https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

On the reconstruction of ocean interior variables: a feasibility data-driven study with simulated surface and water column observations

Aina García-Espriu[1], Cristina González-Haro[1], and Fernando Aguilar-Gómez[2]
[1]Institute of Marine Sciences (ICM), CSIC
[2]Instituto de Física de Cantabria (IFCA), CSIC
Correspondence: Aina García-Espriu (ainagarcia@icm.csic.es)

Abstract. This work uses data-driven approaches to study the feasibility of reconstructing ocean interior variables (temperature and salinity) from surface observations provided by satellites and interior observations provided by buoys. The feasibility of the approach is based on an Observing System Simulation Experiment (OSSE) in which we use the outputs from an ocean numerical model as the ground truth, and simulate a real observing system of the ocean, taking the surface of the model as a simulation of satellite observations, and vertical profiles in the same locations as the real buoys. We implemented different models based on Random Forest Regressors and Long-Short Term Memory networks which were trained with the simulated observations and validated against the complete numerical model results. We obtain high spatial and temporal correlation using both technologies and an accurate description of the annual variability of the data accompanied by small biases.

1 Introduction

The ocean serves as Earth's primary climate regulator, functioning as a massive heat sink and carbon dioxide reservoir and distributing its thermal energy globally through currents (Webster, 1994). Ocean monitoring relies on two complementary approaches: satellite-based remote sensing, which provides extensive surface-level data, and in-situ observations through programs like the Argo float network (Argo, 2009), which offers deep-water measurements up to 2000 m. Satellite missions such as SMOS (Kerr et al., 2010), Aquarius (Lagerloef et al., 2008), and SMAP (Entekhabi et al., 2010) provide global sea surface salinity (SSS) measurements, while AVHRR (Casey et al., 2010), MODIS (Kilpatrick et al., 2015), and Sentinel-3 (Donlon et al., 2012) deliver sea surface temperature (SST) observations. These measurements are processed into various operational products including filled-gap SST products such as OSTIA (Good et al., 2020), MUR (Chin et al., 2017), and ESA CCI (Merchant et al., 2019), as well as SSS products including BEC (Olmedo et al., 2021), SMAP Level 3 (Meissner et al., 2018), and ESA CCI (Boutin et al., 2021). On the other hand, the Argo program provides 300-400 daily vertical profiles worldwide. However, the current monitoring infrastructure cannot completely reconstruct the ocean layers. This challenge is faced in the Digital Twin of the Ocean (DTO) concept, where the application of Artificial Intelligence (AI) models to synthesize surface and deep-water measurements can achieve a 4D (3D ocean reconstruction + time variability) reconstruction of ocean dynamics.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

The integration of different data sources along with advanced modeling techniques presents a promising way to improve our understanding and prediction of oceanic changes in the context of global climate dynamics.

The integration of AI and data-driven methodologies has transformed physical oceanography, which theoretical models and limited observational data traditionally dominated. Recent research has demonstrated significant progress in reconstructing 2D and 3D ocean variables. Subsurface salinity has been reconstructed at 0.25ºx0.25º resolution in Tian et al. (2022) by enhancing coarser resolution products (1ºx1º at a monthly scale). Subsurface temperature and subsurface temperature anomalies were reconstructed in Su et al. (2015, 2018, 2021) using monthly gridded Argo data and AI models such as Support Vector Machines and Long-Short Term Memory (LSTM) Neural Networks. They used monthly gridded Argo data at 1ºx1º resolution to perform their reconstruction. Buongiorno Nardelli (2020) proposed a deep learning network to reconstruct the ocean water column using combined satellite and in-situ measurements. Their innovative approach consists of using the potential of the LSTM not to predict a time series, but to predict a depth series. As in the previous studies, a gridded in-situ dataset with monthly temporal resolution was used.

It has been demonstrated by the state of the art that different AI models can derive oceanic data from in-situ at a large temporal scale and using gridded (and interpolated) in-situ (*i.e*: 1-month Argo gridded products). Here, we want to assess if it is feasible to reconstruct the ocean interior variables at the native resolution of the surface products while maintaining a daily temporal resolution, which is suitable for studying temporal mesoscale ocean dynamics, combined with non-gridded in-situ data to avoid interpolations and to respect the spatial variability of the measurements. The paper is structured as follows: in Sect. 2 we introduce the datasets used for the feasibility study. Then, in Sect. 3 we introduce the methodology, including an analysis of the limitations of the dataset, the selection of variables and model architectures, the generation of the simulated observation system, and the implementation and training of the models. The results are shown in Sect. 4 where we analyze the explainability, accuracy, and errors of the models' predictions with different tools and perspectives. In Sect. 5 we discuss the obtained results and in Sect.6 we summarize the conclusions and propose different paths to explore in this field.

2 Datasets

The data-driven approach presented in this study relies on two complementary datasets. The first one is the in-situ measurements from the international Argo float network, which provides vertical profiles of the physical properties of the ocean (temperature and salinity). The second dataset is the CMEMS Global Ocean Ensemble Reanalysis product, produced with a numerical ocean model constrained with data assimilation of satellite and in situ observations. It provides a complete picture of ocean state variables. The specific products that are used in the study are:

- **Argo floats**. We use Argo profilers (Argo, 2025) to determine the positions that our simulated observing system has to sample. Argo data are collected and made freely available by the International Argo Program and the national programs that contribute to it (http://www.argo.ucsd.edu). The Argo Program is part of the Global Ocean Observing System. We use all available buoys from 2010 to 2022 but only consider ones that reach a minimum depth of 1000 meters and have good quality measurements according to their quality control standard.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

– **CMEMS Global Ocean Ensemble Reanalysis**. This study has been conducted using E.U. Copernicus Marine Service Information (https://doi.org/10.48670/moi-00024). The CMEMS Global Ocean Ensemble Physics Reanalysis product (Mercator Océan International, 2025) is given at 0.25ºx0.25º resolution and contains daily temperature, salinity, currents, and ice variables for 75 vertical levels. We use this reanalysis to simulate both the in-situ observations and the satellite surface data.

3 Methodology

This work uses a reanalysis of daily 3D gridded data to simulate the current ocean observation system and assess the feasibility of a 4D reconstruction of the ocean variables. Using a reanalysis model instead of in-situ data enables us to access the locations that are not sampled by the in-situ measurements. The sampled locations are used in the training datasets of our models (both in train and test split). In contrast, the unsampled locations validate how the model extrapolates to regions not seen by the profilers.

Argo floats cover a small percentage of the ocean, offering 400 profiles per day on average. We can increase the number of observations by a factor of 10 if we use a 10-day time window, which corresponds to the profiler cycle length. This work aims to study the feasibility of a daily reconstruction and spatial resolution provided by the microwave sensors on board of satellites (about 0.25º). In Fig. 1 we show the number of Argo floats for a 10-day window at 1 and 5-degree resolution and the intra-pixel standard deviation as seen by the reanalysis model. To obtain almost complete coverage of the globe, we would need to resort to a resolution of 5ºx5º as seen in the top-right panel in Fig. 1. The intra-pixel variability of each 5ºx5º box is shown in the bottom panels in Fig. 1 and can reach values of more than 1.5 ºC in temperature and 0.5 (g/kg) in salinity. This variability would worsen the reconstructions in areas where the intra-pixel variability is very high (which are the regions of more activity and thus, zones of interest). So, to be able to work with daily data, we opt to use a sparse data approach instead of gridded datasets. With this methodology, the input of the models is a structure that contains the surface variables along with the acquisition condition identifier (*i.e.* latitude, longitude, time of the year) as predictors and the salinity and temperature vertical profiles as predicted variables.

3.1 Variables selection

Satellite observations offer measurements of multiple variables that can be interconnected and describe different processes of interest. However, not only do the measured values offer important information, but also their acquisition conditions such as the acquisition time or the geolocation. As this information also helps in the description and modelization of the oceanic processes, considering them in our models can help in the understanding of the relationships between different water masses or seasonal patterns.

In this work, we consider as observation coordinates the latitude, longitude, depth, and time of the measurements. Surface temperature shows a latitudinal variation pattern, being the equator region the warmest. It also presents daily and seasonal cycles, both of which present latitudinal variations. The longitude coordinate does not directly impact the measured variables

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

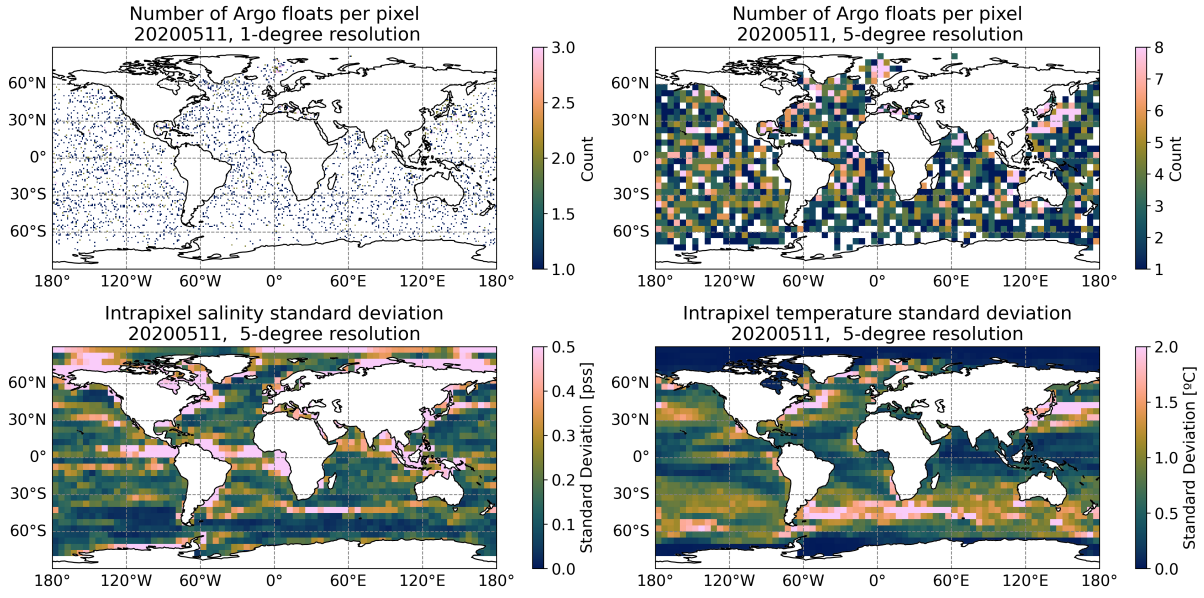

Figure 1. Argo floats coverage at 1-degree (top-left) and 5-degree resolution (top-right). Intra-pixel standard deviation of the salinity (bottom-left) and temperature (bottom-right) measurements as seen by the reanalysis model at 5-degree resolution.

but, along with the latitude, can help the model understand the earth's topology and determine the position of different water masses. The depth of the observations also affects its measured values, as in the deeper layers the ocean processes occur at a much larger time scale. In contrast, in the upper layers, the time scales of the oceanographical processes rapidly increase and the variation is greater. Finally, the observation time affects the measurements at different scales as both, salinity and temperature, present a seasonal cycle, so knowing at which time of the year the measurement was taken can give information on those patterns.

The surface measurements include SSS, SST, sea surface height (SSH), mixed-layer depth (MLD), and information about the currents which are represented as the vertical and horizontal components of the velocity vectors (UO and VO respectively). The combination of SSS and SST defines the sea surface density. These relations contribute to the stratification and vertical mixing of the ocean—, which in turn affect MLD stability. Increased heating at the surface can make the MLD deeper, while increased freshwater input can make it shallower.

Thus, the variables that are used in the training of the models are a combination of the aforementioned ones. However, one key aspect when selecting the model predictors is knowing the limitations of the data. The number of input variables can positively and negatively affect the model's outputs. On the one hand, if too few variables are used, we may not be able to describe the oceanographical processes of interest. On the other hand, if too many variables are used, the model will not have enough data to describe all the possible combinations and its quality will rapidly degrade. Thus, the specific selection of the variables used on each model will be further discussed in Subsect. 3.4

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

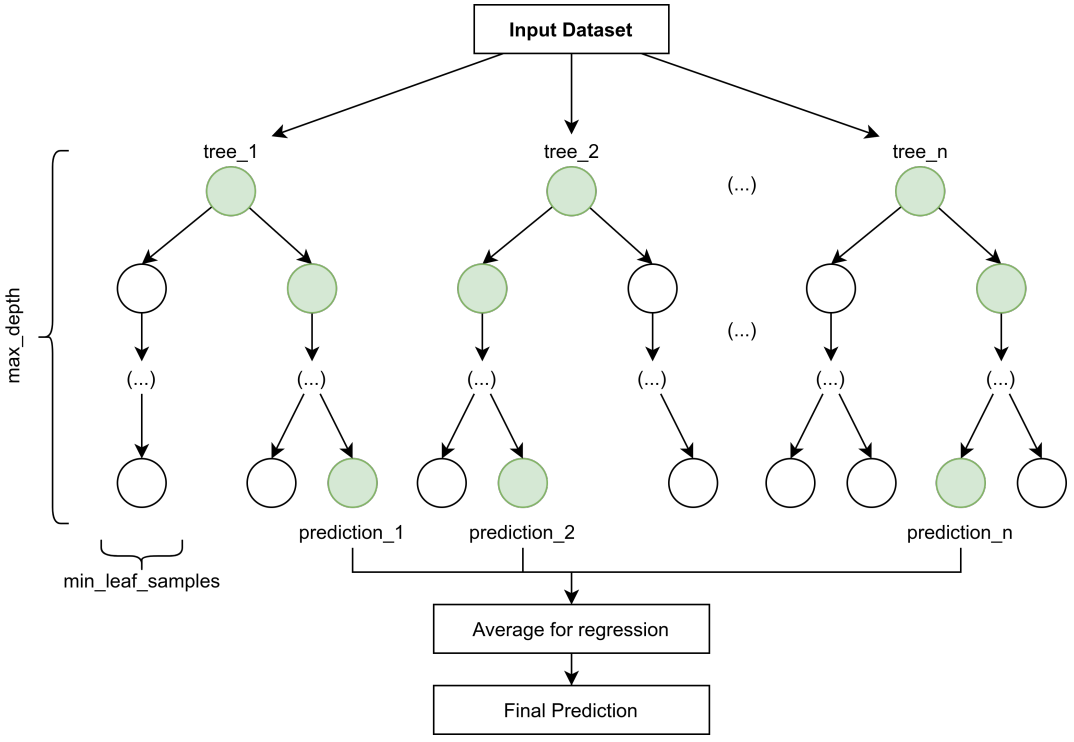

Figure 2. Random Forest Regressor structure. The final prediction is the average of all individual tree predictions, reducing overfitting and variance compared to using a single decision tree.

3.2 Models selection

Two model architectures were chosen for this study: the Random Forest Regressor (RFR) (Breiman, 2001) due to its algorithmic simplicity and training cost and the Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) due to its already demonstrated utility in oceanographic applications.

On one hand, the RFR can model non-linear relations between the variables, which is a key aspect when studying oceanographic processes. Furthermore, it does not need large datasets to produce good predictions. It works by constructing multiple decision trees during training and outputting the mean prediction of all trees for a more robust result, as illustrated in Fig. 2. Each tree is built using a random subset of the training data and considers a random subset of features at each split point. The randomness helps prevent overfitting and the averaging of the predictions gives stability and accuracy in the results.

On the other hand, LSTM architectures have proved to produce promising results in the field (Buongiorno Nardelli, 2020; Su et al., 2021). This type of architecture can handle long-term dependencies while maintaining stable gradients and mitigating the vanishing gradient problem. LSTMs can remember important information and forget irrelevant details through three main gates: the forget gate decides what information to discard, the input gate determines what new information to store, and the output gate controls what parts of the cell state should be output. It is well-suited for tasks that generate data sequences as in

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

this case, where a vertical profile is generated for each surface data point. By comparing these two models, we can not only determine which one has better performance but also discuss if we have enough data to use deep learning mechanisms and if the improvement regarding simpler models is worth the cost.

3.3 Simulated profiles construction

To simulate the current observation system we need to determine the positions in which an in-situ observation was taken and retrieve the equivalent data from the reanalysis, as described in Fig. 3. First, we determine in which positions we had in-situ measurements and need to be part of the simulated dataset (*Generate sampling* in Fig. 3). We use the positions of the Argo floats that offer a homogeneous sampling of good-quality data. We only use profiles that measure at least up to 1000 m depth to guarantee homogeneity in the input dataset and to assure that the profiles have measurements in the complete interest range of the water column. Furthermore, it removes the points close to the coast that typically present different dynamics than the open ocean. To guarantee the quality of the observations, we only use profiles that offer the variables given in adjusted mode that have the best possible quality control (see "1: Good" in (Argo Data Management Team, 2024), pg. 105). Curated and calibrated profiles give a more faithful representation of the measurements. Finally, we only use profiles contained in the 60 ºS - 60 ºN latitude range as the polar region's dynamics greatly differ from the open ocean ones. The positions of the profiles that fulfill the aforementioned checks are then collocated to the reanalysis grid to generate a daily mask of "seen" observations.

Then, we generate the simulated profiles by combining the daily reanalysis datasets and the daily sampling mask (*Generate simulated profiles* in Fig. 3). For each profile, we select the temperature and salinity for all depths up to 1000 m and then, add the surface variables of the corresponding geographical position (first layer of the reanalysis).

3.4 Implementation and Training

We implemented different models using RFR and LSTM architectures, varying the input variables and the configurable parameters such as the number of trees, the number of layers, the number of units in each layer, etc. However, in this work, we only present the two RFR and two LSTM configurations that are interesting for the discussion. In the case of the RFR, we trained one model (RFRv1) with only the SSS and the SST as surface variables to see to which extent the interconnection of these variables could predict their vertical profiles. Then, we trained another model (RFRv2) with the complete set of surface variables to compare it with the predictions made by the LSTM models.

In the case of the LSTM, we trained different models varying their architecture configuration and hyperparameters shown in Fig. 4 such as the activation function, the number of layers, the number of units, the learning rate, etc. For the discussion, we selected the same configuration used in Buongiorno Nardelli (2020) to check whether it could be extrapolated to our problem (LSTMv1). We also selected a coupled model (LSTMv2) which contains the best-performing salinity and temperature predictors. The exact set of predictors and tuning parameters of each model are the ones as follows:

- **RFRv1:** Salinity and Temperature as surface variables. Maximum of 100 decision trees, a minimum of 10 measurements per leaf, and a maximum tree depth of 20.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

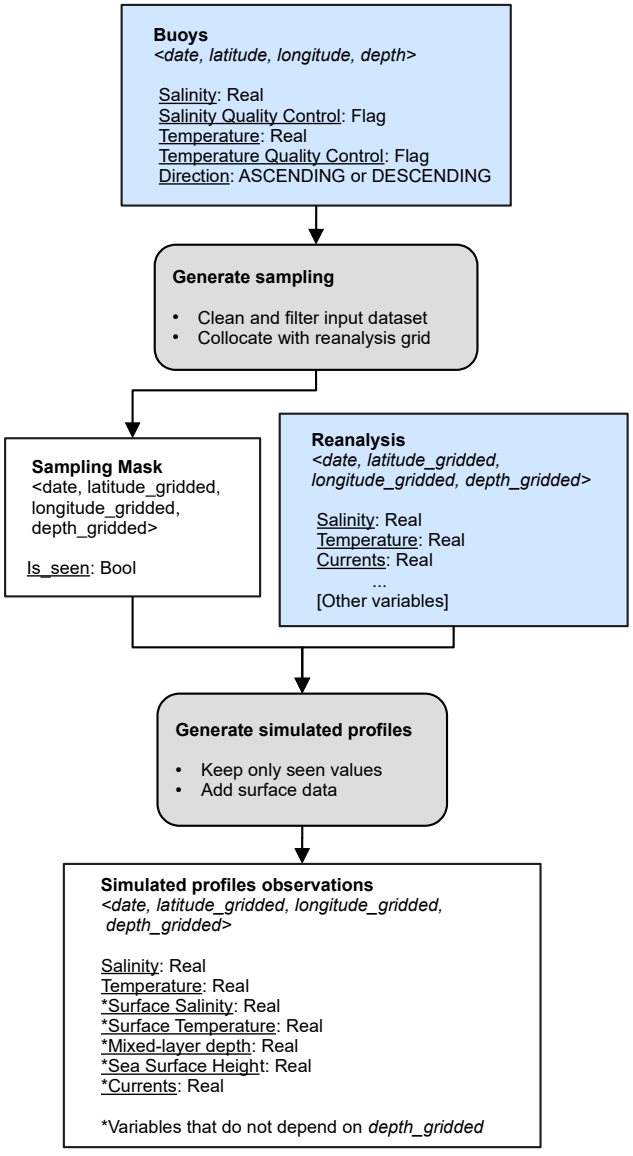

Figure 3. Observing system generation from Argo floats and CMEMS reanalysis data. Boxes in blue indicate the input datasets. Boxes in white are datasets and auxiliary files derived from the input data. Grey boxes indicate a data cleaning or data processing step. Variables with an asterisk indicate surface variables, whereas the ones without an asterisk have data through all the vertical profiles.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

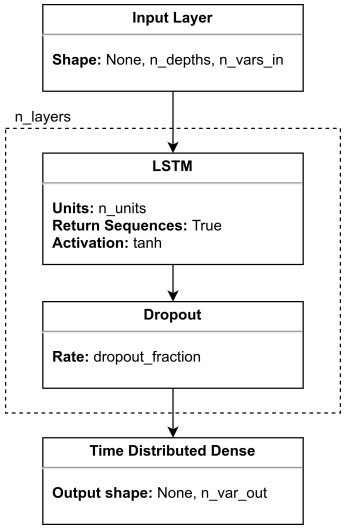

Figure 4. Model architecture using LSTM layers. The model takes as input a list of points to predict, where each point is a matrix of *n_depths* x *n_vars_in* positions. Then, it connects with *n_layers* groups of LSTM + Dropout layers. Finally, a Time distributed Dense layer that produces the final predictions.

- **RFRv2:** Salinity, Temperature, Currents, MLD, SSH, and latitude as surface variables. Maximum of 100 decision trees, a minimum of 10 measurements per leaf, and a maximum tree depth of 20.
- **LSTMv1:** Salinity, Temperature, Currents, MLD, SSH, day of the year, depth, longitude, and latitude as surface variables. Two LSTM layers with 35 units, a dropout value of 0.2, a learning rate of 1e-5, early stopping, and the hyperbolic tangent (tanh) as the activation function. Trained with at least 100 epochs using Early Stopping functions to avoid overfitting.
- **LSTMv2:** Salinity, Temperature, Currents, MLD, SSH, day of the year, depth, longitude, and latitude as surface variables. For salinity prediction: three LSTM layers with 1024 units, a dropout value of 0.2, a learning rate of 1e-5, and the Softsign as the activation function. For temperature prediction: two LSTM layers with 512 units, dropout 0.2, a learning rate of 1e-5, early stopping, and the hyperbolic tangent (tanh) as the activation function.

The models in this work are designed to conduct a reconstruction task. We reflect this in the separation of the training and test datasets. If the model had to predict future events we would need to divide some consecutive years for training and the rest for testing. In this case, we aim to assess if the model can reconstruct what it has already seen on the surface, so the datasets are randomly divided on a daily frequency. The datasets are balanced monthly and yearly to avoid possible biases due to the division imbalance. Furthermore, we use the same train/test splits for all the trained models to ensure that the ingested data is the same.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

Model	R^2 (SAL//TEMP)	MSE (SAL//TEMP)	MAE (SAL//TEMP)
RFRv1	0.88 / 0.76	0.08 / 5.70	0.17 / 1.49
RFRv2	0.95 / **0.85**	**0.04** / **3.16**	**0.11** / **0.86**
LSTMv1	0.87 / 0.82	0.08 / 4.03	0.2 / 1.17
LSTMv2	**0.96** / 0.84	**0.04** / 3.76	0.13 / 1.01

Table 1. R^2, mean squared error and mean absolute error metrics for salinity and temperature for the different models. The best-performing model is highlighted in bold text for each metric.

Both models are implemented in Python (Van Rossum and Drake, 2009) using standard libraries such as Tensorflow (Abadi et al., 2015) and Scikit-learn (Pedregosa et al., 2011). The datasets are preprocessed from the original netCDF (Rew et al., 1997) files to Feather format, which is a column-oriented binary disk-based format based on Apache Arrow and supported by Python. This optimizes the data ingestion which can be one of the major bottlenecks in the training process.

4 Results

We ran the four models with the complete test dataset (which is common for all models). Table 1 shows each model's accuracy and error metrics, both in temperature and salinity predictions. We obtain accuracies that range between [0.75 - 0.96], making the RFRv2 and LSTMv2 better than their v1 counterparts. Furthermore, the salinity predictions are more accurate and less prone to errors than the temperature ones.

Figs. 5 and 6 show the same metrics as in Table 1 but as a function of the depth of the measurement. Salinity models display a robust prediction between depths, as the slope of the functions/metrics with respect to depth is almost vertical. There, we can observe that RFRv2 provides a higher R^2 and smaller errors than the rest of the models, closely followed by the LSTMv2 approach. In the case of temperature, we can see that metrics decline at about 200 m depth where the predictions start to fail and errors grow larger. In salinity, we can also observe the slope change to a lesser extent, but the scores remain stable in the lower depths. Reconstructing the temperature field proves more challenging and requires further work to achieve the quality of the salinity reconstruction. In Appendix A we provide the complete validation of the temperature, but due to its lower quality when compared to the salinity reconstruction, we decided to focus on the validation of the salinity from now on.

4.1 Model explainability

Machine learning-based models, as opposed to physical models, pose a problem when interpreting the results obtained, as they are sometimes treated as black boxes where we know the input information and the output produced but we do not understand the processes involved in obtaining the results. Understanding how these outputs are made helps us understand the causes and improve the models by focusing on the essentials. Although Random Forest-based models are more transparent and interpretable than those based on LSTM networks, a method of interpretability comparable to both has been chosen to improve the analysis.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

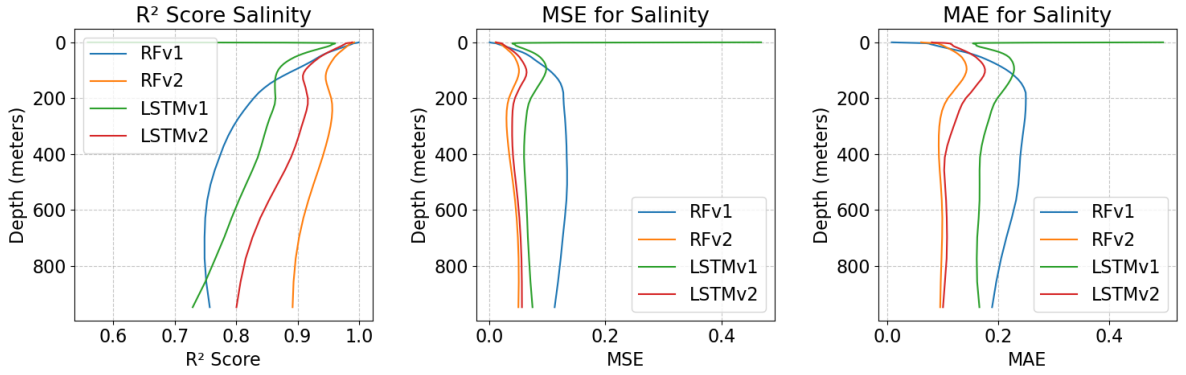

Figure 5. Predicted salinity R^2, mean squared error, and mean absolute error (from left to right) per depth for the different models: RFRv1 (blue), RFRv2 (orange), LSTMv1 (green), LSTMv2 (red).

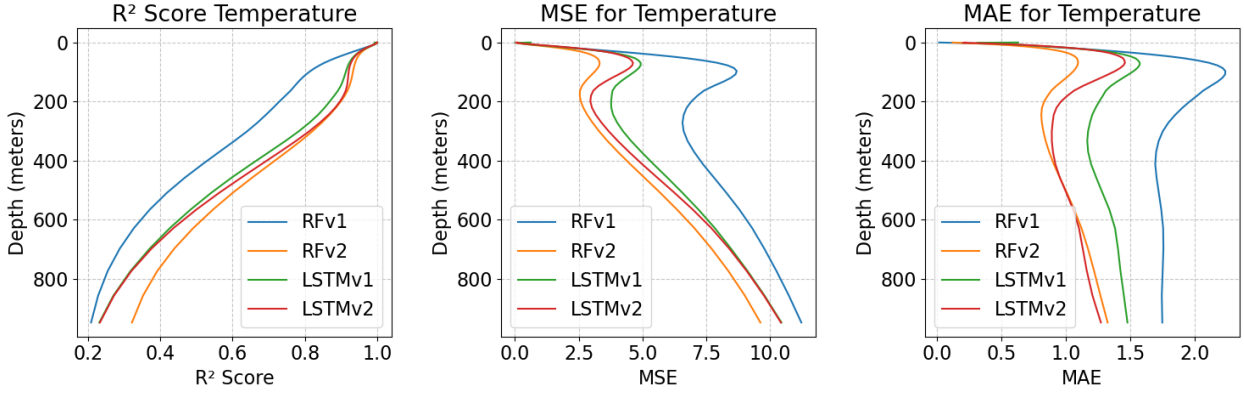

Figure 6. Predicted temperature R^2, mean squared error, and mean absolute error (from left to right) per depth for the different models: RFRv1 (blue), RFRv2 (orange), LSTMv1 (green), LSTMv2 (red).

SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) is a framework for interpreting machine learning model predictions based on cooperative game theory principles. It is based on the Shapley value concept from game theory to allocate contributions of individual features to the output of a model fairly and consistently. They quantify how much each feature contributes to moving the model prediction from a baseline (typically the mean prediction) to the actual output for a specific data point. We have computed the SHAP values for each depth of the predictions to see how the contribution of each variable changes in the vertical profile. Figs. 7 and 8 show the percentage of each variable contribution for the RFR and LSTM models respectively.

For the RFR (Fig. 7), we observe that the SSS is the dominant predictor which provides a climatological reference. Then, variables like SST, latitude, and SSH contribute by capturing temporal and regional variations in the signal. The LSTM model

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

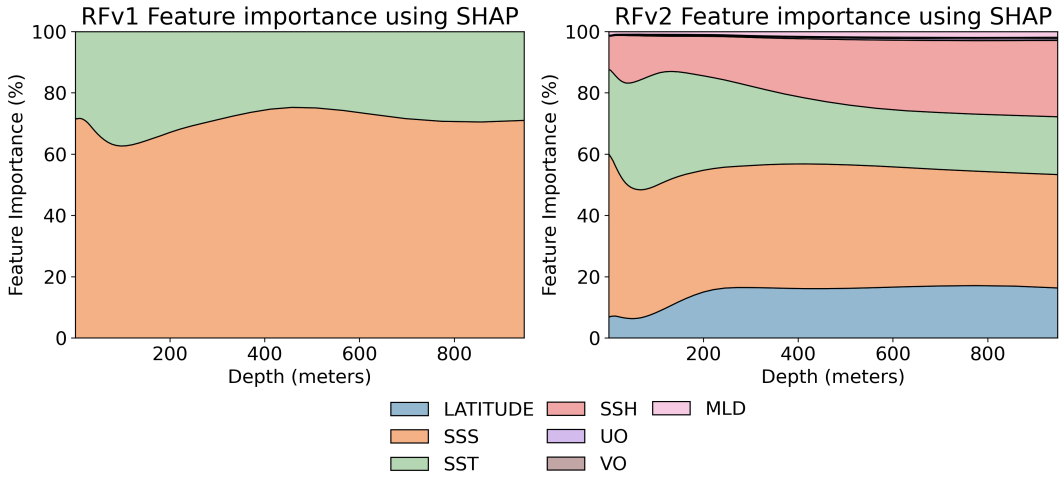

Figure 7. Feature importance percentage using SHAP values for RFRv1 (left) and RFRv2 (right) salinity models.

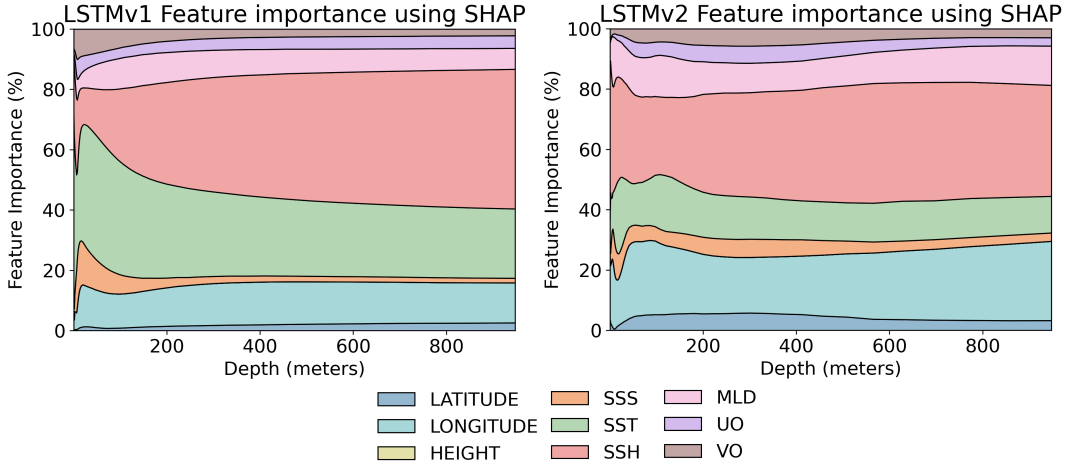

Figure 8. Feature importance percentage using SHAP values for LSTMv1 (left) and LSTMv2 (right) salinity models.

(Fig. 8), however, shows a different learning strategy as it does not rely on SSS. Instead, it derives it from other variables such as longitude and temperature. The latitude variable is also mostly accounted for in the inherent temperature's latitudinal structure. This approach captures both the climatological baseline and dynamic components without depending directly on SSS measurements. This analysis shows the fundamental differences between traditional machine learning and deep learning approaches in ocean reconstruction applications. While RFR excels at direct variable relationships, LSTM infers complex spatiotemporal patterns from indirect indicators. This capability becomes particularly valuable in regions with sparse SSS measurements or when reconstructing historical salinity patterns where direct observations may be limited.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

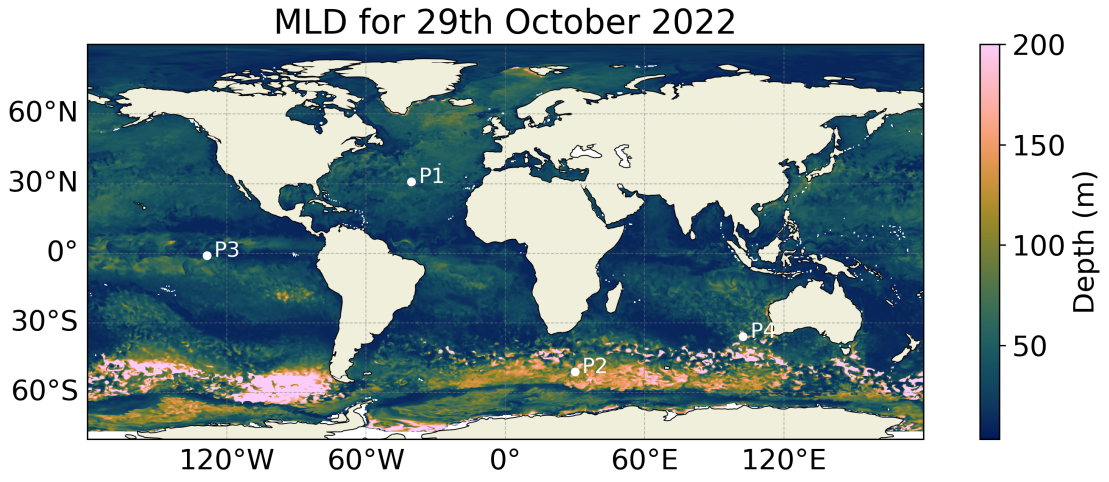

Figure 9. Mixed-Layer Depth as seen by the reanalysis on the 29th October 2022. Points used in the validation shown in Figs. 10 and A3 are referenced here in white.

4.2 Validation with the test split dataset

To gain insights into the vertical reconstruction, we show in Fig. 10 the reconstructed vertical profiles of the points shown in Fig. 9 for the 29th October 2022 compared to the ground truth. We can observe a better fit in the case of the RFRv2 and LSTMv2 as they present a good correspondence with the ground truth curve. The models present some difficulties when predicting sharp transitions, as seen in the third and fourth points, as they tend to smooth the vertical profile. Sharp transitions, however, can be due to the quality of the numerical model or a well-mixed layer that produces abrupt changes. Further studies should be conducted in these difficult regions when using real in-situ data and this metric should be revisited.

We have used the test split dataset aggregated as daily 5-degree maps (to have complete coverage) to validate the reconstruction's spatial coherence. We aggregated the predicted profiles in each grid cell and subtracted the aggregation of the true values in the same cell. In Fig. 11 we can see an example for the 29th of October 2022. We can observe low biases which are about ± 0.1 pss in the major part of the map for all models except LSTMv1, which displays a latitudinal pattern. In broad strokes, the bias patterns are similar among the different model configurations and are in the range of observed biases in satellite products of the same resolution for this region as reported in Figure 6 of Olmedo et al. (2021),

Through this first validation, we saw that the RFRv1 misses key surface variables that can potentially improve the quality of the reconstruction. Variables such as SSH and latitude have a high contribution to the predictions as seen in the SHAP values of the RFRv2 (Fig. 7, right panel). We also saw that the architecture presented in Buongiorno Nardelli (2020) requires some fine tuning to optimize its performance to the given challenge. In subsection 4.3 we conduct a more in-depth validation of the best-performing models (RFRv2 and LSTMv2).

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

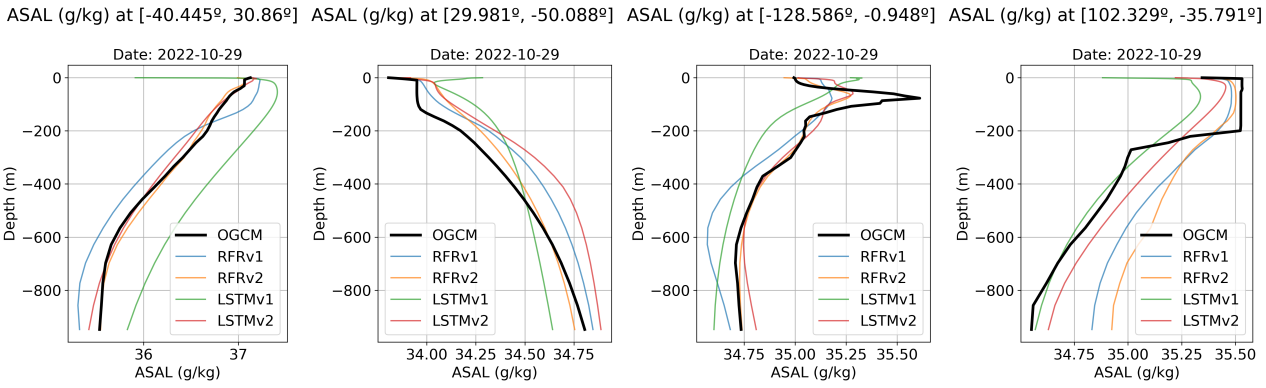

Figure 10. Predicted salinity vertical profiles for four different points (see Fig. 9 for its location) and the different models: RFRv1 (blue), RFRv2 (orange), LSTMv1 (green), LSTMv2 (red). The date is 29th October 2022.

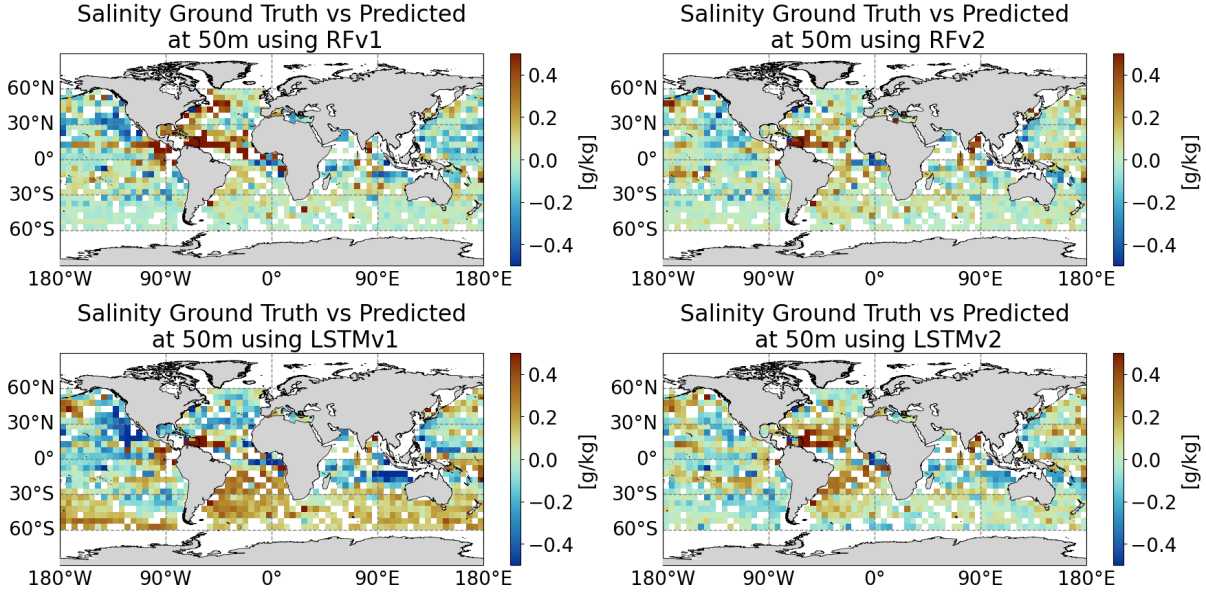

Figure 11. Predicted salinity minus Ground Truth (50 m depth) at a 5-degree grid (from top to bottom and left to right): RFv1, RFv2, LSTMv1, LSTMv2. The date is 29th October 2022.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

4.3 Validation with the reanalysis dataset

We chose the region with longitudes ranging between 80 ºW and 40 ºW and latitudes ranging between 25 ºN and 44 ºN as it is a highly dynamic area that comprises the Gulf Stream current. Notice how the region's sampling is non-homogeneous and focuses on more dynamical regions, *i.e.* where the Gulf Stream is located. The validation is conducted with the predictions of both RFRv2 and LSTMv2 for the 2008-2009 time period, as it does not overlap with the training dataset. Three different levels of depths are considered to give an insight into the column variability: 5, 50, and 500-meter depths.

First, in Fig. 13, we assess the temporal biases of the 2-year averaged regional maps by computing the difference between predicted values and the reanalysis (ground truth) for the selected depths during both years. We can observe that the biases are in the 0.1-0.2 pss range at most. The high biases observed in the region close to the coast in Fig. 13 are because the model predicted values outside of the training region (profilers did not reach the 1000 m specified in the data preparation as seen in Fig. 12).

Then, in Fig. 14 we show the temporal variability seen by the numerical model (our ground truth) and we use it as a reference to assess the differences with our predictions. We performed a significance test using a Fisher F-test for the difference of the variances at a 95% confidence level. Areas with significant differences are indicated as contours. The models are losing a variability range of about 0.1 g/kg and display spatial patterns similar to the ones observed in Fig. 13. LSTMv2 model has more regions where the variability differences are not significant, mostly in the first layer but overall, the differences are statistically too large. However, the differences do not grow larger in highly dynamic areas, implying that the models can capture the dynamical fluctuations of the salinity through the time series (although with less intensity), and thus, it separates from the climatological value of the predicted variable.

The temporal mean squared error (MSE) helps to understand which water masses the model has more difficulty representing. A zone with a high MSE could indicate that the model has not learned how to describe its dynamics, which can be due to under-sampling of that region or highly complex dynamics that the model was not able to learn. Figure 15 shows the temporal MSE for both models which have small values up to 0.1 (g/kg). This low value means that both models have a good characterization of the dynamics. In the case of the RFRv2, the maps are more stable in terms of MSE both horizontally and vertically, but the values are low on both models. As also observed in Fig. 13, the high MSE values observed in the region close to the coast are because we are predicting values outside of the training region, and should be discarded.

In Fig. 16, we compute the temporal correlation coefficient, computed as the Pearson correlation between the proposed model and ground truth time series for each gridpoint. This metric provides information about the capability of the respective models to properly describe the temporal cycles of the variables. As shown before in Fig. 14 the models can capture the same spatial pattern of temporal variability seen by the reanalysis (ground truth). However, we further analyze the temporal correlation to quantify to which extent both, the reconstructed and the ground truth show the same temporal variability evolution. As can be seen, both models properly describe the temporal variability of the reconstructed salinity for the shallower layers (5 m and 50 m). For the deeper layer considered here (500 m) the reconstruction is degraded in the southern part of the region, as we separate from the more dynamic region of the Gulf Stream Current. We further analyzed the impact of MLD on the

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

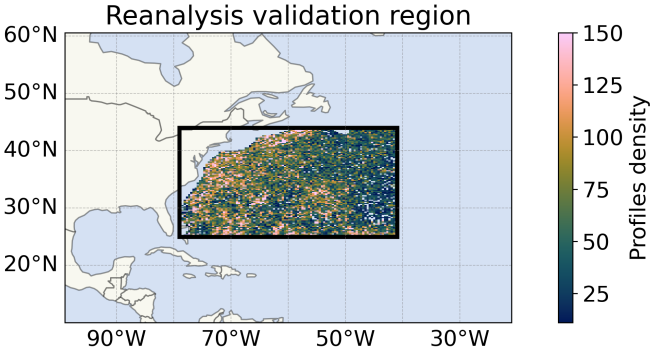

Figure 12. Number of Argo profiles at 1/100 region-size resolution from 2010 to 2022 in the region where the models have been compared to the reanalysis product (black rectangle).

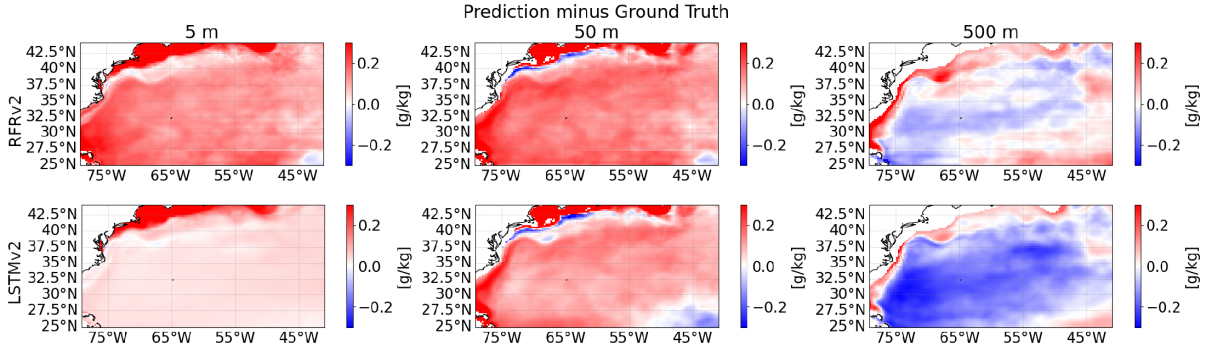

Figure 13. Biases at different depths (from left to right: 5 m, 50 m, and 500 m) between predicted salinity and ground truth for the two proposed models RFRv2 (top row), LSTMv2 (bottom row).

reconstruction to check to which extent a deeper mixed layer provided better reconstruction, but no conclusive results were found (not shown).

It is also important to remark on the artifact that we can observe in Fig. 16 in the RFRv2 at a 5-meter depth, where horizontal lines appear due to the inclusion of the latitude coordinate in the model. Although RFRv2 produces great statistics (the variability is well captured, the MSE is low and the correlation is high), it also produces artifacts in the latitudinal direction due to the binarization of the decision trees. This should be further investigated if we want to use this model architecture in the future.

Finally, we compute the temporal series of the spatial correlation coefficient to understand if there is any temporal or seasonality in the performance of our models. We can observe in Fig. 17 that both models present a good correlation at all studied depths, with an R^2 score higher than 0.92. Furthermore, we do not observe any tendency in the data or seasonal biases, meaning that our models can capture both possible tendencies and seasonal variability of the reconstructed variable.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

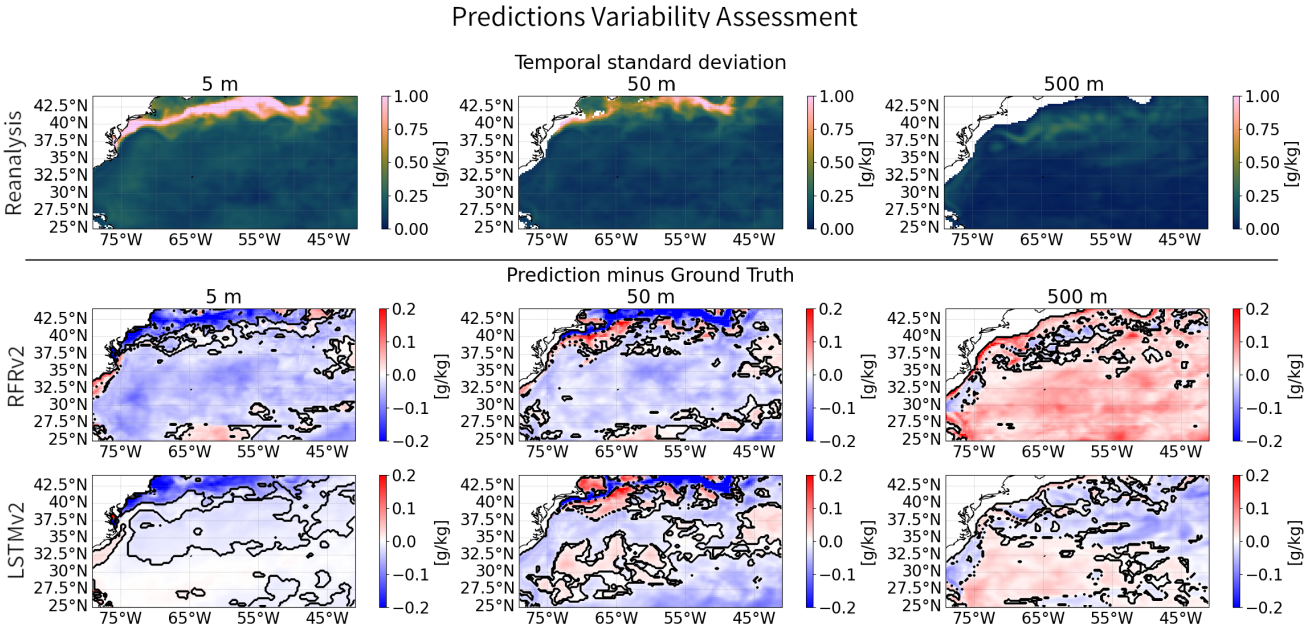

Figure 14. Temporal variability assessment between the predictions and the ground truth (reanalysis) salinities at different depths (from left to right: 5 m, 50 m, and 500 m). The standard deviation of the reanalysis (taken as our ground truth) for the 2008-2009 period (first row). Differences between the standard deviation of the ground truth and each model RFRv2 (second row), and LSTMv2 (third row). Solid lines delimit statistically significant areas at a 95% confidence level using an F-test for the difference of variances.

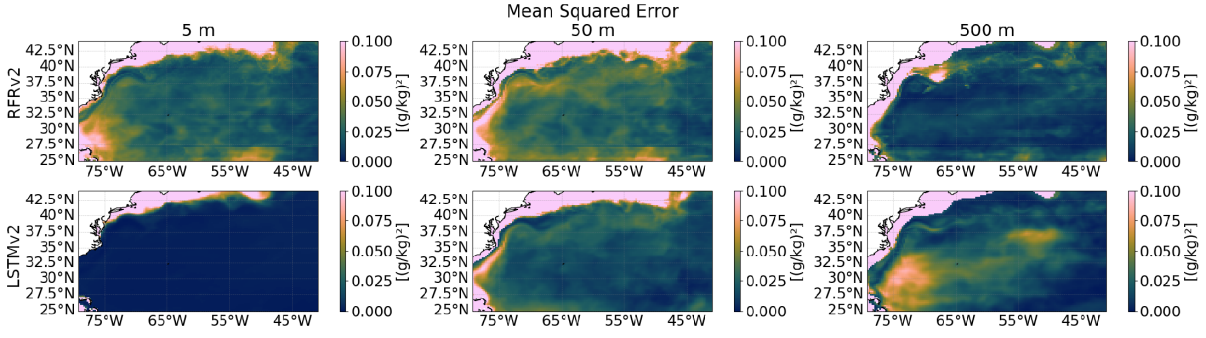

Figure 15. Mean Squared Error at different depths (from left to right: 5 m, 50 m, and 500 m) between predicted salinity and ground truth for the two proposed models RFRv2 (top row), LSTMv2 (bottom row).

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

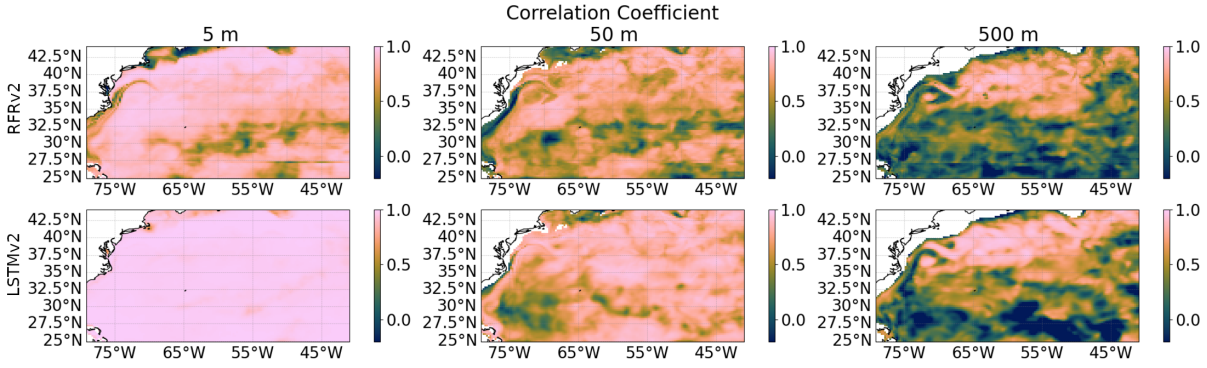

Figure 16. Temporal Correlation Coefficients at different depths (from left to right: 5 m, 50 m, and 500 m) between ground truth and predicted salinity for the two proposed models RFRv2 (top row), LSTMv2 (bottom row).

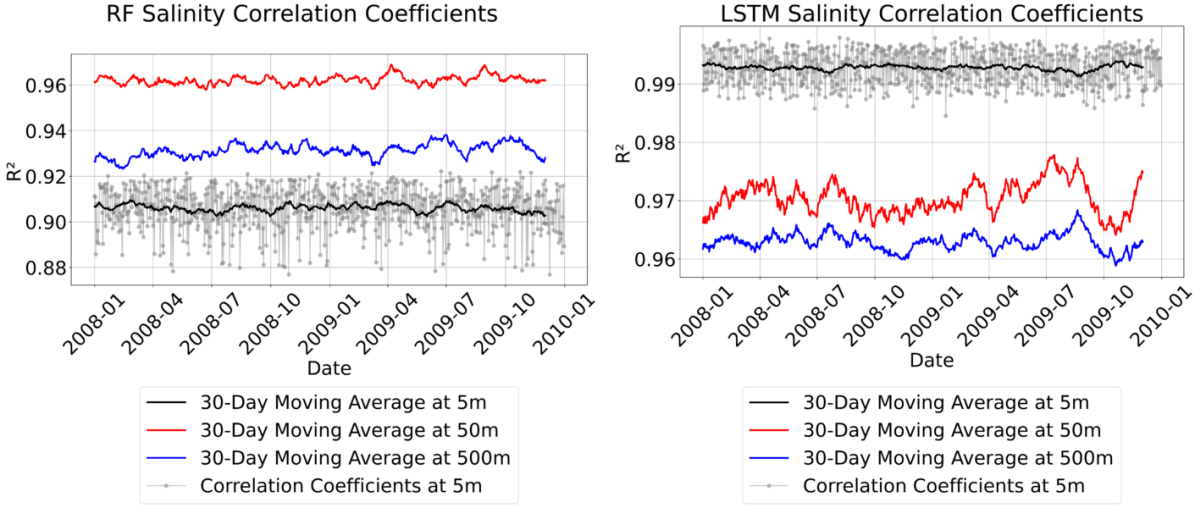

Figure 17. Salinity spatial correlation time series using a moving average of 30 days at different depths (5m black solid line, 50 m red solid line, and 500 m blue solid line) for the two proposed models for RFRv2 (left) and LSTMv2 (right).

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

5 Discussion

We implemented two approaches to study the feasibility of the 4D ocean reconstruction using the actual sampling capabilities provided by satellite and in-situ profilers, each of them with its own challenges and insights. The RFR exhibited artifacts due to the inclusion of latitude among input variables, suggesting that alternative techniques, such as incorporating neighboring region measurements instead of geographical coordinates, might be more effective for spatial contextualization. However, the representativity of the data might not be sufficient with the current ocean's sampling, as the high number of predictors relative to the limited spatial and temporal coverage of marine observations could lead to overfitting in the random forest model, potentially reducing its ability to generalize to undersampled regions or time periods. In the LSTM approach, we explored various configurations by adjusting the number of layers, activation functions, and units per layer. The final implementation consisted of a combined model for salinity and temperature predictions. Both architectures demonstrated particular strength in salinity reconstruction, achieving high accuracy in first and intermediate depths. However, correlation decreased in deeper layers with minimal ocean variations where climatological values might suffice. Temperature reconstruction showed superior results with the RFR compared to LSTM. Nevertheless, the artifacts produced by input variables limit their application in 4D reconstruction. The RFR might still be valuable in studies with less critical horizontal dimensions.

The validation with the test split dataset demonstrated that an increased number of surface variables improved the reconstruction of the water column, as the vertical profiles adjusted more faithfully to the ones of our ground truth and the spatial biases were smaller (Figs. 10 and 11). In the case of the LSTM, it also demonstrated that each reconstructed variable (salinity and temperature) requires a different tuning of the model. Overall, the RFR performed better than the LSTM with the test split dataset but when validating with the complete reanalysis product we observed better extrapolation and representativity of the data with the LSTM model in terms of correlation and variability assessment (Figs. 14, 16, and 17).

The contributions of each variable (SHAP values in Figs. 7 and 8) in the models have a geophysical meaning. The RFR excels at direct data relationships, thus, it uses the SSS as a base reference and modulates the variations with the rest of the variables such as SST, latitude, or SSH. In the case of the LSTM, we can observe that it can derive those relationships with the intrinsical patterns of the data, such as the latitudinal dependence of the SST. Furthermore, it tries to balance the weight of each of the input variables, giving enough weight to each of them to make them important in the decision.

We observed some biases in the validation with the reanalysis in Fig. 13. These biases can originate from the irregular sampling of the ocean as highly dynamic areas can attract Argo floats, making those regions more sampled. Even if the latitude and longitude coordinates are set as predictors, there is no smooth spatial transition between high-dynamical areas and calm waters. Another part of this bias is due to the well-known Bias/Variance trade-off present in these methodologies (Geman et al., 1992), where, to capture the variability of the data (which is what is interesting as it is the dynamical part of the data), one has to deal with higher biases. However, the bias appears to have smooth spatial gradients, making it easy to study and correct in future works if needed. The absence of seasonal variation in the spatial correlation indicates that the models can represent the seasonal cycles of the variables and the lack of variation (constant) of the time series indicates that the reconstruction is not affected by unaccounted trends.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

Both models were able to reconstruct the spatial and temporal variability patterns observed by the numerical model at different depths, which is a key aspect of studying the ocean's dynamics. However, we observed in Fig. 14 that both models tend to underestimate the variability range in the upper layers, losing part of the observed variability. The MSE values across different depths (Fig. 15) show consistently low values for both models throughout the different studied depths. This indicates that the model successfully captured different temporal dynamics intensities (calm waters vs. dynamical regions). This also suggests that the current sampling of the ocean provides adequate coverage of different water types for the model to learn the underlying patterns of their variability. The spatial correlations seen in Fig. 17 were also high in both models, achieving an R^2 score higher than 0.9 in both models in all the studied depths, indicating that the seasonal cycles and tendencies are well-captured by both of our models.

Overall, both models can capture the spatial and temporal variability of the ocean as seen by the reanalysis, with high correlations and accurate representations of seasonal cycles. However, the variability range is underestimated and should be improved in future work. The results obtained with these models offer promising prospects for ocean reconstruction with the current observing system, although improvements in the specific architectures can be made. For example, we could integrate them with other architectures such as diffusion networks or encoders/decoders which are specially used for high-resolution image generation. Using these technologies can provide a new perspective on how we observe and study the ocean.

6 Conclusions

Our study successfully demonstrated the feasibility of 4D ocean reconstruction using data-driven approaches and current observing systems although there is still much room for improvement in future works. The complexity of ocean dynamics across multiple dimensions presents significant challenges, requiring careful treatment of the input data and model architecture selection. While our models showed promising results in capturing ocean dynamics, particularly in vertical reconstruction, the horizontal dynamics reconstruction can be further improved.

Future work should focus on several aspects: investigating how to improve the temporal variability characterization, analyzing the evolution of biases to determine if constant corrections can preserve reconstructed variability, exploring alternative deep learning architectures for improved multi-dimensional reconstruction, and applying these models to real in-situ and satellite data. These findings contribute to our understanding of ocean reconstruction methodologies while highlighting the potential for further improvements in capturing the complex dynamics of ocean systems across all dimensions. This data-driven approach also contributes to further exploiting the synergy of the different and complementary ocean observation systems.

Code availability. The code used in this study is publicly available at GitHub (https://github.com/ainagarciaes/SSS-SST-4D-Reconstruction/tree/v1) under a GNU general public license and has been archived on Zenodo (DOI: 10.5281/zenodo.11487678). The repository contains all scripts necessary to reproduce the analyses presented in this paper.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

Data availability. This study has been conducted using the CMEMS Global Ocean Ensemble Physics Reanalysis (Mercator Océan International, 2025) dataset. It is accessible through the E.U. Copernicus Marine Service Information webpage (https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_ENS_001_031/description). The Argo profilers dataset (Argo, 2025) from which we derive the observed points are available through the SEANOE webpage (https://www.seanoe.org/data/00311/42182/) or all the alternative access to data options provided there. Mixed Layer Depth climatology dataset (de Boyer Montégut, 2023) can be downloaded from the Seanoe webpage https://www.seanoe.org/data/00806/91774/.

Appendix A: Temperature Validation

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

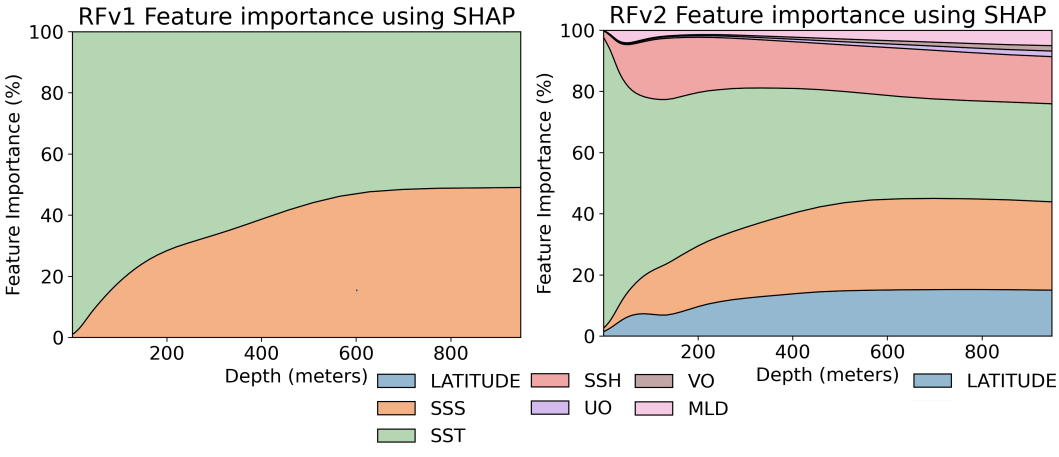

Figure A1. Feature importance percentage using SHAP values for RFRv1 (left) and RFRv2 (right) temperature models.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

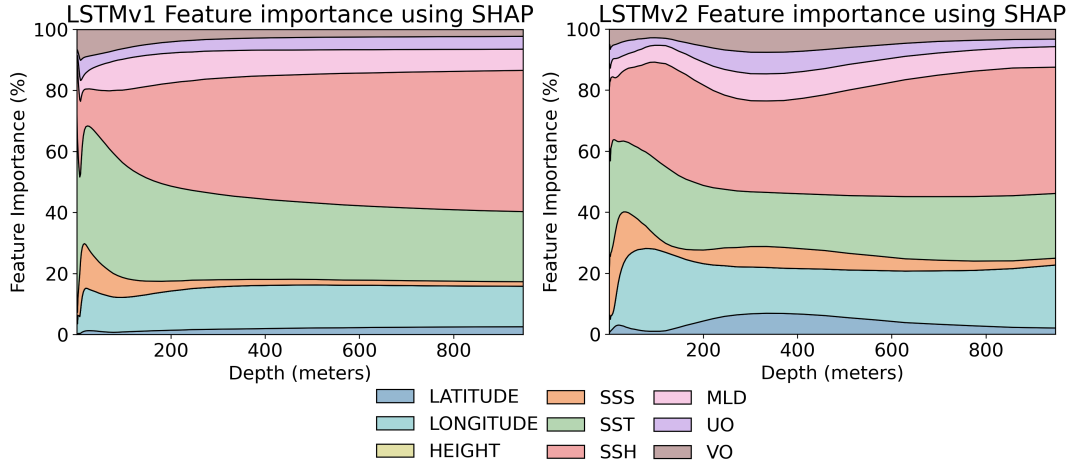

Figure A2. Feature importance percentage using SHAP values for LSTMv1 (left) and LSTMv2 (right) temperature models.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

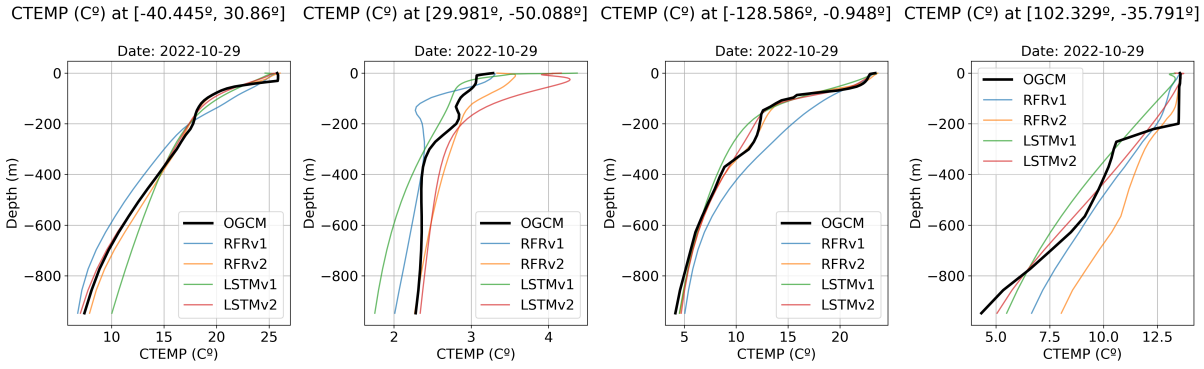

Figure A3. Predicted temperature vertical profiles for four different points (see Fig. 9 for its location) and the different models: RFRv1 (blue), RFRv2 (orange), LSTMv1 (green), LSTMv2 (red). The date is 29th October 2022.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

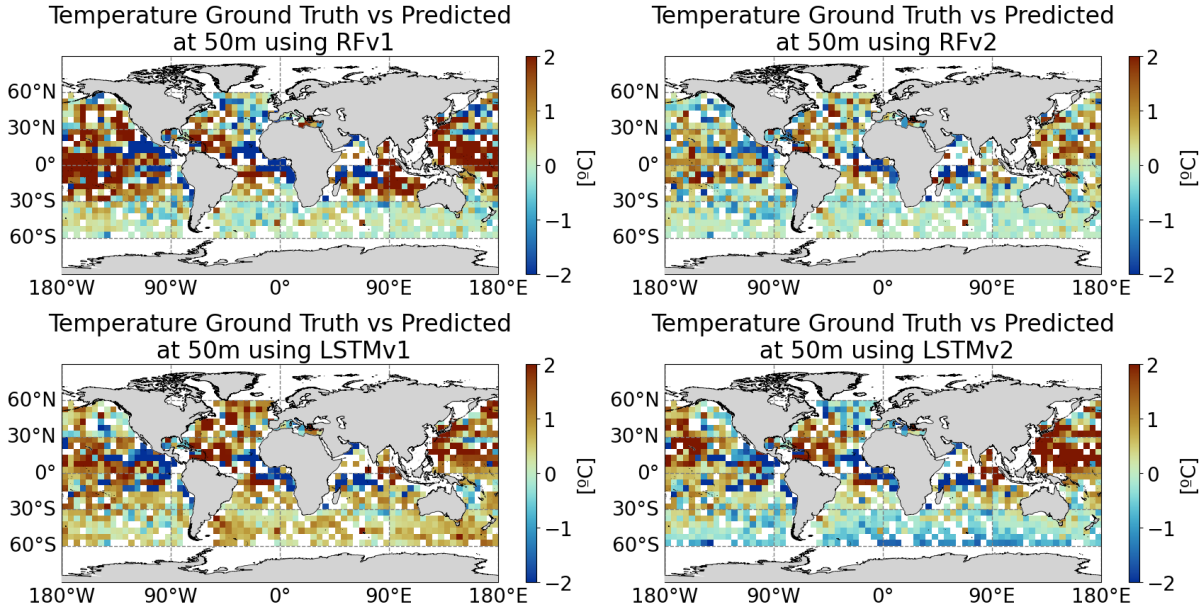

Figure A4. Predicted temperature minus Ground Truth (50 m depth) at a 5-degree grid for (from top to bottom and left to right): RFv1, RFv2, LSTMv1, LSTMv2. The date is 29th October 2022.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

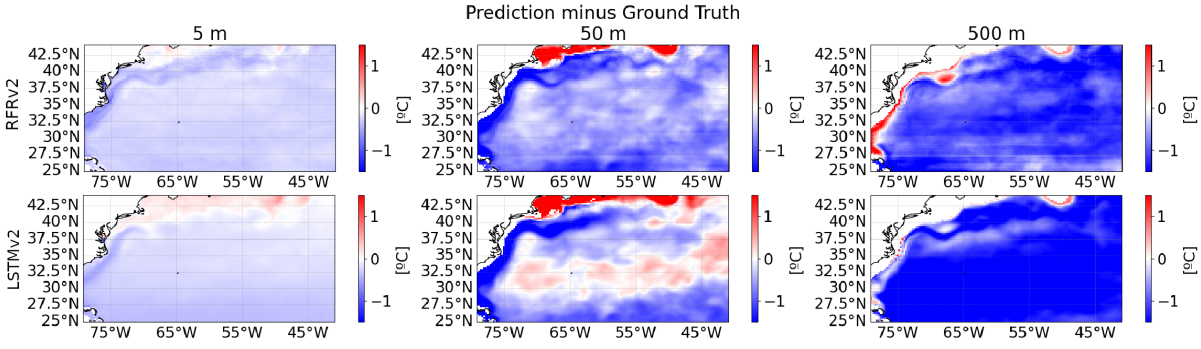

Figure A5. Biases at different depths (from left to right: 5 m, 50 m, and 500 m) between predicted temperature and ground truth for the two proposed models RFRv2 (top row), LSTMv2 (bottom row).

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

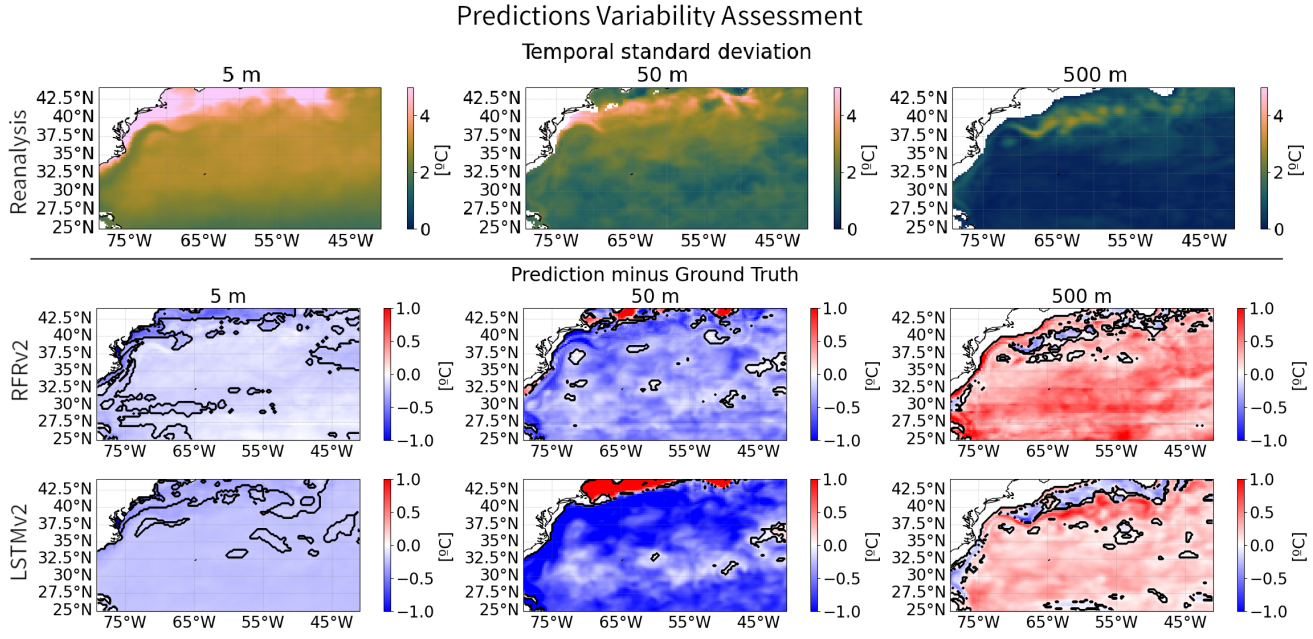

Figure A6. Variability assessment between the predictions and the ground truth (reanalysis) temperatures at different depths (from left to right: 5 m, 50 m, and 500 m). The standard deviation of the reanalysis for the 2008-2009 period (first row). Differences between the standard deviation of the ground truth and each model RFRv2 (second row), and LSTMv2 (third row). Solid black lines delimit statistically significant areas at a 95% confidence level assessed using an F-test for the difference of variances.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

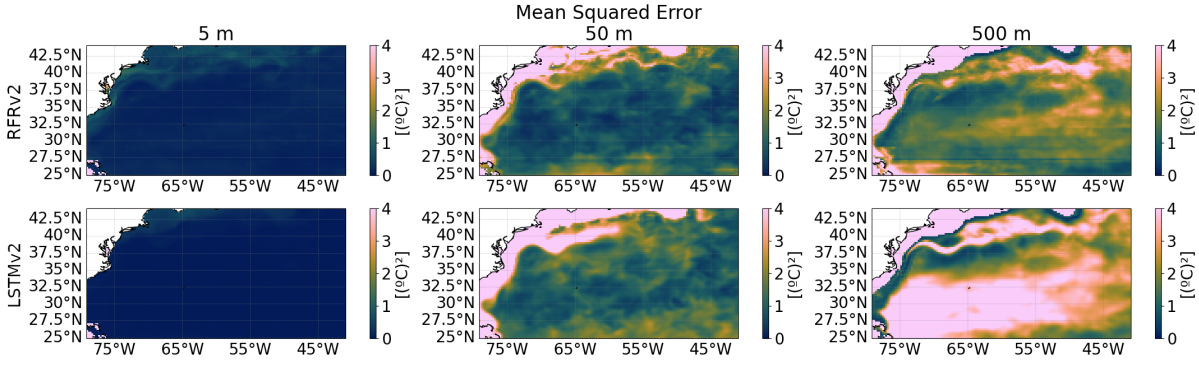

Figure A7. Mean Squared Error at different depths (from left to right: 5 m, 50 m, and 500 m) between ground truth and predicted temperature for the two proposed models RFRv2 (top row), LSTMv2 (bottom row).

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

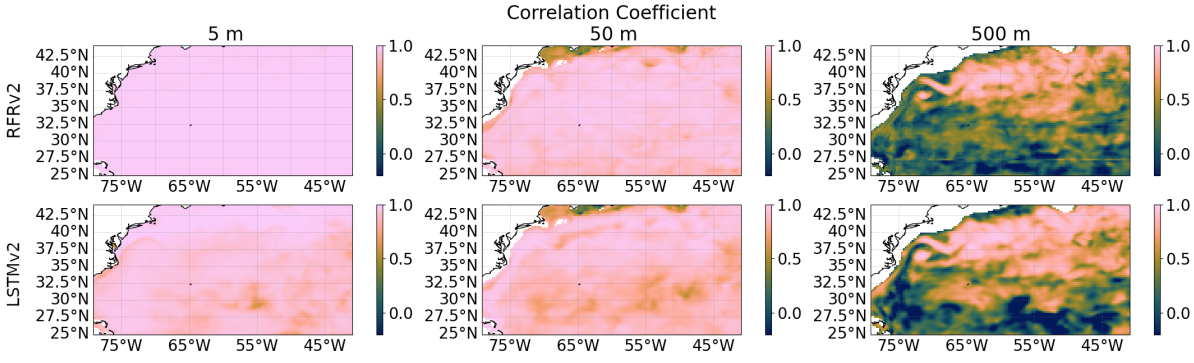

Figure A8. Temporal Correlation Coefficients at different depths (from left to right: 5 m, 50 m, and 500 m) between ground truth and predicted temperature for the two proposed models RFRv2 (top row), LSTMv2 (bottom row).

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

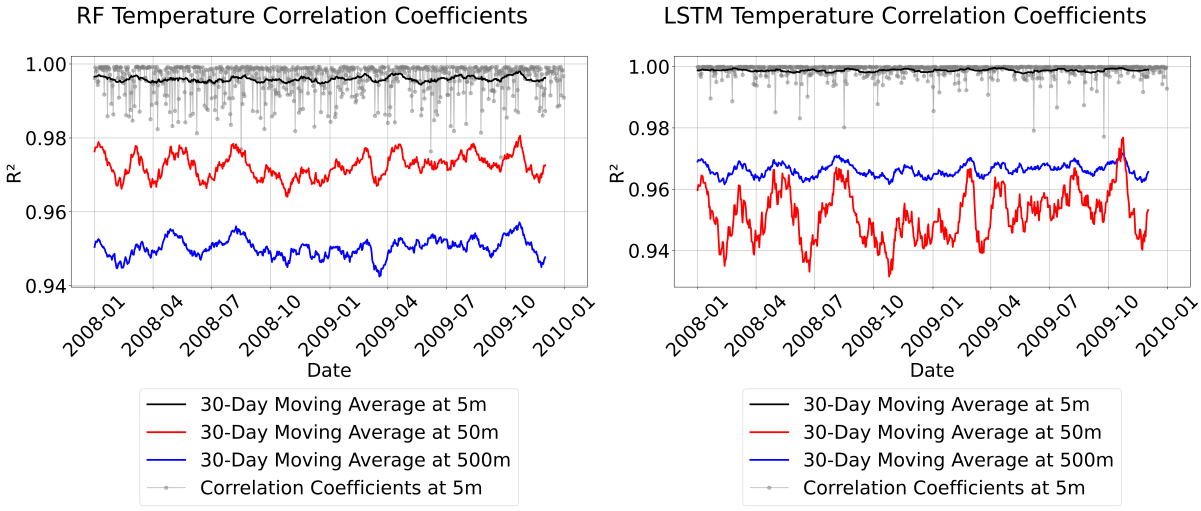

Figure A9. Temperature spatial correlation time series using a moving average of 30 days at different depths (5m black solid line, 50 m red solid line, and 500 m blue solid line) for the two proposed models for RFRv2 (left) and LSTMv2 (right).

Author contributions. Conceptualization by CGH; methdology by AGE and FA. Fundings and computational resources were acquired by FA. AGE did the data curation, formal analysis of the data, and visualization of the datasets and the obtained results. AGE, CGH, and FA did research and further experiments. The software was developed by AGE with the support of FA. The results were validated by AGE and CGH. This work was supervised by CGH and FA. The original draft was written by AGE and revised by CGH and FA.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the EO4TIP project, grant PID2023-149659OB-C21, funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU. This work was also supported by the European Maritime, Fisheries and Aquaculture Fund (EMFAF). The authors would like to thank the support and computing resources from AI4EOSC platform that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101058593. This work also acknowledges the "Severo Ochoa Centre of Excellence" accreditation, grant CEX2019-000928-S funded by MICIU/AEI/10.13039/501100011033. This work is moreover a contribution to CSIC PTI Teledetect.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/, software available from tensorflow.org, 2015.

Argo: Argo: Global array of profiling floats observing the ocean in real-time, Journal of Atmospheric and Oceanic Technology, 26, 2070–2079, https://doi.org/10.1175/2009JTECHO654.1, 2009.

Argo: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC), https://doi.org/10.17882/42182, 2025.

Argo Data Management Team: Argo Quality Control Manual for CTD and Trajectory Data, Tech. rep., Ifremer, 2024.

Boutin, J. et al.: ESA Sea Surface Salinity Climate Change Initiative (Sea Surface Salinity CCI), Earth System Science Data, 13, 2021.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Buongiorno Nardelli, B.: A Deep Learning Network to Retrieve Ocean Hydrographic Profiles from Combined Satellite and In Situ Measurements, Remote Sensing, 12, 3151, https://doi.org/https://doi.org/10.3390/rs12193151, 2020.

Casey, K. S. et al.: The past, present, and future of the AVHRR Pathfinder SST program, Oceanography from Space, 2010.

Chin, T. M. et al.: Multi-scale Ultra-high Resolution (MUR) analysis of sea surface temperature, Remote Sensing of Environment, 200, 2017.

de Boyer Montégut, C.: Mixed layer depth climatology computed with a density threshold criterion of 0.03kg/m^3 from 10 m depth value, https://doi.org/10.17882/91774, 2023.

Donlon, C. et al.: The Global Monitoring for Environment and Security (GMES) Sentinel-3 mission, Remote Sensing of Environment, 120, 2012.

Entekhabi, D. et al.: The soil moisture active passive (SMAP) mission, Proceedings of the IEEE, 98, 2010.

Geman, S., Bienenstock, E., and Doursat, R.: Neural Networks and the Bias/Variance Dilemma, Neural Computation, 4, 1–58, https://doi.org/10.1162/neco.1992.4.1.1, 1992.

Good, S., Fiedler, E., Mao, C., and Martin, e. a.: The Current Configuration of the OSTIA System for Operational Production of Foundation Sea Surface Temperature and Ice Concentration Analyses, Remote Sens., 12, 720, https://doi.org/10.3390/rs12040720, 2020.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 9, 1735–1780, 1997.

Kerr, Y., Waldteufel, P., Wigneron, J.-P., Martinuzzi, J., Font, J., and Berger, M.: The SMOS Mission: New Tool for Monitoring Key Elements of the Global Water Cycle, Proceedings of the IEEE, 98, 666–687, https://doi.org/10.1109/JPROC.2010.2043032, 2010.

Kilpatrick, K. A. et al.: A decade of sea surface temperature from MODIS, Remote Sensing of Environment, 165, 2015.

Lagerloef, G. et al.: The Aquarius/SAC-D mission: Designed to meet the salinity remote-sensing challenge, Oceanography, 21, 2008.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems (NIPS), vol. 30, pp. 4765–4774, Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf, 2017.

Meissner, T. et al.: Remote Sensing Systems SMAP Ocean Surface Salinities, Version 4.0, Remote Sensing Systems Technical Report, 2018.

Mercator Océan International: Global Ocean Ensemble Physics Reanalysis, E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS), https://doi.org/10.48670/moi-00024, [Online; accessed 28. Jan. 2025], 2025.

Merchant, C. J. et al.: Satellite-based time-series of sea-surface temperature since 1981 for climate applications, Scientific Data, 6, 2019.

https://doi.org/10.5194/egusphere-2025-705
Preprint. Discussion started: 24 February 2025

Olmedo, E., González-Haro, C., Hoareau, N., Umbert, M., González-Gambau, V., Martínez, J., Gabarró, C., and Turiel, A.: Nine years of SMOS sea surface salinity global maps at the Barcelona Expert Center, Earth System Science Data, 13, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Rew, R. K., Davis, G. P., Emmerson, S., and Davies, H.: NetCDF User's Guide for C, An Interface for Data Access, 1997.

Su, H., Wu, X., Yan, X.-H., and Kidwell, A.: Estimation of subsurface temperature anomaly in the Indian Ocean during recent global surface warming hiatus from satellite measurements: A support vector machine approach, Remote Sens. Environ., 160, 63–71, https://doi.org/10.1016/j.rse.2015.01.001, 2015.

Su, H., Huang, L., Li, W., Yang, X., and Yan, X.-H.: Retrieving Ocean Subsurface Temperature Using a Satellite-Based Geographically Weighted Regression Model, Journal of Geophysical Research: Oceans, 123, 5180–5193, https://doi.org/https://doi.org/10.1029/2018JC014246, 2018.

Su, H., Zhang, T., Lin, M., Lu, W., and Yan, X.-H.: Predicting subsurface thermohaline structure from remote sensing data based on long short-term memory neural networks, Remote Sens. Environ., 260, 112 465, https://doi.org/10.1016/j.rse.2021.112465, 2021.

Tian, T., Cheng, L., Wang, G., Abraham, J., Wei, W., Ren, S., Zhu, J., Song, J., and Leng, H.: Reconstructing ocean subsurface salinity at high resolution using a machine learning approach, Earth Syst. Sci. Data, 14, 5037–5060, https://doi.org/10.5194/essd-14-5037-2022, 2022.

Van Rossum, G. and Drake, F. L.: Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, ISBN 1441412697, 2009.

Webster, P. J.: The role of hydrological processes in ocean-atmosphere interactions, Reviews of Geophysics, 32, 427–476, https://doi.org/https://doi.org/10.1029/94RG01873, 1994.