Reviewer 1:

García-Espriu conduct an observing system simulation experiment (OSSE) to evaluate the feasibility of reconstructing ocean interior temperature and salinity from in situ observational data and satellite observational data products. The authors leverage output from the CMEMS Global Ocean Ensemble Reanalysis product to conduct this experiment, and they subsample the product at times and locations where Argo float profiles are available. They then use these subsampled synthetic profiles to train machine learning models, which they apply to satellite products to reconstruct ocean interior properties, and they compare these reconstructions against the reanalysis "truth" to evaluate the skill of the reconstruction methods.

The authors find that the more complex versions of their random forest regression (RFRv2) model and Long-Short Term Memory (LSTMv2) network are able to reproduce ocean temperature with an R^2 of around 0.85 and salinity with an R^2 of around 0.95. They validate their models with synthetic profiles withheld from model training and by using a regional subsection of the reanalysis dataset. They report validation statistics spatially and by depth, concluding that the RFRv2 model performed better in terms of the evaluation statistics against the test dataset but the LSTMv2 model was better able to represent the data in terms of variability over time and space. The authors also use SHapley Additive exPlanations (SHAP) to interpret their trained models.

Overall, I support the approach this manuscript takes to the question of how in situ and satellite observing systems can be leveraged to reconstruct ocean interior properties. However, it falls short in its execution and interpretation of the analysis. Most importantly, the authors could attempt to remedy or discuss more extensively the shortcomings of the models to predict ocean interior variables from primarily surface data, and the results could be better placed into context among similar studies that reconstruct ocean interior properties from observational data.

We would like to thank the reviewer for her/his valuable comments. Our response is given in blue, and the number of lines corresponds to those of the new manuscript with track changes.

General suggestions

One aspect that I think is missing from the manuscript is the contextualization of the authors' results with similar methodologies that have been applied to map salinity and temperature from observations (a few of which are referenced in the introduction). Although not all studies that reconstruct ocean interior properties from observations include a reanalysis-based evaluation of mapping accuracy (as is the focus of this manuscript), many report error statistics of their reconstructions evaluated against independent data. Su et al. (2018), for example, evaluate their reconstructed subsurface temperature anomalies using root mean squared error and R^2 as metrics, and the results of the OSSE reported here could be evaluated against those results.

We have included some contextualization in the discussion section as suggested by the reviewer. (lines 403-412)

In general, I was surprised to see such high disagreement with the test data at depth, when temperature and salinity should be more constant in space and time, and therefore relatively easier to reconstruct than at the surface. Buongiorno Nardelli (2020), for example, retrieves minimum errors for temperature and salinity at depth. This, in my opinion, points to an aspect of the methodology that can be significantly improved. It is not particularly surprising that a model based primarily on surface characteristics would struggle to estimate temperature and salinity at 1000 meters. I suspect a strategy of somehow de-emphasizing the impact of the surface predictor datasets as depth increases might improve these high offsets at depth. In any event, this is another instance where contextualization of the results of this OSSE would be helpful.

We understand the concern of the reviewer. The main question of this work, as he /she pointed out previously, is to assess how in situ and satellite observing systems can be leveraged to reconstruct ocean interior properties. However, the in situ profiles are only used for training our models, and then the reconstruction of subsurface fields is done using only surface fields. This is why our metrics degrade with depth. We have tried to clarify this point in the new version of Figure 3 and make the text clearer (lines 170-174). In addition, we contextualized the results in the discussion section (lines 403-411) and pointed out other possible strategies to improve this point in further work.

Lastly, the authors miss an opportunity to incorporate uncertainties into their experiment, or at least to discuss their implications. OSSEs present an opportunity to mimic real-world conditions; in reality, satellite observations are not perfect, nor are temperature and salinity measurements from profiling floats. Incorporating measurement uncertainty estimates in the analysis would be an important piece for answering the central question of how feasible it is to use satellite and in situ data to reconstruct ocean interior properties.

We would like to thank the reviewer, especially for this point. We have re-done the study, taking into account the typical uncertainties of each observable variable used in the study (taken from some of the products detailed in tables 1 to 3), and repeated the validation of the models. We provide a comparison of the general metrics with and without uncertainties in Table 5. We could conclude that even if the metrics are slightly degraded when uncertainties are included, as expected, the main conclusions of our work are still valid. We have modified the text accordingly.

Line-by-line comments

Abstract: I would suggest defining the simulated in situ measurement platforms as “Argo floats” or “profiling floats” rather than buoys in the abstract.

Following the reviewer’s comment, we have changed both appearances of “*buoys*” to “*Argo floats*”.

28: Presumably, this should say “subsurface temperature and salinity”?

The reviewer is right. We changed the text from *“subsurface temperature and subsurface temperature anomalies”* to *“subsurface temperature and salinity anomalies”*.

86: Awkward phrasing in reference to the equatorial region.

We changed the wording to “*equatorial region*”.

97: punctuation issue here.

We updated accordingly.

161-166: I’m not sure I understand the training and test split. Are you withholding some percentage of the dataset on a daily frequency (if so, what percentage?) for testing during model training? How does this differ from the ground truth dataset that is being used for evaluation?

Our input data are only the vertical profiles of the reanalysis model (with their associated surface information) for the points where there was an Argo profile that day for the 2010-2022 period. This data is stored as daily files, which are then divided into an 80/20 split. The rest of the points of the reanalysis, where there were no Argo profiles registered, can be used for validating the model, as they will not be seen in the training of the models. However, we further independentized it in the validation section and used the 2008-2009 period, but we could have made it using any period of the 2010-2022 as most of the points are not seen by Argo and thus, not included in the training of the models.

We have changed Fig. 3 to clarify how the datasets are constructed and how the train/test split was divided. The text explaining the separation into train and test splits has also been updated and now contains the following:

“Finally, we separate our datasets into a train/test split, which will be common for all the trained models. This separation is made using an 80/20 ratio, where 80\% of data will be used for training and 20\% for validation as usual in machine learning models. We generate one dataset (or datafile) for each day. As the objective of our study is to analyze the feasibility of the reconstruction using current sampling of the ocean (and not predicting future trends and events), the separation is done by randomly separating the dates, but ensuring that each month is represented equally in both datasets. This avoids adding imbalances due to seasonal cycles that must be accounted for.”

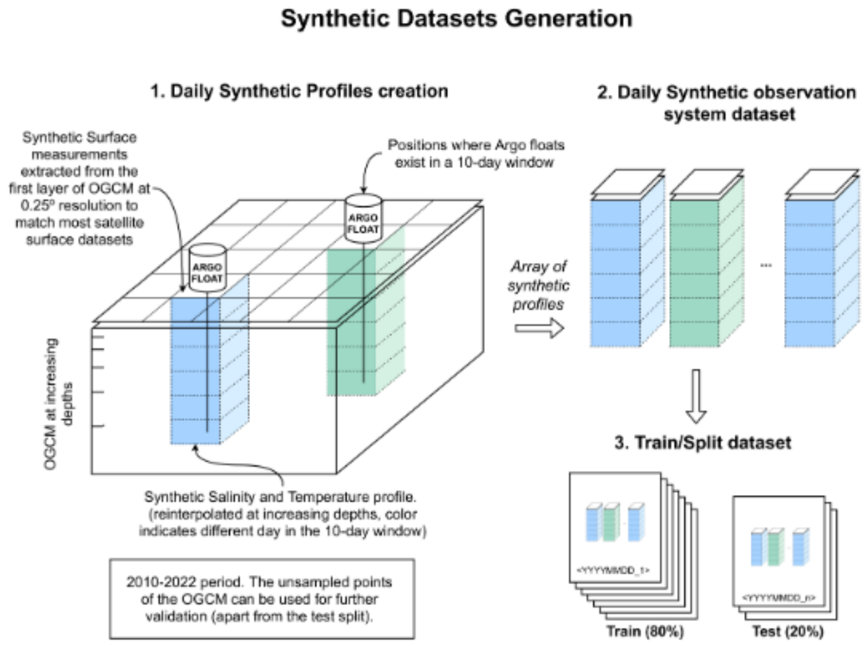
Synthetic Datasets Generation
1. Daily Synthetic Profiles creation
2. Daily Synthetic observation system dataset
Synthetic Surface measurements extracted from the first layer of OGCM at 0.25º resolution to match most satellite surface datasets
Positions where Argo floats exist in a 10-day window
ARGO FLOAT
ARGO FLOAT
Array of synthetic profiles
...
OGCM at increasing depths
Synthetic Salinity and Temperature profile. (reinterpolated at increasing depths, color indicates different day in the 10-day window)
2010-2022 period. The unsampled points of the OGCM can be used for further validation (apart from the test split).
3. Train/Split dataset
<YYYYMMDD_1>
<YYYYMMDD_n>
Train (80%)
Test (20%)

Figure 3. Observing system datasets generation from ~~Argo floats and~~ Copernicus Marine Service reanalysis data using Argo floats and surface satellite measurements sampling. For each day, the 10-day windowed simulated profiles are collocated with the central date data, generating a daily array of synthetic profiles. The different colors indicate different days in the 10-day window of a specific central date. ~~Boxes in blue indicate the input datasets. Boxes in white are datasets and auxiliary files derived from the input data. Grey boxes indicate a data cleaning or data processing step. Variables with an asterisk indicate surface variables, whereas the ones without an asterisk have data through all the vertical profiles.~~

173: It would be helpful to specify the metric you are referring to when discussing “accuracies”.

We changed “*Table 1 shows each model’s accuracy and error metrics*” to “*Table 1 shows each model’s accuracy (R2) and error metrics (MSE, MAE)*” to make it clearer when reading.

239: What is meant by “it does not overlap with the training dataset”? There are no Argo profiles from 2008-2009 in this region?

We used the 2010 to 2022 Argo floats for the training of our models because, from 2010 onwards, the number of profiles increased significantly. The specific filtering criterion and time overage for the training dataset are specified in Section 2.

“... We use all available profiles from 2010 to 2022, but only consider those that reach a minimum depth of 1000 meters and have good quality measurements according to their quality control standard.”

272: should be “…each of them with their own…”

We updated the phrase accordingly.