eddies are an intrinsic component of the signal. No action is taken for this comment.

Relatedly, given that the method predicts the climatological annual cycle, I think the paper would be stronger if results were compared to a climatology obtained by objective mapping or optimal interpolation, e.g. updated Roemmich and Gilson 2009 gridded Argo climatology or the mean of the CORA gridded product.

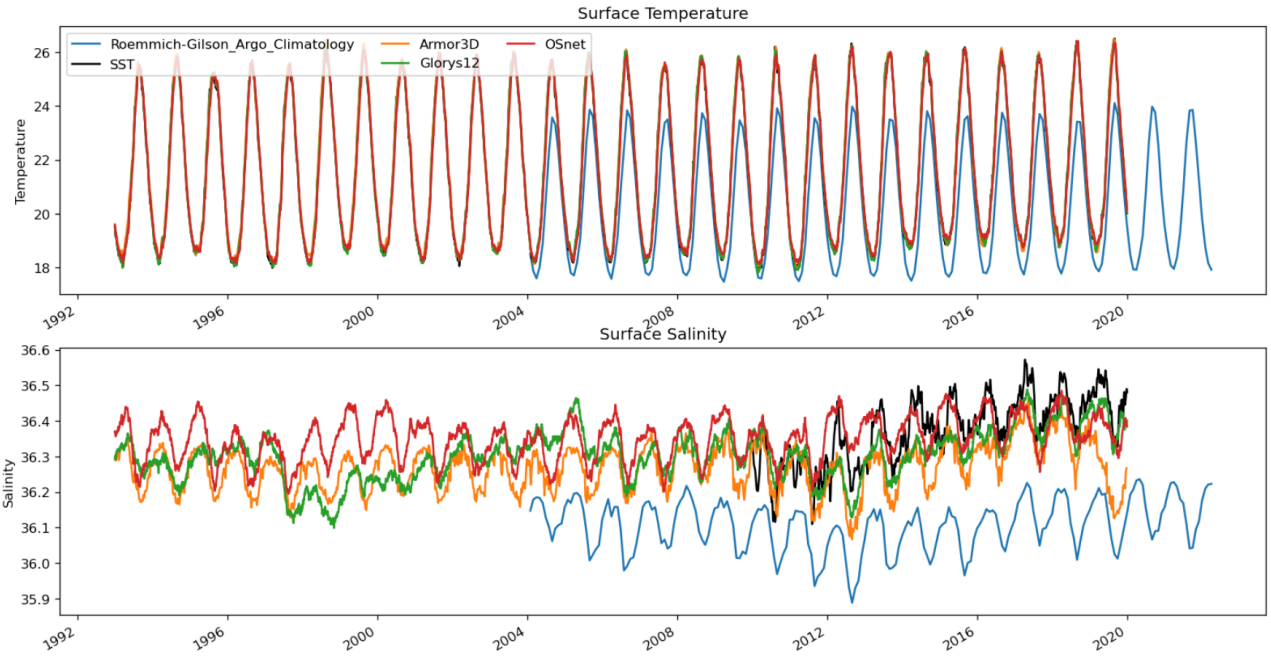
—> L116 We added this explanation “We compare OSnet gridded fields to Armor3D and Glorys12 because they are the only ocean products, to our knowledge, that extend from 1993 to today with a spatial resolution of at least 1/4 of a degree and a frequency under the month (weekly for Armor3D and daily for Glorys12).”

We tried the comparison with Roemmich and Gilson 2009 Argo (RG2009) climatology but their spatial resolution of 1/2 a degree makes it hard to have the same spatial coverage than us, especially near the shelf where several cold and fresh grid points are shifting their mean value compared to ours (see the figure below with the surface temperature and salinity mean timeseries where we see the RG2009 temperature colder and salinity fresher than all other products).

More importantly, the RG2009 climatology only extends from 2004 to today, so their annual cycle would be biased towards those years. Or we could compute the annual cycle for OSnet only for

2004-2019 to validate the RG climatology which is not the scope of our paper. Similarly the ISAS gridded fields (https://www.seanoe.org/data/00412/52367/) extends from 2002 to 2020 with 1/2 degree resolution and monthly means only.

Finally we believe that if OSnet compares well with Glorys12, it has to compare well with any climatology as well, as Glorys12 is constrained by the same observations than other climatologies.



In the training, it seems that the selection of cross-validation data does not account for spatial and temporal autocorrelation. It is not clear that the testing data are independent of the training data. Perhaps this is ok, given that you're trying to predict or map the climatology. But, the paper would be stronger if more explicit effort was made to train and test on truly independent data (at least with regard to modelling the anomalies).

It is true that our test and train data are not independent.
—> L152 We added a mention of that caveat in the method "Be aware that the train and test data are not truly independent, the selection is random without accounting for spatial and temporal autocorrelation."
However if the *test* and *train* datasets were not completely independent, we would have exactly the same RMSE for both, but in fact there is a small difference for each model. Predictions from *test* data are always worst than those from the *train* ones.

About "you're trying to predict or map the climatology.", we believe there is a confusion here, we predict daily ocean stratification on a grid, from which we can extract averages (the climatology).

Confidence intervals or uncertainty. I'm a bit confused about how these are calculated and thus how to interpret them. The paper would be stronger if this was clearer.

The confidence intervals are the standard deviation of the 15 predictions (from the 15 models of the bootstrap). This is stated in the method L166 :"Overall, given 15 trained models, we compute the mean T, S and K profiles for each input data and their standard deviation (Fig. \ref{RMSE}, grey). The latter deliver an estimate for a confidence interval."
—> We changed the wording to "confidence interval" instead of uncertainty when it appears in the text to clarify.

—> We also added a text in the caption of the figures when the confidence intervals appear to repeat the information “[…] i.e. the standard deviation of the 15 bootstrapped models.”

The main quantitative metric used is root-mean-square-error in physical units. I appreciate that this is physically intuitive, but this may obfuscate the generic statistical properties of the predictions. The paper would be stronger if normalized error metric were included, e.g. some sort of relative error and correlation.

Thank you for this proposition. We agree and changed all occurences of RMSE to a normalized RMSE. We divide the RMSE by the standard deviation of the measured property, by depth. It is a percentage that can now be compared with predictions from other regions or other dataset.

The word “coherence” is used a lot to refer to a desirable property of the 4-D gridded fields. Is this related to the frequency/waveform of the signal?? I’m not sure I understand exactly what is meant by coherence and why it is a valuable property of the predicted field. For example, in some cases, it may be that "smoothness" is unrealistic, e.g. in MLD predictions from GLORYS. Is coherence related to smoothness?

We use the words “vertical coherence” 3 times to talk about the MLD prediction accuracy and the absence of density inversions. We explain each time what we mean: e.g. “ the presence of density inversions and the accuracy of the MLD prediction.”
We also use the word coherence L294 to characterise the ARMOR3D fields of MLD that are often patchy (Figure 8d), it is explained in the sentence.

Be more specific about what properties of a gridded T/S dataset make it useful for interpreting local oceanographic measurements or for process studies. I’m not sure what you mean? Low error? Correlation with real variability

We want to obtain a product that is as close as possible from observations, while being physically consistent. We already introduce this goal and how it is important to have a proper MLD for climate studies (paragraph L55 to L65).
—> L357 we added a sentence to repeat this goal after the “for interpreting local oceanographic measurements or for process studies” sentence : “The goal is to be as close as possible from observation, while being physically consistent.”

There are several areas where minor typographical and grammar issues need to be corrected.

We corrected several english mistakes and typos thanks to the second reviewer and our careful re-reading of the manuscript.