Neural Synthetic Profiles from Remote Sensing and Observations (NeSPReSO) - Reconstructing temperature and salinity fields in the Gulf of Mexico.

Jose R. Miranda[a,b], Olmo Zavala-Romero[a,b], Luna Hiron[b], Eric P. Chassignet[b], Bulusu Subrahmanyam[c], Thomas Meunier[d,e], Robert W. Helber[f], Enric Pallas-Sanz[g], Miguel Tenreiro[g]

[a]*Department of Scientific Computing, Florida State University, Tallahassee, FL, United States*
[b]*Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, FL, United States*
[c]*School of the Earth, Ocean, and Environment, University of South Carolina, Columbia, FL, United States*
[d]*Woods Hole Oceanographic Institution, Woods Hole, MS, United States*
[e]*Laboratoire d'Océanographie Physique et Spatiale, Ifremer / UBO / CNRS / IRD, Plouzané, Brittany, France*
[f]*Ocean Sciences Division, US Naval Research Laboratory, Hancock County, MS, United States*
[g]*Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico*

Keywords: Synthetic temperature and salinity, machine learning, Gulf of Mexico, Loop Current, Data Assimilation

1. Introduction

Accurate representation of the Gulf of Mexico (GoM) circulation in numerical models is of great importance for the scientific community and holds operational significance for fisheries, hurricane prediction, and oil and gas companies (Jaimes et al., 2016; Koch et al., 1991; National Academies of Sciences, Engineering, and Medicine, 2018). The GoM Loop Current (LC) is part of the Atlantic western boundary current system and plays an important role in the transport of heat from the Caribbean Sea to the Atlantic Ocean, contributing to climate regulation. The LC also holds strong currents (up to 2 ms^{-1}) (Forristall et al., 1992; Sturges et al., 2005; Hiron et al., 2021) and is very dynamic, shedding large (≈200-400 km) warm eddies at an irregular

rate of 6 to 17 months (Vukovich, 1988; Behringer et al., 1977; Sturges and Leben, 2000). Loop Current Eddies (LCE) affect oil and gas activities in the GoM due to their strong peripheral velocities, and they can also fuel hurricane intensification by releasing heat to the atmosphere during storm passage (Shay and Uhlhorn, 2008; Shay, 2010; Jaimes et al., 2016). Cold-core, frontal eddies present in the vicinity of the LC contribute to the detachment of the LCEs and can enhance activity across the trophic chain by pumping deep-water nutrients to the upper ocean (Hiron et al., 2020, 2022; Suthers et al., 2023). Although recent model advancements have improved the representation of this complex system, a key limitation across ocean models remains the scarcity of in situ data to effectively constrain the models.

Temperature and salinity observations are two essential variables to be assimilated in numerical models, as density gradients, driven by these variables and pressure, govern large-scale ocean circulation. The ocean surface is well constrained in models, thanks to global satellite-derived sea surface height (SSH) and sea surface temperature (SST) data. However, subsurface observations are scarcer. The Argo program supports almost 4,000 floats worldwide that provide valuable information about the subsurface temperature and salinity structure of the ocean since 2005 (Roemmich and Gilson, 2009). In the GoM, the NAS-funded LC-floats and the UGOS 3 program are significant initiatives in subsurface observation. The LC-floats, supported by the National Academy of Sciences (NAS), are designed for oceanographic research in the GoM. Since June 2019, these floats have played a key role in collecting data on subsurface temperature and salinity structures. The UGOS 3 program, focusing on the GoM region, involves specialized floats that have contributed to more than 7,000 profiles sampled since the same period.

Despite their significance in constraining subsurface models, these measurements are too sparse, limiting the accurate representation of subsurface mesoscale circulation. Techniques such as Multiple Linear Regression (Carnes et al., 1994) Gravest Empirical Modes (GEM) method (Watts et al., 2001; Sun and Watts, 2001; Meunier et al., 2022) and the Improved Synthetic Ocean Profile (ISOP) system (Helber et al., 2013; Townsend et al., 2015; Helber et al., 2022) have been employed to generate synthetic temperature and salinity profiles for data assimilation in large-scale and regional ocean models. Those synthetic profiles rely on past observations and are generated mainly from altimetry SSH fields, based on the presumed relationship between SSH values and subsurface temperature and salinity, valid

for large-scale flows (geostrophic adjustment). Although promising, these methods can be computationally demanding and may not capture complex, non-linear relationships between surface and subsurface ocean fields.

In recent years, there has been significant advancement in deriving temperature and salinity profiles from ocean surface data using machine learning (ML) and artificial intelligence (AI) approaches. These models aim to bridge the gap between sparse in-situ measurements and satellite observations, enabling more comprehensive ocean monitoring. For instance, Chen et al. (2022) developed a machine learning-based assimilation system that uses a generalized regression neural network with fruit fly optimization to reconstruct T/S profiles from satellite observations, significantly improving the simulation of subsurface structures compared to direct assimilation of satellite data alone. Similarly, Tian et al. (2022) employed a feed-forward neural network to generate a high-resolution (0.25° x 0.25°) global subsurface salinity dataset by merging in-situ profiles with satellite altimetry, sea surface temperature, and wind data. Mao et al. (2023) developed a dual-path convolutional neural network to reconstruct ocean subsurface temperature and salinity from sea surface information, demonstrating improved accuracy over traditional methods. Pauthenet et al. (2022) reconstructed four-dimensional temperature, salinity, and mixed-layer depth in the Gulf Stream using neural networks, combining remote-sensing and in situ observations. These AI-based methods have shown promise in capturing mesoscale features and improving upon traditional interpolation techniques, offering new possibilities for generating comprehensive ocean T/S datasets with enhanced spatial and temporal resolution.

In the Gulf of Mexico, machine learning has been used in numerous applications, such as forecasting LCE shedding events (Zeng et al., 2015; Wang et al., 2019)), predicting hurricane wave height (Mafi and Amirinia, 2017), and estimating spatial and temporal variation in dissolved carbon dioxide near the Mississippi river outflow (Fu et al., 2020). Meng et al. (2021) developed a convolutional neural network (CNN) method using satellite-observed sea surface data (SSH, SST, sea surface salinity (SSS), and surface wind speed) and ocean subsurface temperature and salinity from Argo to obtain three-dimensional salinity fields from 0-2000 m depth. Despite these advancements, research with ML for subsurface modeling in the Gulf of Mexico is ongoing, as traditional methods still face challenges in efficiency, accuracy, and capturing the complex dynamics of the Gulf's circulation, especially at submesoscale.

In this study, we introduce NeSPReSO (Neural Synthetic Profiles from Remote Sensing and Observations), a method to effectively estimate subsurface temperature and salinity profiles using satellite-derived absolute dynamic topography (ADT), SST, and SSS by leveraging in-situ Argo data and Principal Component Analysis (PCA). Unlike previous methods, NeSPReSO focuses specifically on the Gulf of Mexico, utilizing a neural network architecture optimized for this region's oceanographic features. Our approach advances the field by combining PCA to reduce the dimensionality of the T/S profiles while capturing most of their variability, and a neural network that maps surface observations to these principal components. This methodology allows for efficient computation while capturing the complex, non-linear relationships between surface and subsurface ocean fields, thereby improving upon traditional methods and previous ML approaches in terms of accuracy and computational cost.

This study aims to address the following questions: How effectively can ML techniques, specifically neural networks (NN), be utilized to synthesize temperature and salinity profiles in the Gulf of Mexico? Can NeSPReSO provide an improvement over state-of-the-art methods? How do these synthetic profiles compare against independent measurements? Applications of this study include investigating the effects of assimilating the synthetic subsurface temperature and salinity profiles into hindcast and forecast numerical models in the Gulf of Mexico to determine whether they improve forecast accuracy. Additionally, we plan to provide a system through which the scientific community can request synthetic profiles for specific locations and time periods (depending on satellite data availability) to foster further research and applications.

2. Data

Our ML approach builds upon in situ observations and satellite-derived measurements. The following subsections details the specifics of each dataset, specifically Argo float, glider, and satellite datasets, as well as the ISOP statistics used as benchmark.

2.1. Argo Data

The main dataset for this study is a total of 4,145 temperature (T) and salinity (S) profiles acquired between 2015 and 2022 in the GoM region, and includes geographical coordinates, date, and time, as well as the estimated

local steric height referenced to 1,950 dbar (SH1950) for each profile. The distribution of these profiles is shown in Figure 1. T and S measurements were taken at one-meter intervals from the surface to a depth of 2,000 meters, capturing both major upper-ocean water masses present in the GoM: the warm and salty North Atlantic Subtropical Underwater (NASUW), typical of the Loop Current (SH1950 $\geq$ 0.17 m), and the fresher Gulf Common Water (GCW), representative of the Gulf waters (SH1950 $<$ 0.17 m) (e.g., Hiron et al. (2022)).

The dataset, described in detail by Meunier et al. (2022, 2023, 2024), includes a mixture of real-time and delayed mode profiles, re-processed without using the standard quality control (QC) flags. Outliers, defined as values outside four standard deviations, were removed, as well as profiles showing biased salinity at depth. Although these profiles could potentially be recovered with further processing, they were excluded from this analysis to maintain data consistency.

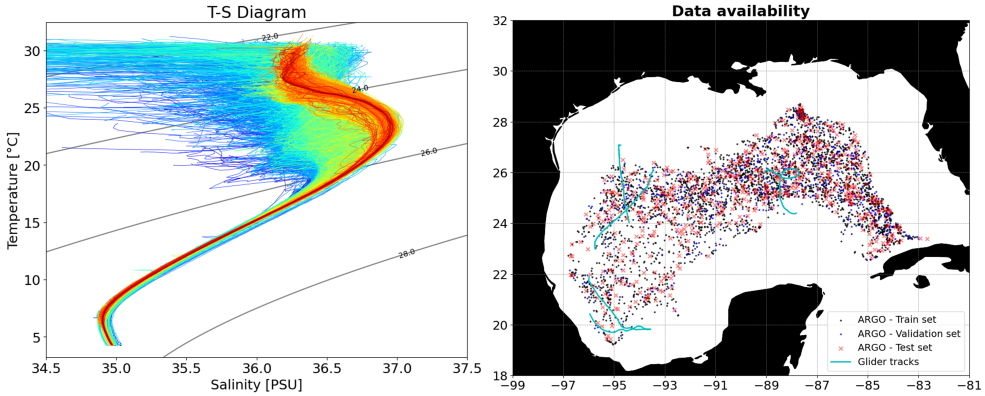


Figure 1: Temperature-Salinity (T-S) diagram (left) and spatial distribution (right) of glider tracks and Argo profiles used in this study. The T-S diagram identifies key water masses, including Gulf Common Water (GCW), North Atlantic Subtropical Underwater (NASUW), and Sub-Antarctic Intermediate Water (SAAIW). The spatial distribution uses markers/colors to represent dataset categories (train, validation, and test).

ISOP statistics are limited to the 0 to 1,800-meter range. Given that our Argo database has missing data beyond 1,800 meters, we restricted our dataset for model training, testing, and validation to this range.

2.2. Glider dataset

This dataset comprises T and S profiles from three missions (0006, 0010, and 0012) conducted between June 2017 and October 2018, targeting various mesoscale structures within the Gulf of Mexico by the glider oceanographic monitoring group (GMOG) from Cicese. These missions, executed using Seagliders equipped with a Seabird free-flow CT-sail, aimed to capture the vertical thermohaline variability associated with these mesoscale features. Data were collected at an averaged vertical resolution of 1 m and horizontal resolution of 3 km.

Missions 0006 and 0012 sampled old and young LCEs, respectively, and mission 0010 targeted a cyclonic eddy in Campeche Bay. During post-processing, data was vertically binned at 5 m intervals, and temperature adjusted for thermal lag, while thermal-inertia effects on conductivity were corrected following the methodology of Lueck and Picklo (1990). A fourth-order low-pass Butterworth filter with a cut-off frequency of $\frac{1}{48}h^{-1}$ was applied to smooth out high-frequency, near-inertial gravity waves. Missing segments were linearly interpolated to maintain the integrity of the profiles.

The gliders sampled contrasting thermohaline structures critical for assessing the reconstruction algorithm's proficiency. Significant differences in salinity ($\Delta S = 0.2$) and temperature ($\Delta T = 2°\mathrm{C}$) anomalies were observed between the eddies, with variations in the depth of the 26°C isotherm between young and old LCEs indicative of the effect of eddy age on thermohaline structure. However, large discrepancies are anticipated at the peripheries of the eddies due to submesoscale processes like density-compensated T and S layering and intrusions, which are not captured by the satellite fields, challenging the model's predictive capability in these areas.

2.3. Satellite data

Satellite-derived Absolute Dynamic Topography (ADT), sea surface temperature (SST) and salinity (SSS) were sourced from CMEMS, OISST, and SMAP, respectively. The Copernicus Marine Environment Monitoring Service (CMEMS) archives, validates, and interprets oceanographic satellite data. We utilized ADT, available since 1993, serving as a proxy for SSH. CMEMS provides an ADT gridded product with a daily resolution and a horizontal grid-spacing of approximately $\frac{1}{4}$ degrees (Copernicus Marine Service, 2024).

Optimum Interpolation Sea Surface Temperature (OISST) is a long-term climate data record that incorporates observations from different sources to

provide a high-resolution analysis of sea surface temperatures. It uses an optimal interpolation technique to combine data from satellites, ships, buoys, and other sources to create a consistent and accurate record of sea surface temperatures. Analysed SST is available since 1981 on a daily basis, with a resolution of approximately $\frac{1}{4}$ degrees (Good et al., 2020).

Finally, SMAP, or "Soil Moisture Active Passive", is a NASA satellite mission that uses active and passive microwave sensors to provide high-resolution measurements of soil moisture, freeze/thaw state, and ocean surface salinity. SMAP SSS has been available since 2015 on a daily basis and has a resolution of 40 km (Meissner et al., 2018).

The ADT, SST, and SSS fields are interpolated to each location of the Argo and glider databases using bicubic interpolation, and together with spatial and temporal information, serve as input to the proposed neural network as described in Section 3.2. Following Leben (2005) and Hiron et al. (2020), the daily mean of ADT over the GoM deep waters (> 200 m) is removed from the ADT field for each day. This removes the variations in ADT associated with thermal expansion and contraction of the upper ocean due to seasonal variability.

2.4. ISOP statistics

ISOP projects surface ocean data downward, generating T and S profiles across the global ocean using surface observations and a mixed-layer depth (MLD) estimate. Optionally, a prior forecast of T and S profiles can be used. The creation of these synthetic profiles plays an important step in the Navy's operational forecasting and is seamlessly integrated into their data assimilation workflows. ISOP divides the ocean's depth into 78 fixed levels, extending from the surface to 6600 meters. The process begins with the compilation of a T and S covariance matrix and climatology database from a comprehensive set of in-situ observations, followed by the application of a multilayered approach that considers three different dynamics zones within the ocean subsurface. These regions include the *mixed layer*, extending from the surface to the MLD; the *thermocline layer*, reaching from the MLD down to 1000 meters; and the *deep ocean layer*, below 1000 meters.

For the *mixed layer*, there are two options. One option adjusts the initial estimated profile to align with the surface potential density at 4 meters depth and ensures consistency with the potential density and its gradient at the MLD within the *thermocline layer*. The second option for the *mixed layer* shifts the prior forecast profile (if provided) to match the input SST value.

The *thermocline layer* prediction employs a variational method, leveraging climatological T and S values and the first vertical Empirical Orthogonal Functions (EOFs), or modes, extracted from historical data to constrain the forecast. Detailed descriptions of the each term involved in this variational approach is available in reference Helber et al. (2013). Finally, the prediction within the *deep ocean layer* involves modifying a decay function based on climatological data and the T and S readings from the *thermocline layer* at 1000 meters depth. This function accounts for the variance between climatological values and the 1000-meter predictions, ensuring a coherent transition into the deep ocean predictions. The inputs for ISOP's predictive models include SST and sea surface height anomaly (SSHA), along with uncertainty estimates, an MLD estimation, and an (optional) T and S profile can be obtained from either climatological data or model outputs. In this work, the synthetics used climatological data for estimating the initial MLD and T and S profiles, along with Argo-derived SST and SSH.

The ISOP data used in this work was generated by the US Navy and corresponds to the entire Argo dataset (4,145 profiles). The provided data included only the average vertical statistics and binned spatial statistics of the ISOP synthetics relative to the Argo profiles (no individual profiles were provided). These statistics were used as benchmark for the other methods.

3. Methods

In this section, we detail our methodology for training and validating a multilayer perceptron (MLP) to predict subsurface T and S profiles using surface data. The model is designed to learn the nonlinear functions that associates the ocean surface, through satellite observations, with subsurface information from a comprehensive dataset of Argo profiles. NeSPReSO uses PCA to focus the model on the main variability within the subsurface profiles, while also reducing the data's dimensionality and improving the efficiency of computation and training. Lastly, we assess the model's performance using unseen Argo profiles (15% of the dataset, randomly selected) and compare it with MLR, GEM and ISOP methods. The four unseen glider transects in the GoM were also reconstructed using our method, and compared with the original glider data.

The Argo float dataset, consisting of T and S profiles, is inherently high-dimensional, containing 1801 measurements (from 0 to 1800 meters at 1-meter intervals) for each parameter. In order to obtain an efficient model that

captures the overall shape of the profiles, we applied PCA to the data sets of the T and S profiles separately. By doing so, we can express each profile with a significantly reduced number of variables while retaining over 99% of the original data variability. Utilizing this transformation of data, we train the neural network to estimate the 30 most significant principal component scores (PCS) for each profile in the Argo dataset used for training, which are used to reconstruct the profiles using the inverse PCA.

Combining PCA with neural networks is an effective strategy for handling high-dimensional output spaces, as it reduces computational complexity and can improve prediction accuracy (Howley et al., 2006; Sun et al., 2023). PCA captures the most significant features in the data, and the neural network learns to predict these features from the inputs. This methodology has been successfully applied in various fields, including meteorology and oceanography (Preisendorfer and Mobley, 2023), finance (Sarıkoç and Celik, 2024), and engineering (Sun et al., 2023).

Figure 2 shows a general diagram of our methodology and the main components of the proposed neural network.

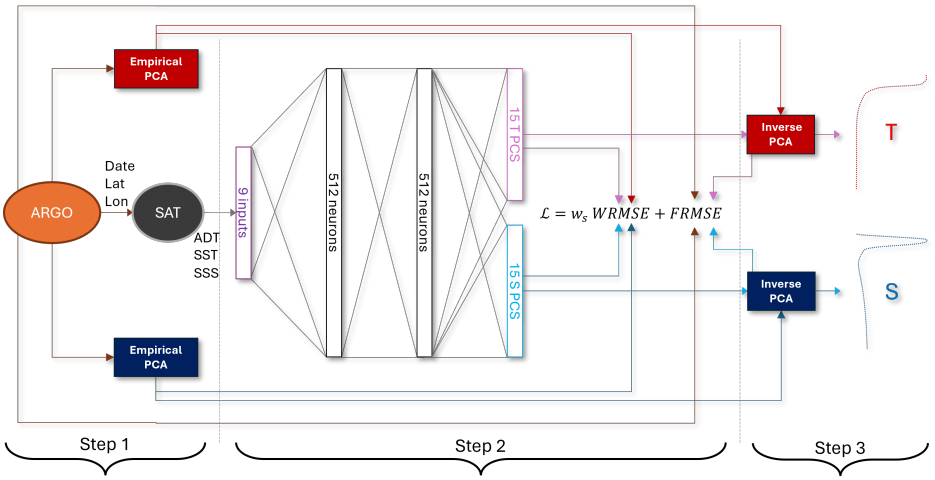


Figure 2: General diagram of NeSPReSO. Step 1 computes the empirical PCA of the Argo database. Step 2 trains a dense neural network from interpolated SST, SSH and SSS satellite data, location and date to predict the PCS. Step 3 reconstruct the profiles using the predicted PCS and inverse PCA.

3.1. Principal Component Analysis

Principal Component Analysis (PCA) is employed in various fields for dimensionality reduction of large datasets while preserving most of the original data variability. This method identifies orthogonal axes, known as principal components (PC), each representing a direction in which the data's variance is maximized.

Given a centered data matrix $\mathbf{Y}$ of size $n \times p$, where n is the number of observations (profiles) and p is the number of variables (measurements).

A covariance matrix $\mathbf{S}$ is computed as:

$$\mathbf{S} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}, \tag{1}$$

which captures the variances (in the diagonal) and the covariances (off-diagonals).

The next step involves solving the eigenvalue problem for $\mathbf{S}$:

$$\mathbf{SV} = \mathbf{DV}, \tag{2}$$

where $\mathbf{V}$ and $\mathbf{D}$ are the eigenvector matrix and eigenvalue diagonal matrix of $\mathbf{S}$, respectively. These eigenvectors define the directions of maximum variance in the data, and the eigenvalues indicate the magnitude of variance in these directions.

The eigenvectors and eigenvalues are arranged in descending order based on the magnitude of the eigenvalues. The first eigenvector, associated with the largest eigenvalue, becomes the first principal component (PC), and so forth. The eigenvector matrix $\mathbf{V}$, which is the concatenation of all $\mathbf{v_i}$ eigenvectors, is used to project the centered data matrix $\mathbf{Y}$ into the principal component space:

$$\mathbf{Z} = \mathbf{YV}, \tag{3}$$

where $\mathbf{Z}$ is a matrix of principal component scores (PCS), each column representing a principal component. To reduce dimensionality, $\mathbf{V}$ can be truncated, keeping only the eigenvectors corresponding to the largest eigenvalues.

The PCA transformation is linear and reversible. The inverse transformation, which approximates the original data from its reduced principal component representation, is given by:

$$\hat{\mathbf{Y}} = \mathbf{ZV}^T, \tag{4}$$

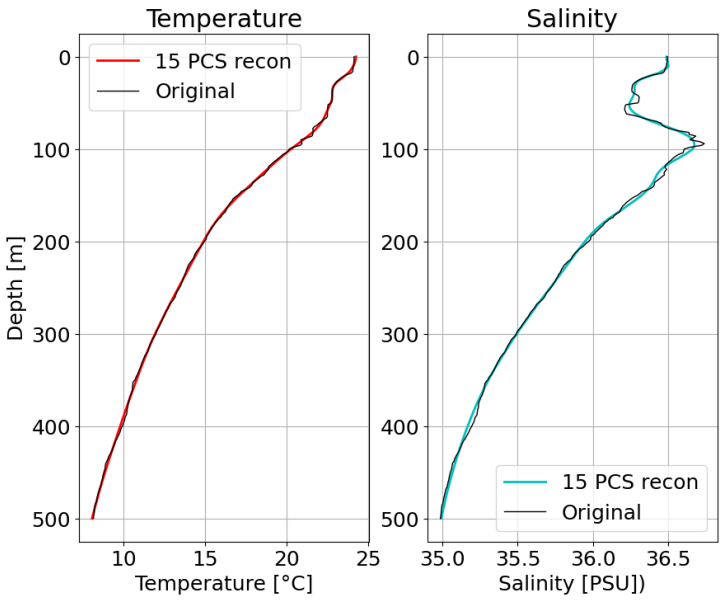


Figure 3: Example of reconstruction of temperature and salinity profiles using 15 PCS. The profile were truncated at 500 meters to emphasize the differences, which occur mostly in the upper ocean.

where $\hat{\mathbf{Y}}$ is the reconstructed data. Note that if $\mathbf{V}$ is truncated, this reconstruction is an approximation with some loss of information.

We applied PCA to the T and S datasets, reducing the dimensionality of the data (from 1801 to 15) by transforming the raw measurements ($\mathbf{Y}$) into PCS ($\mathbf{Z}$), while retaining most of the variance: 99.8% for temperature and 99.4% for salinity. Figure 3 illustrates the first 500 meters of a temperature and salinity profile and its reconstruction using 15 PCS.

Our proposed model is then trained to generate these 30 PCS for each Argo location in our training set. Next we describe NeSPReSO's architecture and training.

3.2. NeSPReSO

Let $X \subset \mathbb{R}^{d_X}$ denote our input space, representing spatial and temporal information along with surface measurements (e.g., sea surface temperature, salinity, and height), and let $Y \subset \mathbb{R}^{d_Y}$ be the output space consisting of the corresponding vertical profiles of temperature and salinity that we aim to predict. Our objective is to construct a mapping $\Phi : X \to Y$ such that for

each input vector $x \in X$, the predicted profile $y = \Phi(x)$ approximates the true profile $y \in Y$.

Due to the high dimensionality of the vertical profiles, directly predicting y with a neural network can be computationally intensive, inaccurate, and prone to overfitting. To address this, we employ Principal Component Analysis (PCA) for dimensionality reduction, focusing on modeling the most significant features of the profiles (Jolliffe and Cadima, 2016; Preisendorfer and Mobley, 2023). Formally, we encode the output space Y into a lower-dimensional space $Z \subset \mathbb{R}^{d_Z}$, where $d_Z \ll d_Y$, using an encoder E_Y such that $z = E_Y(y)$, and reconstruct the profiles with a decoder D_Y such that $y \approx D_Y(z)$.

Applying PCA to the profiles in Y yields the PCS z and defines the decoder operator $D_{PCA}(z) = z\mathbf{V}^{\mathrm{T}}$, where $\mathbf{V}$ is the matrix of eigenvectors from the PCA decomposition. Here, the encoder E_Y corresponds to the PCA transformation mapping profiles y to their PCS z, and the decoder D_Y corresponds to the inverse PCA transformation reconstructing y from z.

To predict z from the surface measurements x, we design a neural network $\zeta : X \to Z$ that approximates the mapping from the input space to the PCA space. This approach leverages the ability of neural networks to model complex nonlinear relationships between inputs and outputs. By training the neural network to predict z, we can reconstruct the full profiles using the inverse PCA transformation. Combining PCA with neural networks is a common practice in machine learning for handling high-dimensional outputs (Howley et al., 2006; Sun et al., 2023), as PCA reduces the output dimmensionality and the neural network captures the nonlinear relationships between inputs and principal components.

In designing the loss function for training the neural network ζ, we consider the accuracy of the reconstructed profiles. Specifically, we minimize the difference between the reconstructed PCS $\hat{z}$ and the true PCS z, and difference between the reconstructed profiles $\hat{y} = D_{PCA}(\hat{z})$ and the true profiles y. Our approximation process can be formalized as:

$$\min_{\zeta} \mathcal{L} = \frac{1}{nL_W} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{d_z} \frac{v_j}{\sigma_z^2} \left(\hat{z}_{ij} - z_{ij}\right)^2}_{WMSE} + \frac{1}{nL_F} \underbrace{\left(\sum_{i=1}^{n} \left(\frac{1}{\sigma_T^2} \sum_{k=1}^{d_Y} \left(\hat{Y}_{ik}^T - Y_{ik}^T \right)^2 + \frac{1}{\sigma_S^2} \sum_{k=1}^{d_Y} \left(\hat{Y}_{ik}^S - Y_{ik}^S \right)^2 \right) \right)}_{FMSE} \tag{5}$$

where $\mathcal{L}$ denotes the total loss function, n is the number of profiles (indexed by i), d_z is the number of principal components used (indexed by j),

and d_Y is the number of depth levels in each profile (indexed by k). The first term is the weighed mean squared error (WMSE) of the PCS, weighted by the variance captured by each component v_j, where $\hat{z}_{ij}$ and z_{ij} represent the predicted and true PCS for sample i and component j, respectively. The second term represents the functional mean squared error (FMSE), which is computed for both temperature and salinity profiles. Specifically, $\hat{Y}_{ik}^T$ and Y_{ik}^T denote the predicted (after inverse PCA transformation) and true temperature values, respectively, at depth k for sample i. Similarly, $\hat{Y}_{ik}^S$ and Y_{ik}^S represent the predicted and true salinity values.

It's important to note that in our model $\mathcal{L}$ accounts for both temperature and salinity predictions simultaneously, which have different scales and units. To ensure that the contributions of these parameters are appropriately scaled in this multi-task model, each mean squared error term is divided by the variance of the respective parameter: σ_z^2 for the PCS, σ_T^2 for temperature, and σ_S^2 for salinity (Zhang and Yang, 2017).

Additionally, training the model using WMSE or FMSE individually results in different loss values, with $L_W \approx 0.0255$ for WMSE and $L_F \approx 2.8294$ for FMSE. These values are used to normalize each term when combining them in the final loss function.

The neural network used in this study consists of a simple multilayer perceptron, suitable for regression tasks involving continuous outputs, with an input layer that receives satellite-derived ADT, SST, and SSS bi-cubicly interpolated to the location of each Argo profile. It also receives spatial information coming from the latitude and longitude. Recognizing that latitude and longitude represent angular measurements with cyclical properties, we compute the sine and cosine harmonics for each normalized temporal and spatial inputs ($2\pi\frac{lat}{180}$, $2\pi\frac{lon}{360}$ and $2\pi\frac{day}{365}$), helping the network to capture the cyclical nature of these parameters, which has been shown to improve model performance in previous studies (Thottakkara et al., 2016). The output layer produces the predicted PCS, which are then used to reconstruct the full temperature and salinity profiles using the inverse PCA transformation.

We use 2 fully connected hidden layers with 512 neurons each, employing the Rectified Linear Unit (ReLU) activation function to reduce computational complexity and mitigate vanishing gradients (Dubey et al., 2022; Nguyen et al., 2021). To prevent over-fitting, we apply a dropout rate of 20%, randomly disabling neurons during training, which encourages the network to learn more robust features (Zhang et al., 2024). Additional training parameters include using a batch size of 300, a maximum number of 8000

epochs and an early stopping mechanism of 500 epochs, if the loss value in the validation set is not improved.

The training of the neural network involves an iterative process where the model learns to approximate the PCS through exposure to different subsets of the data. The model is trained using 70% of the profiles (2,895 in total), while its performance is continuously monitored against a separate validation set comprising 621 (15%) profiles, which effectively determines when the training should stop. Training the model on this setting took 8 minutes using a single GPU. Evaluation of the model's accuracy is conducted on the remaining 15% of the data (621 profiles), the test set, to assess its predictive capabilities.

NeSPReSO is compared against standard models for creating synthetic profiles: Multiple Linear Regression, Gravest Empirical Modes and Improved Synthetic Ocean Profile.

3.3. Multiple Linear Regression Approach

In addition to the neural network architecture, we explore a MLR model as a baseline method for predicting the PCS from surface measurements (Carnes et al., 1994). The MLR serves to assess the effectiveness of the neural network by comparing its performance with a simpler, linear approach.

Let us consider the same input space $X \subset \mathbb{R}^{d_X}$, output space $Y \subset \mathbb{R}^{d_Y}$ and the reduced-dimensional space $Z \subset \mathbb{R}^{d_Z}$, where $d_Z \ll d_Y$, along with the encoder E_Y and decoder D_Y mappings. The MLR model aims to establish a linear relationship between the input variables in X and the PCS in Z. Specifically, we model each principal component score z_j as a linear combination of the input features:

$$\hat{z}_j = \beta_j + \sum_{i=1}^{d_X} \beta_{ij} x_i, \tag{6}$$

where $\hat{z}_j$ is the predicted PCS for component j, β_j is the intercept term, β_{ij} are the regression coefficients, and x_i represents the input features from X. The regression coefficients β are then estimated by solving the least squares problem:

$$\beta = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}, \tag{7}$$

where $\mathbf{Z}$ is the matrix of true PCS obtained from PCA, and $\mathbf{X}$ is the expanded feature matrix. The inverse operation $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ denotes the pseu-

doinverse when $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible. This estimation provides the exact least squares solution for the regression coefficients.

The MLR model predicts the PCS by applying the estimated coefficients to new input data:

$$\hat{\mathbf{Z}}_{MLR} = \mathbf{X}_{\mathrm{new}}\boldsymbol{\beta}, \tag{8}$$

where $\mathbf{X}_{\mathrm{new}}$ contains the polynomial features of the new input samples. The predicted PCS $\hat{\mathbf{Z}}_{MLR}$ are then used with the decoder D_Y to reconstruct the full temperature and salinity profiles:

$$\hat{Y}_{MLR} = D_Y(\hat{\mathbf{Z}}_{MLR}) = \hat{\mathbf{Z}}_{MLR}\mathbf{V}^{\mathrm{T}}, \tag{9}$$

where $\mathbf{V}$ is the matrix of eigenvectors from the PCA decomposition.

In our implementation, we include the same inputs as in our NN approach: spatial and temporal harmonics of latitude, longitude, day of the year, and satellite SST, SSH and ADT. The MLR model is trained using the combined training and validation datasets, comprising 3,516 profiles (85% of the total data), to ensure sufficient data for estimating the regression coefficients accurately. Fitting the model took 180 milliseconds on a single GPU.

The remaining 15% of the data (621 profiles) is used as a test set to evaluate the model's predictive performance. By comparing the MLR results with those of the neural network, we can assess the benefits of incorporating nonlinear activation functions and deeper architectures in capturing complex relationships within the data, and by comparing with GEM, we can assess the advantages of operating in a reduced dimensional space.

It's important to note that we initially experimented with polynomial expansions up to degree 3 to capture potential nonlinear relationships between the input variables and the PCS. However, these higher-degree models exhibited significant issues:

- Computational Challenges: The inclusion of polynomial terms up to degree 3 dramatically increased the dimensionality of the feature matrix. With a large number of samples and input variables, the feature matrix became extremely large. This led to high memory consumption ($\approx$ 80GB) and computational inefficiency and instabilities during model fit.

- Numerical Instability: The large size of the matrices exacerbated numerical issues, such as difficulty in inverting matrices during the estimation of regression coefficients. This instability adversely affected the model's ability to learn accurate relationships.

- Overfitting: The expanded feature space increased the risk of overfitting, where the model captured noise rather than underlying patterns, resulting in poor generalization to unseen data.

- Multicollinearity: Higher-degree polynomial terms introduced strong correlations among predictor variables, destabilizing coefficient estimates and reducing the reliability of the model.

As a result of these challenges, the higher-degree polynomial models were unstable, producing predictions that were too inaccurate for practical application. Therefore, we opted to use the degree 1 MLR model, which captures linear relationships between the input variables and the PCS.

3.4. Gravest Empirical Modes

The GEM method is a technique extensively utilized in oceanography for the generation of synthetic temperature and salinity profiles. The GEM method is based on the establishment of an empirical relationship between dynamic height and other oceanographic parameters, capturing the essential spatiotemporal patterns of oceanic temperature and salinity, making it a valuable tool for studying and simulating these parameters. This method has been applied to various oceanic regions, contributing to a better understanding of ocean dynamics and climate processes (Watts et al., 2001; Liu et al., 2021; Meunier et al., 2022).

The implementation of the GEM method is described as follows:

A. The steric height is computed for each in situ profile of temperature and salinity.

B. All profiles are sorted according to their steric height, and grouped by month.

C. A regular pressure grid is defined (0–1800 dbar) with a vertical grid-step of 1 dbar. For each reference pressure value and for each month, a cubic smoothing spline is fitted to the functions $T(\zeta)|_{p,m}$ and $S(\zeta)|_{p,m}$, where T and S are temperature and salinity, ζ is ADT, p is the pressure at which the variables are evaluated, and m is the month.

The process of fitting GEM to the dataset took 3 seconds on CPU.

4. Results

In this section we analyze the performance of NeSPReSO with respect to 621 Argo profiles in our test dataset (15% of the dataset, randomly selected, not used in training), and compare its performance against GEM, MLR and ISOP methods. We also generate NeSPReSO synthetics to reconstruct four glider sections in the GoM.

The average processing time per profile on CPU is around 60µs for NeSPReSO, 20µs for MLR and 11600µs for our GEM implementation, when generating synthetics for our test set. However, it's important to note that in an operational setting, where profiles are generated on the fly, the time to extract the satellite information from stored data is the limiting factor for generating synthetics, regardless of the method used (0.5s per day of interest, regardless of the number of profiles).

NeSPReSO and MLR synthetics were generated using satellite surface information (ADT, SST and SSS) interpolated to the locations of the measurements, location, and day of the year, while GEM synthetics used month and ADT. ISOP utilized climatological MLD and profile-derived SSH and SST, with only statistical summaries of the ISOP synthetics being available, rather than individual profiles. This limitation, along with the fact that ISOP synthetics was not derived from satellite sources like the other methods, may skew the comparison in the upper ocean.

4.1. Test set

We use root mean square error (RMSE) and bias as analysis metrics to evaluate the performance of our model relative to observations. RMSE, measuring precision and accuracy, indicates the model's prediction consistency and closeness to observed values. RMSE penalizes larger deviations and reflects the average prediction error, with lower RMSE indicating more reliable predictions. Bias measures the average deviation from observed values, showing if the model consistently overestimates or underestimates the variable under consideration. Both statistics are given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}, \tag{10}$$

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right), \tag{11}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and N is the number of observations. For calculations at each depth level, N represents the number of profiles at that depth. When computing RMSE and bias over a depth range, the statistics are averaged over all depths within that range.

The Pearson correlation coefficient (R^2) quantifies the degree of linear correlation between the predicted and observed values, with values closer to 1 indicating a stronger correlation. It is calculated as:

$$R^2 = \left(\frac{\sum_{i=1}^{N} (y_i - \bar{y}) \left(\hat{y}i - \bar{\hat{y}} \right)}{\sqrt{\sum i = 1^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N} \left(\hat{y}_i - \bar{\hat{y}} \right)^2}} \right)^2, \tag{12}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the mean values of the observed and predicted data, respectively. The R^2 metric assesses the proportion of variance in the observed data that is predictable from the predicted data. Since we don't have access to individual ISOP synthetics, we could not calculate R^2 for ISOP.

The statistics of the profiles in the test set are shown on table 1, calculated using predictions at the same depths as ISOP, for fairness. For temperature, the RMSE values indicate that NeSPReSO consistently outperforms the GEM predictions across all depth ranges, MRL below 20 meters and ISOP below 100 meters. However, it is difficult to draw comparisons with ISOP near the surface, given that it uses Argo SST, but we observe a more accurate estimation of temperature profiles compared to the GEM method, which we attribute to the use of satellite SST. Bias values for temperature are comparable between all methods, implying that the methods exhibit a similar direction and magnitude of systematic error in temperature estimation. For salinity, NeSPReSO also demonstrates lower RMSE and bias values than the other methods for most of the depth ranges, indicating superior performance in salinity predictions.

The Pearson correlation coefficient (R^2) values for both T and S predictions are higher for NeSPReSO compared to GEM across all depths, and particularly pronounced in the upper 100 meters. NeSPReSo also overperforms MLR in most cases, except for T on the range from 0 to 20 meters. This improvement in R^2 signifies a stronger correlation between predictions and observations, meaning a better characterization of the upper-ocean.

Table 1: Statistics (RMSE, Bias, and R^2) by depth range. Best results in bold.

Depth range		0-20	20-100	100-200	200-500	500-1000	1000-1800	0-1000	0-1800
T	NeSPReSO	0.430	**0.816**	**0.802**	**0.587**	**0.301**	**0.083**	0.682	0.637
RMSE	GEM	1.468	1.419	1.094	0.854	0.394	0.125	1.195	1.116
(°C)	MLR	0.380	1.031	0.944	0.699	0.357	0.087	0.823	0.768
	ISOP	**0.140**	0.835	0.917	0.756	0.360	0.111	**0.673**	**0.598**
T	NeSPReSO	0.047	**-0.038**	**0.015**	0.016	**0.005**	0.003	**0.001**	**0.001**
BIAS	GEM	-0.043	-0.153	-0.059	-0.036	0.006	0.006	-0.077	-0.067
(°C)	MLR	**-0.011**	-0.041	0.016	**-0.001**	-0.010	**0.000**	-0.014	-0.012
	ISOP	0.022	0.186	0.203	0.137	-0.057	-0.074	0.127	0.102
T	NeSPReSO	0.983	**0.956**	**0.971**	**0.986**	**0.987**	**0.973**	**0.995**	**0.997**
R^2	GEM	0.773	0.870	0.949	0.970	0.978	0.941	0.986	0.991
	MLR	**0.986**	0.929	0.960	0.980	0.981	0.970	0.993	0.996
S	NeSPReSO	**0.280**	**0.139**	**0.116**	**0.088**	**0.032**	**0.009**	**0.154**	**0.143**
RMSE	GEM	0.478	0.193	0.163	0.122	0.046	**0.009**	0.241	0.225
(PSU)	MLR	0.299	0.154	0.155	0.112	0.044	**0.009**	0.173	0.162
	ISOP	0.604	0.229	0.160	0.147	0.049	0.015	0.240	0.210
S	NeSPReSO	**0.012**	**-0.002**	0.005	0.003	**-0.001**	**0.000**	**0.003**	**0.002**
BIAS	GEM	-0.036	-0.010	-0.014	-0.005	0.002	**0.000**	-0.013	-0.011
(PSU)	MLR	-0.021	-0.007	**0.002**	**0.001**	**-0.001**	**0.000**	-0.005	-0.005
	ISOP	-0.092	-0.086	-0.033	0.023	-0.009	-0.010	-0.048	-0.043
S	NeSPReSO	**0.829**	**0.729**	**0.887**	**0.985**	**0.977**	**0.861**	**0.962**	**0.975**
R^2	GEM	0.337	0.411	0.789	0.971	0.958	0.833	0.905	0.939
	MLR	0.803	0.654	0.786	0.975	0.957	0.857	0.952	0.969

Figure 4 presents the average T and S RMSE and bias per depth for all methods. In general, NeSPReSO yields better approximations compared to the other methods, as indicated by the lower RMSE and bias values overall. The improved prediction of upper-ocean temperature and salinity profiles in our model compared to GEM is likely due to the use of satellite SST and SSS, which offer additional information about the upper thermal and haline structures that might not be captured in the ADT fields, such as low salinity due to river outflow.

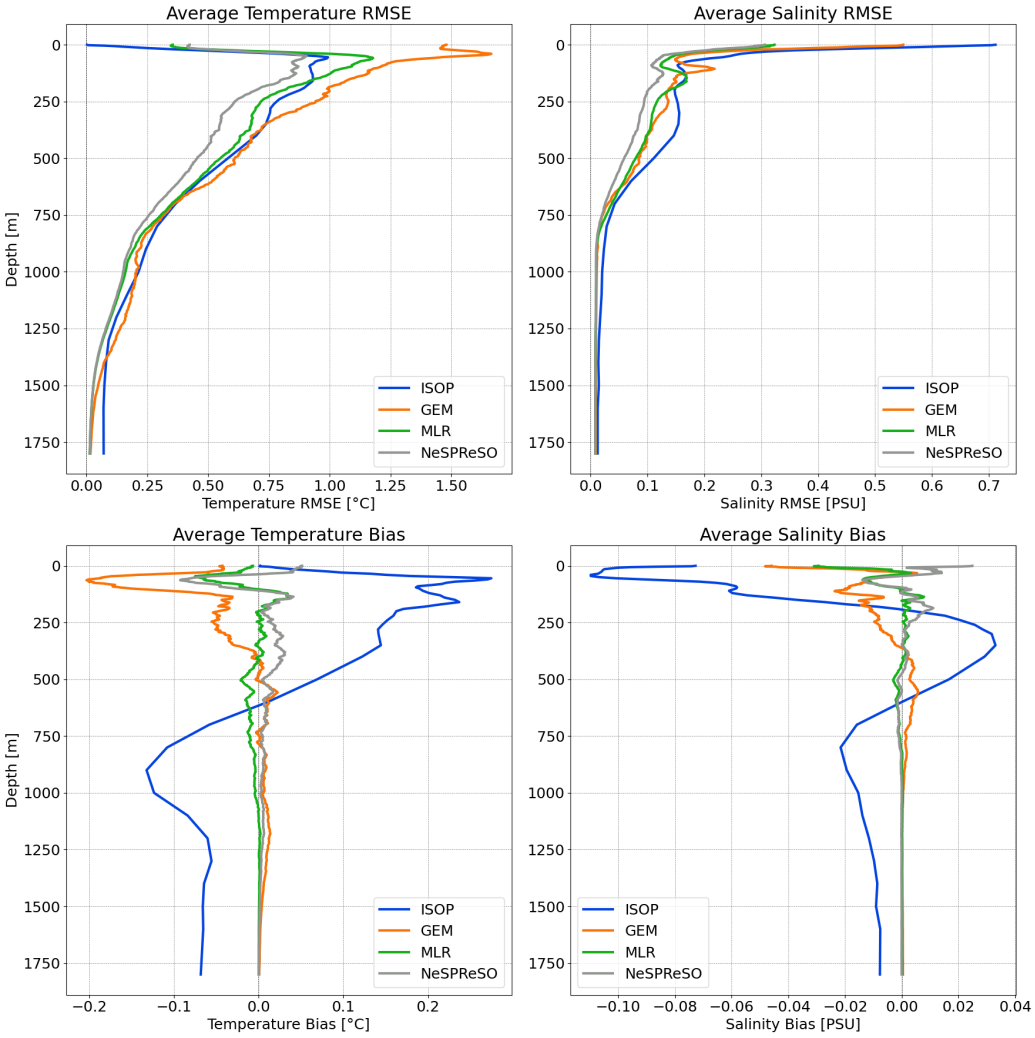


Figure 4: Average RMSE for temperature and salinity predictions (top), and average bias (bottom) as a function of depth.

The synthetic profiles were aggregated spatially into 1-degree latitude by 1-degree longitude grid cells to assess the methods' performance in predicting T and S across the area of study. Figures referenced as 5 through 8 present the spatial distribution of RMSE and bias for T and S. The statistics were calculated using predictions at the same depths as ISOP for a fair comparison.

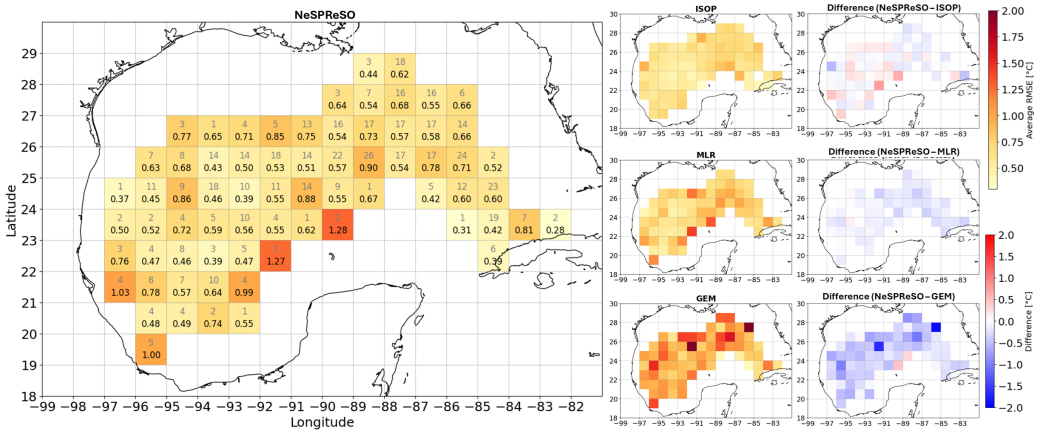


Figure 5: Distribution of average temperature RMSE for predictions down to 1,800m for NeSPReSO (left), with the number of profiles in each bin is displayed in gray, and RMSE values in black. Statistics for ISOP (top), MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO are shown on the right column (blues indicate NeSPReSO performs better, and reds indicate NeSPReSO performs worse).

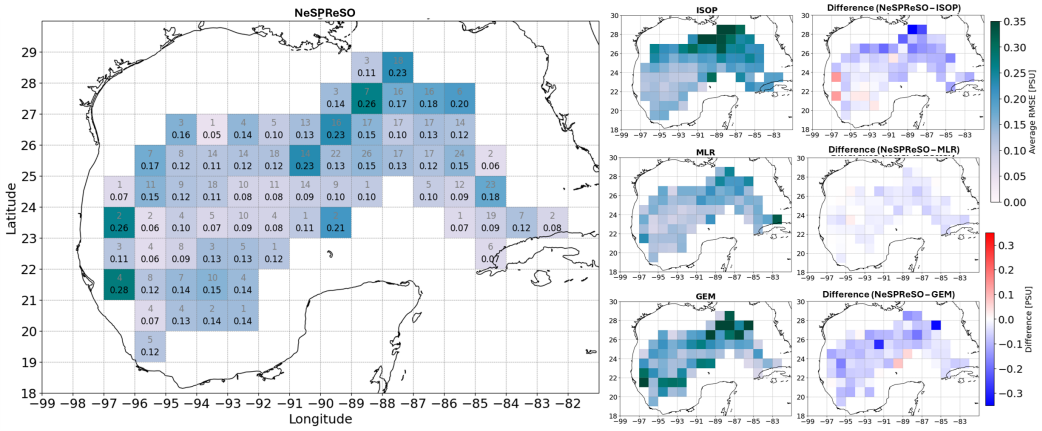


Figure 6: Distribution of average salinity RMSE for predictions down to 1,800m for NeSPReSO (left), with the number of profiles in each bin is displayed in gray, and RMSE values in black. Statistics for ISOP (top), MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO are shown on the right column (blues indicate NeSPReSO performs better, and reds indicate NeSPReSO performs worse).

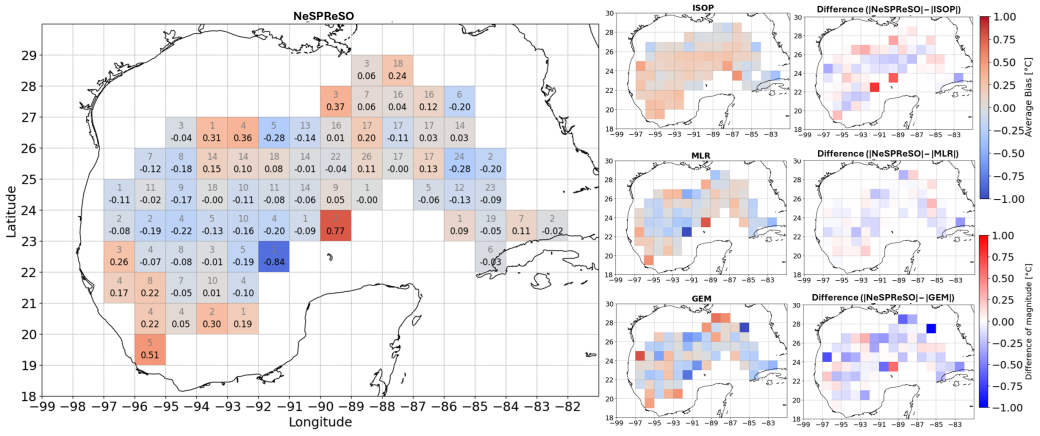


Figure 7: Distribution of average temperature bias for predictions down to 1,800m for NeSPReSO (left) with the number of profiles in each bin is displayed in gray, and bias values in black. Statistics for ISOP (top), MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO are shown on the right column (blues indicate NeSPReSO performs better, and reds indicate NeSPReSO performs worse).

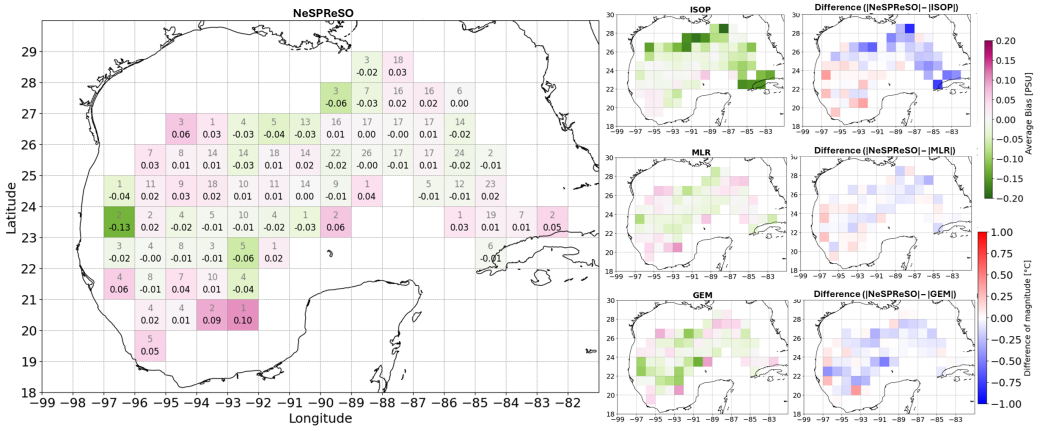


Figure 8: Distribution of average salinity bias distribution for predictions down to 1,800m for NeSPReSO (left), with the number of profiles in each bin is displayed in gray, and bias values in black. Statistics for ISOP (top), MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO are shown on the right column (blues indicate NeSPReSO performs better, and reds indicate NeSPReSO performs worse).

The results indicate a robust performance of NeSPReSO in real-world

scenarios and applications, as NeSPReSO has lower overall RMSE for both T and S predictions across the entire GoM region, with a few exceptions. NeSPReSO shows a spatial distribution of bias predominantly of low magnitude and somewhat homogeneous (no apparent predominant bias). MRL has a very similar spatial distributions as NeSPReSO, with slightly higher magnitudes. GEM also demonstrates a relatively homogeneous distribution, but with even higher magnitude on average. Meanwhile, ISOP exhibits a clear warmer and low magnitude trend for T and fresher for S, with greater magnitudes in the eastern portion of the GoM. Notably, in regions adjacent to the Mississippi River, ISOP demonstrates increased errors.

4.2. Glider tracks

This section presents a comparative analysis of processed glider tracks against the reconstructions from NeSPReSO, offering a direct assessment of the model's performance by replicating independent observations.

Figures 9 to 12 illustrate four different processed glider crossings with the corresponding synthetic reconstructions and the differences. Overall, the displacement of isothermals and isohalines are in agreement with the observations, and the reconstructed fields are smoother, as expected.

Table 2 shows the RMSE, bias, and R^2 for each LCE crossing. The T and S RMSE closely aligns with those derived from the test set ([0-1000] range on Table 1). The bias for T and S exhibits a larger magnitude relative to the test set across each crossing, with variations between positive and negative biases. One possible explanation for these variations is related to the temporal and spatial resolution of satellite observations, particularly of ADT. These factors may contribute to a consistent directional bias in the model's predictions.

The R^2 values range from 0.996 to 0.998 for T predictions, and from 0.988 to 0.994 for S predictions, meaning NeSPReSO consistently captures around 99% of the T and S variances.

Crossing	T RMSE	T Bias	T R^2	S RMSE	S Bias	S R^2
Mission 0006, crossing #1	0.546	0.070	0.997	0.096	-0.006	0.988
Mission 0006, crossing #2	0.516	-0.119	0.998	0.094	-0.025	0.990
Mission 0010	0.544	0.121	0.996	0.072	0.020	0.992
Mission 0012	0.586	0.003	0.997	0.086	-0.035	0.994

Table 2: RMSE, bias and R^2 between observations and synthetics across mesoscale eddy crossings.

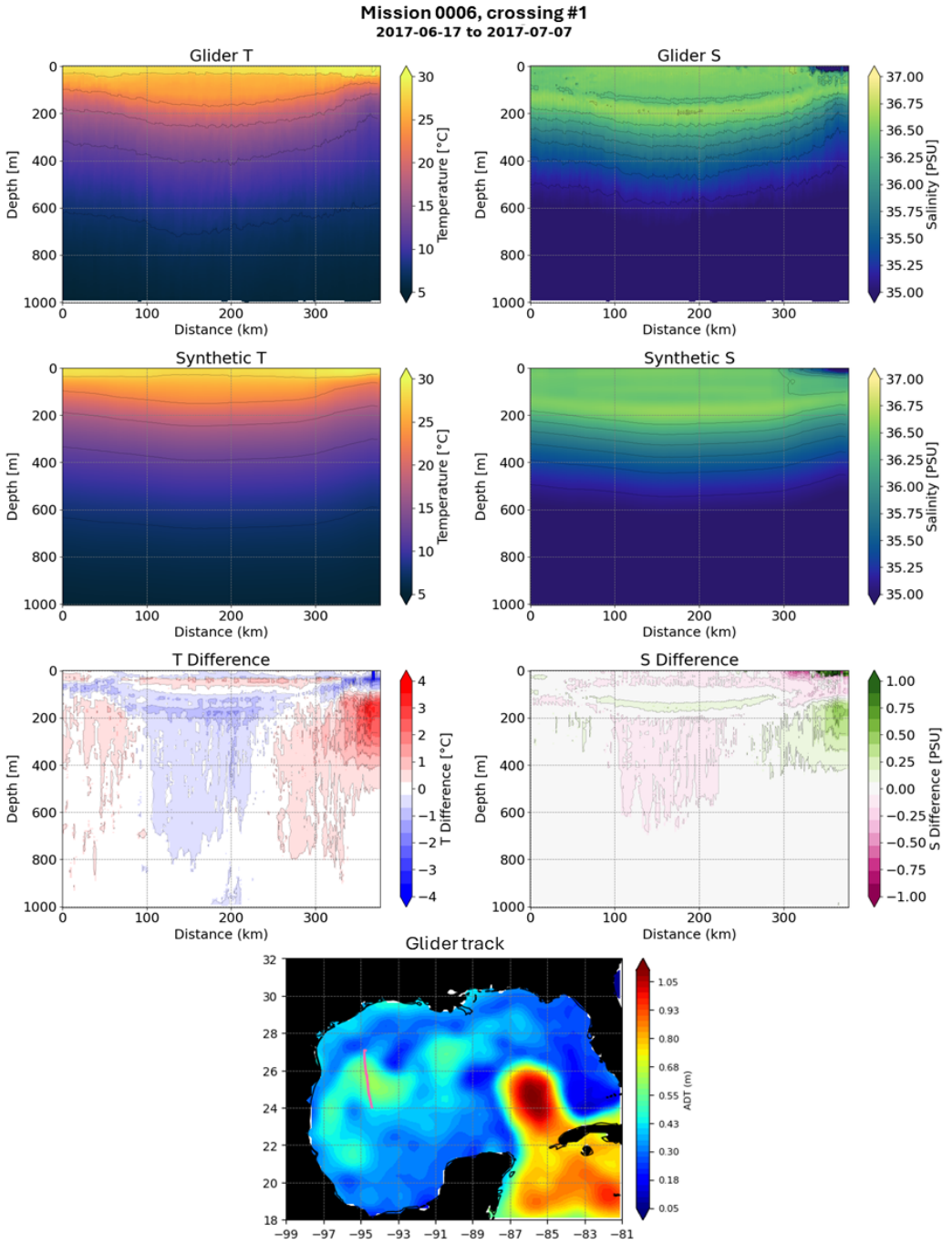


Figure 9: Temperature and salinity sections of mission 0006, crossing #1. First column: Temperature. Second column: Salinity. First row: processed data from glider. Second row: synthetic profiles using NeSPReSO. Third row: differences. Last row: ADT field and position of the glider track.

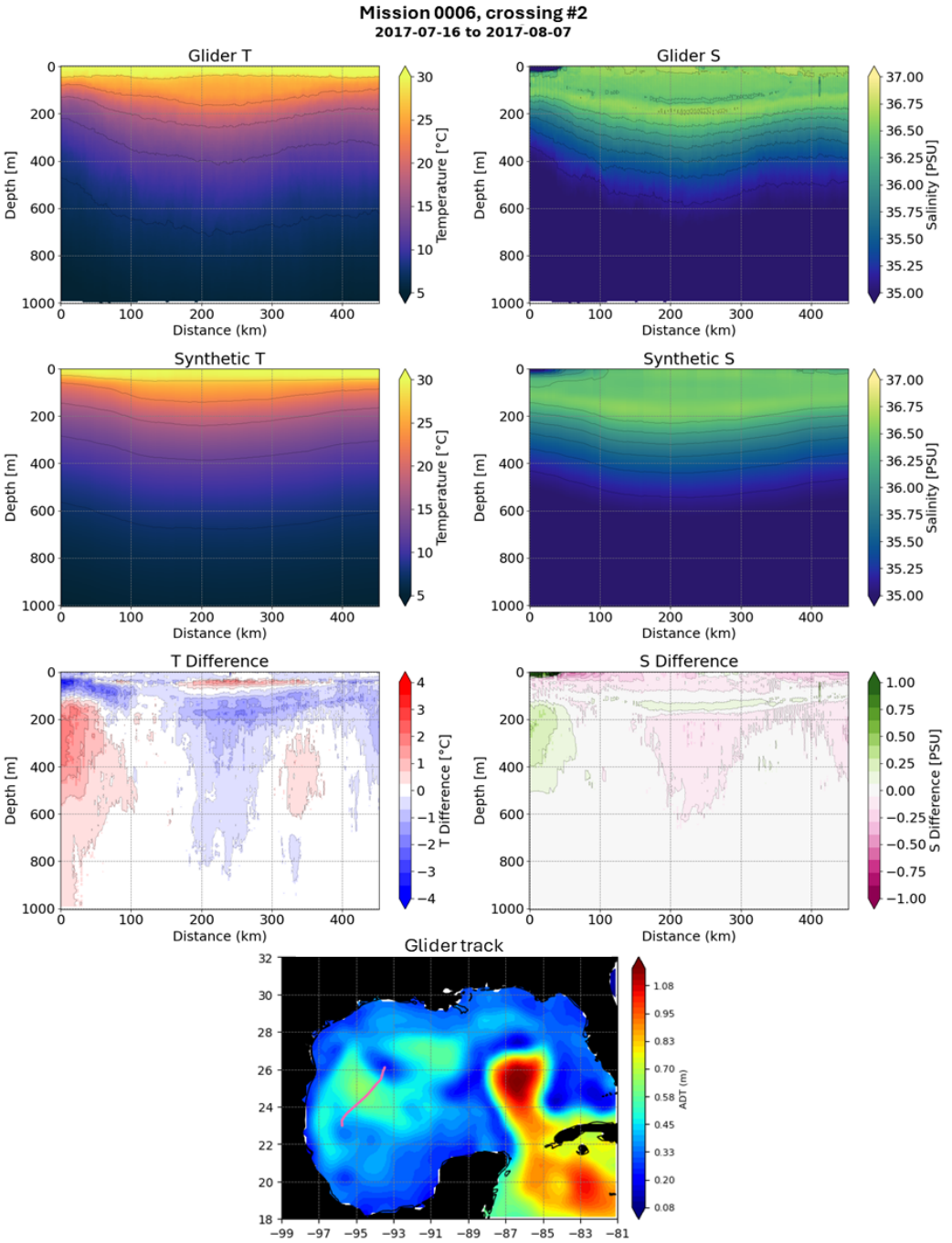


Figure 10: Temperature and salinity sections of mission 0006, crossing #2. First column: Temperature. Second column: Salinity. First row: processed data from glider. Second row: synthetic profiles using NeSPReSO. Third row: differences. Last row: ADT field and position of the glider track.

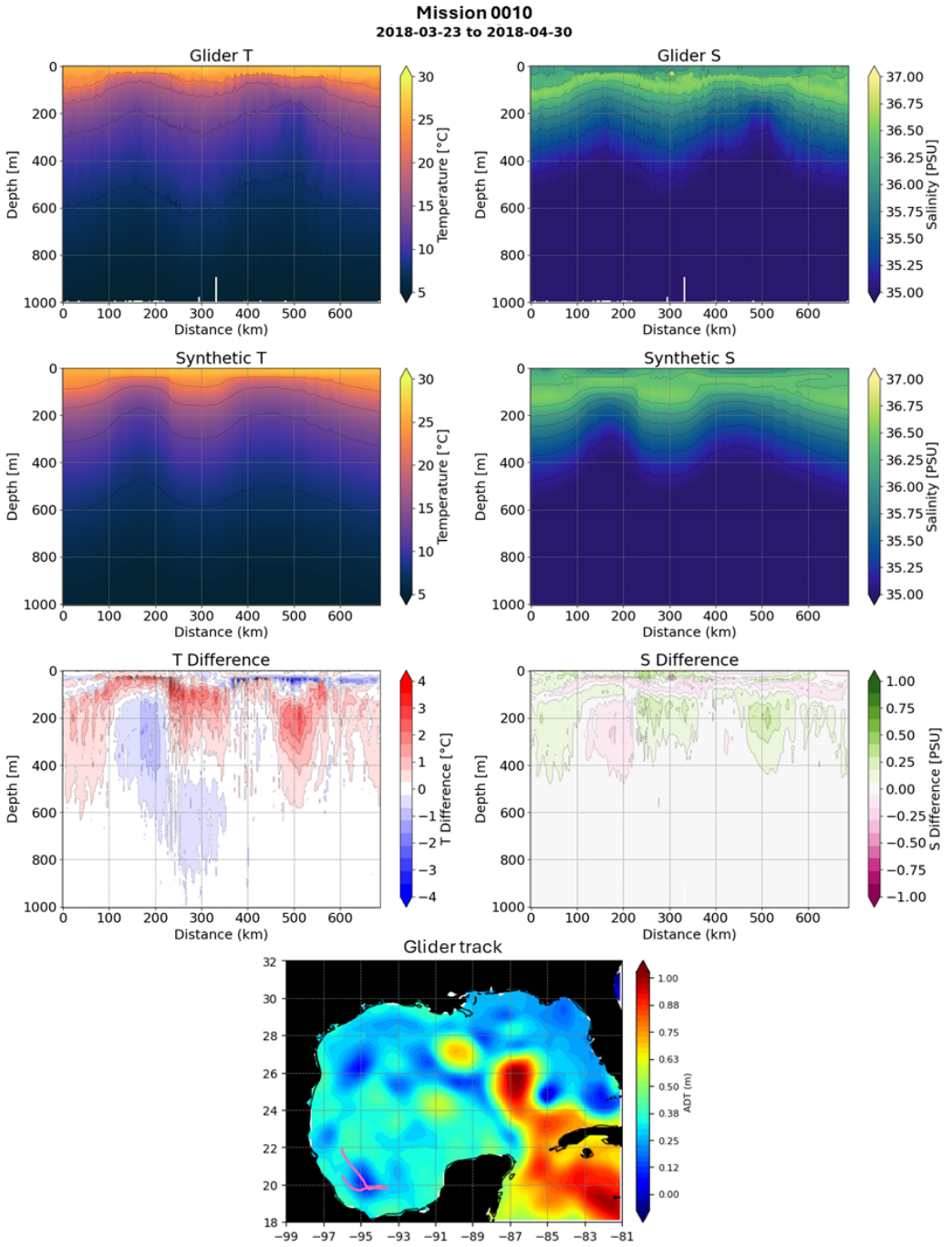


Figure 11: Temperature and salinity sections of mission 0010. First column: Temperature. Second column: Salinity. First row: processed data from glider. Second row: synthetic profiles using NeSPReSO. Third row: differences. Last row: ADT field and position of the glider track.

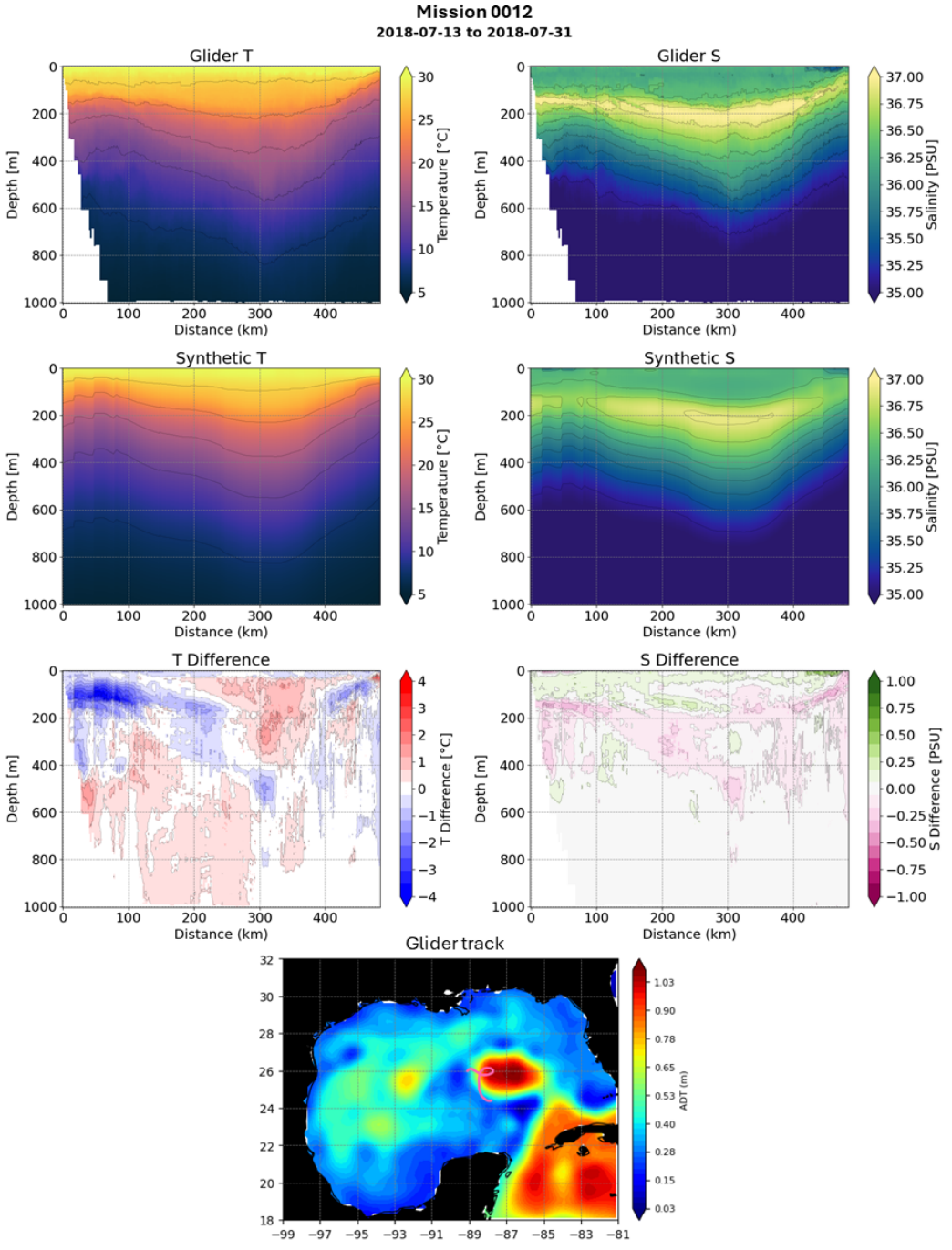


Figure 12: Temperature and salinity sections of mission 0012. First column: Temperature. Second column: Salinity. First row: processed data from glider. Second row: synthetic profiles using NeSPReSO. Third row: differences. Last row: ADT field and position of the glider track.

5. Conclusions

This study underscores the efficacy of machine learning in producing synthetic temperature and salinity profiles for oceanographic data. By integrating Principal Component Analysis (PCA) with neural network models, we successfully generated subsurface profiles from surface data, surpassing traditional methods like MLR, GEM and ISOP in accuracy and reliability.

Our results indicate that the neural network model consistently outperforms other investigated methods in terms of average RMSE, bias, and R^2, suggesting a more accurate representation of the temperature and salinity profiles in the Gulf of Mexico. This improvement is notable given the complex, nonlinear relationships between surface and subsurface properties of the ocean, which machine learning models are particularly adept at capturing.

These results raises several questions that warrant further investigation. For instance, how will NeSPReSO perform in different oceanic regions with distinct hydrodynamic and thermohaline characteristics, and what adaptations might be required for different regional applications? Also, how can NeSPReSO be adapted and trained to effectively generate accurate temperature and salinity profiles in oceanic regions with depths shallower than the model's current maximum depth range?

Future work should focus on addressing these questions, perhaps exploring other machine learning techniques or hybrid models that combine the strengths of various approaches. With the UGOS3 autonomous profiling floats fleet projected to accumulate approximately 1500 profiles annually, the expanding dataset will significantly enhance the model's training and refinement. This expansion is crucial for extending the model's applicability across different oceanic areas, enriching our comprehension of its potential and constraints.

In conclusion, this work lays a precedent for using advanced machine learning methods in oceanographic data synthesis, offering a promising direction for future research in this field. The ability to accurately predict subsurface oceanographic profiles using surface data not only aids in understanding ocean dynamics but also has practical implications in weather forecasting, climate modeling, and resource exploration.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT (3.5, 4, and 4o) in order to improve grammar, clarity and coherence. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Behringer, D.W., Molinari, R.L., Festa, J.F., 1977. The variability of anticyclonic current patterns in the gulf of mexico. Journal of Geophysical Research (1896-1977) 82, 5469–5476. doi:10.1029/JC082i034p05469.

Carnes, M., Teague, W., Mitchell, J., 1994. Inference of subsurface thermohaline structure from fields measurable by satellite. Journal of Atmospheric and Oceanic Technology 11, 551–566. URL: https://journals.ametsoc.org/view/journals/atot/11/2/1520-0426_1994_011_0551_iostsf_2_0_co_2.xml.

Chen, Z., Wang, P., Bao, S., Zhang, W., 2022. Rapid reconstruction of temperature and salinity fields based on machine learning and the assimilation application. Frontiers in Marine Science 9, 985048. doi:10.3389/fmars.2022.985048.

Copernicus Marine Service, 2024. Global ocean gridded l4 sea surface heights and derived variables nrt. URL: https://doi.org/10.48670/moi-00149.

Dubey, S.R., Singh, S.K., Chaudhuri, B.B., 2022. Activation functions in deep learning: A comprehensive survey and benchmark URL: http://arxiv.org/abs/2109.14545.

Forristall, G.Z., Schaudt, K.J., Cooper, C.K., 1992. Evolution and kinematics of a loop current eddy in the gulf of mexico during 1985. Journal of Geophysical Research: Oceans 97, 2173–2184. doi:https://doi.org/10.1029/91JC02905.

Fu, Z., Hu, L., Chen, Z., Zhang, F., Shi, Z., Hu, B., Du, Z., Liu, R., 2020. Estimating spatial and temporal variation in ocean surface pco2 in the gulf of mexico using remote sensing and machine learning techniques. Science of The Total Environment 745, 140965. doi:https://doi.org/10.1016/j.scitotenv.2020.140965.

Good, S., Fiedler, E., Mao, C., Martin, M.J., Maycock, A., Reid, R., Roberts-Jones, J., Searle, T., Waters, J., While, J., Worsfold, M., 2020. The current configuration of the ostia system for operational production of foundation sea surface temperature and ice concentration analyses. Remote Sensing 12, 720. doi:10.3390/rs12040720.

Helber, R.W., Smith, S.R., Jacobs, G.A., Barron, C.N., Carrier, M.J., Yaremchuk, M., Rowley, C.D., Ngodock, H.E., Bartels, B.P., Pasmans, I., et al., 2022. Velocity Assimilation with Improved Synthetic Ocean Profiles (ISOP2): Validation Test Report.

Helber, R.W., Townsend, T.L., Barron, C.N., Dastugue, J.M., Carnes, M.R., 2013. Validation test report for the Improved Synthetic Ocean Profile (ISOP) system, Part I: Synthetic profile methods and algorithm. Naval Res. Lab. Rep. NRL/MR/7320-13-9364 .

Hiron, L., de la Cruz, B.J., Shay, L.K., 2020. Evidence of loop current frontal eddy intensification through local linear and nonlinear interactions with the loop current. Journal of Geophysical Research: Oceans 125, e2019JC015533. doi:https://doi.org/10.1029/2019JC015533. e2019JC015533 10.1029/2019JC015533.

Hiron, L., Miron, P., Shay, L.K., Johns, W.E., Chassignet, E.P., Bozec, A., 2022. Lagrangian coherence and source of water of loop current frontal eddies in the gulf of mexico. Progress in Oceanography 208, 102876. doi:https://doi.org/10.1016/j.pocean.2022.102876.

Hiron, L., Nolan, D.S., Shay, L.K., 2021. Study of ageostrophy during strong, nonlinear eddy-front interaction in the gulf of mexico. Journal of Physical Oceanography 51, 745 – 755. doi:10.1175/JPO-D-20-0182.1.

Howley, T., Madden, M.G., O'Connell, M.L., Ryder, A.G., 2006. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data, in: Macintosh, A., Ellis, R., Allen, T. (Eds.), Applications and Innovations in Intelligent Systems XIII. Springer, London, pp. 209–222. doi:10.1007/1-84628-224-1_16.

Jaimes, B., Shay, L.K., Brewster, J.K., 2016. Observed air-sea interactions in tropical cyclone isaac over loop current mesoscale eddy features. Dynamics of Atmospheres and Oceans 76, 306–324. doi:https://doi.org/10.1016/j.dynatmoce.2016.03.001.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374. doi:10.1098/rsta.2015.0202.

Koch, S., Barker, J., Vermersch, J., 1991. The Gulf of Mexico Loop Current and Deepwater Drilling. Journal of Petroleum Technology 43, 1046–1119. doi:10.2118/20434-PA.

Leben, R.R., 2005. Altimeter-Derived Loop Current Metrics. American Geophysical Union (AGU). pp. 181–201. doi:https://doi.org/10.1029/161GM15.

Liu, H., Zhou, H., Yang, W., Liu, X., Li, Y., Yang, Y., Chen, X., Li, X., 2021. A three-dimensional gravest empirical mode determined from hydrographic observations in the western equatorial pacific ocean. Journal of Marine Systems 214, 103487. doi:https://doi.org/10.1016/j.jmarsys.2020.103487.

Lueck, R., Picklo, J., 1990. Thermal inertia of conductivity cells: Observations with a sea-bird cell. Journal of Atmospheric and Ocean Technology 7, 756–768.

Mafi, S., Amirinia, G., 2017. Forecasting hurricane wave height in gulf of mexico using soft computing methods. Ocean Engineering 146, 352–362. doi:https://doi.org/10.1016/j.oceaneng.2017.10.003.

Mao, K., Liu, C., Zhang, S., Gao, F., 2023. Reconstructing ocean subsurface temperature and salinity from sea surface information based on dual path convolutional neural networks. Journal of Marine Science and Engineering 11, 1030. doi:10.3390/jmse11051030.

Meissner, T., Wentz, F.J., Le Vine, D.M., 2018. The salinity retrieval algorithms for the nasa aquarius version 5 and smap version 3 releases. Remote Sensing 10, 1121. doi:10.3390/rs10071121.

Meng, L., Yan, C., Zhuang, W., Zhang, W., Yan, X.H., 2021. Reconstruction of three dimensional temperature and salinity fields from satellite observations. Journal of Geophysical Research: Oceans 126, e2021JC017605.

Meunier, T., Bower, A., Pérez-Brunius, P., Graef, F., Mahadevan, A., 2024. The Energy Decay of Warm-Core Eddies in the Gulf of Mexico. Geophysical Research Letters 51. doi:10.1029/2023GL106246.

Meunier, T., Le Boyer, A., Molodtsov, S., Bower, A., Furey, H., Robbins, P., 2023. Internal wave activity in the deep Gulf of Mexico. Frontiers in Marine Science 10. doi:10.3389/fmars.2023.1285303.

Meunier, T., Pérez-Brunius, P., Bower, A., 2022. Reconstructing the three-dimensional structure of loop current rings from satellite altimetry and in situ data using the gravest empirical modes method. Remote Sensing 14, 4174.

National Academies of Sciences, Engineering, and Medicine, 2018. Understanding and Predicting the Gulf of Mexico Loop Current: Critical Gaps and Recommendations. The National Academies Press, Washington, DC. doi:10.17226/24823.

Nguyen, A., Pham, K., Ngo, D., Ngo, T., Pham, L., 2021. An analysis of state-of-the-art activation functions for supervised deep neural network URL: http://arxiv.org/abs/2104.02523.

Pauthenet, E., Bachelot, L., Balem, K., Maze, G., Tréguier, A.M., Roquet, F., Fablet, R., Tandeo, P., 2022. Four-dimensional temperature, salinity and mixed-layer depth in the gulf stream, reconstructed from remote-sensing and in situ observations with neural networks. Ocean Science 18, 1221–1244. doi:10.5194/os-18-1221-2022.

Preisendorfer, R.W., Mobley, C.D., 2023. Principal component analysis in meteorology and oceanography. SERBIULA (sistema Librum 2.0). Posthumously compiled and edited by Curtis D. Mobley.

Roemmich, D., Gilson, J., 2009. The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the argo program. Progress in oceanography 82, 81–100.

Sarıkoç, M., Celik, M., 2024. Pca-ica-lstm: A hybrid deep learning model based on dimension reduction methods to predict s&p 500 index price. Computational Economics doi:10.1007/s10614-024-10629-x.

Shay, L.K., 2010. Air-Sea Interactions in Tropical Cyclones. pp. 93–131. doi:10.1142/9789814293488_0003.

Shay, L.K., Uhlhorn, E.W., 2008. Loop current response to hurricanes isidore and lili. Monthly Weather Review 136, 3248 – 3274. doi:10.1175/2007MWR2169.1.

Sturges, W., Leben, R., 2000. Frequency of ring separations from the loop current in the gulf of mexico: A revised estimate. Journal of Physical Oceanography 30, 1814–1819. doi:10.1175/1520-0485(2000)030<1814:FORSFT>2.0.CO;2.

Sturges, W., Lugo-Fernandez, A., Shargel, M.D., 2005. Introduction to Circulation in the Gulf of Mexico. American Geophysical Union (AGU). doi:https://doi.org/10.1029/161GM02.

Sun, C., Watts, D.R., 2001. A circumpolar gravest empirical mode for the southern ocean hydrography. Journal of Geophysical Research: Oceans 106, 2833–2855.

Sun, Y., Zhou, S., Meng, S., Wang, M., Mu, H., 2023. Principal component analysis–artificial neural network-based model for predicting the static strength of seasonally frozen soils. Scientific Reports 13. doi:10.1038/s41598-023-43462-7.

Suthers, I.M., Schaeffer, A., Archer, M., Roughan, M., Griffin, D.A., Chapman, C.C., Sloyan, B.M., Everett, J.D., 2023. Frontal eddies provide an oceanographic triad for favorable larval fish habitat. Limnology and Oceanography 68, 1019–1036. doi:https://doi.org/10.1002/lno.12326.

Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B.B., Rashidi, P., Pardalos, P., Momcilovic, P., Bihorac, A., 2016. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. PLOS ONE 11. doi:10.1371/journal.pone.0155705.

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

Townsend, T., Barron, C., Helber, R., 2015. Ocean prediction with improved synthetic ocean profiles (isop). 2015 NRL Review .

Vukovich, F.M., 1988. Loop current boundary variations. Journal of Geophysical Research: Oceans 93, 15585–15591. doi:10.1029/JC093iC12p15585.

Wang, J.L., Zhuang, H., Chérubin, L.M., Ibrahim, A.K., Muhamed Ali, A., 2019. Medium-term forecasting of loop current eddy cameron and eddy darwin formation in the gulf of mexico with a divide-and-conquer machine learning approach. Journal of Geophysical Research: Oceans 124, 5586–5606. doi:https://doi.org/10.1029/2019JC015172.

Watts, D.R., Sun, C., Rintoul, S., 2001. A two-dimensional gravest empirical mode determined from hydrographic observations in the subantarctic front. Journal of Physical Oceanography 31, 2186–2209.

Zeng, X., Li, Y., He, R., 2015. Predictability of the loop current variation and eddy shedding process in the gulf of mexico using an artificial neural network approach. Journal of Atmospheric and Oceanic Technology 32, 1098 – 1111. doi:10.1175/JTECH-D-14-00176.1.

Zhang, A., Lipton, Z.C., Li, M., Smola, A.J., 2024. Multilayer perceptrons, in: Dive into Deep Learning. URL: https://d2l.ai/chapter_multilayer-perceptrons/mlp.html. retrieved October 8, 2024.

Zhang, Y., Yang, Q., 2017. A survey on multi-task learning. arXiv preprint arXiv:1707.08114 .

Highlights

Neural Synthetic Profiles from Remote Sensing and Observations (NeSPReSO) - Reconstructing temperature and salinity fields in the Gulf of Mexico.

Jose R. Miranda, Olmo Zavala-Romero, Luna Hiron, Eric P. Chassignet, Bulusu Subrahmanyam, Thomas Meunier, Robert W. Helber, Enric Pallas-Sanz, Miguel Tenreiro

- NeSPReSO uses Argo and satellite data to generate accurate T and S profiles.
- Experiments shows NeSPReSO outperforms MLR, GEM and ISOP in the GoM.
- Synthetic profiles will be used to improve ocean models through data assimilation.

Neural Synthetic Profiles from Remote Sensing and Observations (NeSPReSO) - Reconstructing temperature and salinity fields in the Gulf of Mexico.

Jose R. Miranda[a,b], Olmo Zavala-Romero[a,b], Luna Hiron[b], Eric P. Chassignet[b], Bulusu Subrahmanyam[c], Thomas Meunier[d,e], Robert W. Helber[f], Enric Pallas-Sanz[g], Miguel Tenreiro[g]

[a]*Department of Scientific Computing, Florida State University, Tallahassee, FL, United States*
[b]*Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, FL, United States*
[c]*School of the Earth, Ocean, and Environment, University of South Carolina, Columbia, FL, United States*
[d]*Woods Hole Oceanographic Institution, Woods Hole, MS, United States*
[e]*Laboratoire d'Océanographie Physique et Spatiale, Ifremer / UBO / CNRS / IRD, Plouzané, Brittany, France*
[f]*Ocean Sciences Division, US Naval Research Laboratory, Hancock County, MS, United States*
[g]*Ensenada Center for Scientific Research and Higher Education, Ensenada, BC, Mexico*

Keywords: Synthetic temperature and salinity, machine learning, Gulf of Mexico, Loop Current, Data Assimilation

1. Introduction

Accurate representation of the Gulf of Mexico (GoM) circulation in numerical models is of great importance for the scientific community and holds operational significance for fisheries, hurricane prediction, and oil and gas companies ~~(Jaimes et al. (2016); Koch et al. (1991); National Academies of Sciences, Engineering,)~~(Jaimes et al., 2016; Koch et al., 1991; National Academies of Sciences, Engineering, and Medici . The GoM Loop Current (LC) is part of the Atlantic western boundary current system and plays an important role in the transport of heat from the Caribbean Sea to the Atlantic Ocean, contributing to climate regulation. The LC also holds strong currents (up to 2 ms^{-1}~~; Forristall et al. (1992); Sturges et al. (2005); Hiron et al. (2021))~~) (Forristall et al., 1992; Sturges et al., 2005; Hiron et al.

and is very dynamic, shedding large (≈200-400 km) warm eddies at an irregular rate of 6 to 17 months ~~(Vukovich (1988); Behringer et al. (1977); Sturges and Leben (2000))~~(Vukovich, 1988; Behringer et al., 1977; Sturges and Leben, 2000) . Loop Current Eddies (LCE) affect oil and gas activities in the GoM due to their strong peripheral velocities, and they can also fuel hurricane intensification by releasing heat to the atmosphere during storm passage ~~(Shay and Uhlhorn (2008); Shay (2010); Jaimes et al. (2016))~~(Shay and Uhlhorn, 2008; Shay, 201 . Cold-core, frontal eddies present in the vicinity of the LC contribute to the detachment of the LCEs and can enhance activity across the trophic chain by pumping deep-water nutrients to the upper ocean ~~(Hiron et al. (2020); Hiron et al. (2022); Suthers et al. (2023))~~(Hiron et al., 2020, 2022; Suthers et al., 2023) . Although recent model advancements have improved the representation of this complex system, a key limitation across ocean models remains the scarcity of in situ data to effectively constrain the models.

Temperature and salinity observations are two essential variables to be assimilated in numerical models, as density gradients, driven by these variables and pressure, govern large-scale ocean circulation. The ocean surface is well constrained in models, thanks to global satellite-derived sea surface height (SSH) and sea surface temperature (SST) data. However, subsurface observations are scarcer. The Argo program supports almost 4,000 floats worldwide that provide valuable information about the subsurface temperature and salinity structure of the ocean since 2005 ~~(Roemmich and Gilson (2009))~~(Roemmich and Gilson, 2009). In the GoM, the NAS-funded LC-floats and the UGOS 3 program are significant initiatives in subsurface observation. The LC-floats, supported by the National Academy of Sciences (NAS), are designed for oceanographic research in the GoM. Since June 2019, these floats have played a key role in collecting data on subsurface temperature and salinity structures. The UGOS 3 program, focusing on the GoM region, involves specialized floats that have contributed to more than 7,000 profiles sampled since the same period.

Despite their significance in constraining subsurface models, these measurements are too sparse, limiting the accurate representation of subsurface mesoscale circulation. ~~Recent techniques, such as the~~ Techniques such as Multiple Linear Regression (Carnes et al., 1994) Gravest Empirical Modes (GEM) method ~~(Watts et al. (2001); Sun and Watts (2001); Meunier et al. (2022))~~(Watts et al., 2001; Sun and Watts, 2001; Meunier et al., 2022) and the Improved Synthetic Ocean Profile (ISOP) system ~~(Helber et al. (2013); Townsend et al. (2015) ; Helber et al. (2022)),~~(Helber et al., 2013; Townsend et al., 2015; Helber et al., 2022)

have been employed to generate synthetic temperature and salinity profiles for data assimilation in large-scale and regional ocean models. Those synthetic profiles rely on past observations and are generated mainly from altimetry SSH fields, based on the presumed relationship between SSH values and subsurface temperature and salinity, valid for large-scale flows (geostrophic adjustment). Although promising, these methods can be computationally demanding and may not capture complex, non-linear relationships between surface and subsurface ocean fields.

~~There are previous studies using multi-regressions between surface information (SST and SSH) and temperature and salinity at depths (e. g. , the Navy's Modular Ocean Data Assimilation System (MODAS) , Fox et al. (2002)). Machine learning, in particular,~~

In recent years, there has been significant advancement in deriving temperature and salinity profiles from ocean surface data using machine learning (ML) and artificial intelligence (AI) approaches. These models aim to bridge the gap between sparse in-situ measurements and satellite observations, enabling more comprehensive ocean monitoring. For instance, Chen et al. (2022) developed a machine learning-based assimilation system that uses a generalized regression neural network with fruit fly optimization to reconstruct T/S profiles from satellite observations, significantly improving the simulation of subsurface structures compared to direct assimilation of satellite data alone. Similarly, Tian et al. (2022) employed a feed-forward neural network to generate a high-resolution (0.25° x 0.25°) global subsurface salinity dataset by merging in-situ profiles with satellite altimetry, sea surface temperature, and wind data. Mao et al. (2023) developed a dual-path convolutional neural network to reconstruct ocean subsurface temperature and salinity from sea surface information, demonstrating improved accuracy over traditional methods. Pauthenet et al. (2022) reconstructed four-dimensional temperature, salinity, and mixed-layer depth in the Gulf Stream using neural networks, combining remote-sensing and in situ observations. These AI-based methods have shown promise in capturing mesoscale features and improving upon traditional interpolation techniques, offering new possibilities for generating comprehensive ocean T/S datasets with enhanced spatial and temporal resolution.

In the Gulf of Mexico, machine learning has been used in ~~the GoM in~~ numerous applications, such as forecasting LCE shedding events ~~(Zeng et al. (2015) ; Wang et al. (2019)~~(Zeng et al., 2015; Wang et al., 2019)), predicting hurricane wave height ~~(Mafi and Amirinia (2017))~~(Mafi and Amirinia, 2017), and estimating spatial and temporal variation in dissolved carbon dioxide near

the Mississippi river outflow ~~(Fu et al. (2020))~~(Fu et al., 2020). Meng et al. (2021) developed a ~~CNN (~~convolutional neural network (CNN) method using satellite-observed sea surface data (SSH, SST, sea surface salinity (SSS), and surface wind speed) and ~~the~~ ocean subsurface temperature and salinity from Argo to obtain ~~the~~ three-dimensional salinity fields from 0-2000 m depth. ~~Research~~ Despite these advancements, research with ML for subsurface modeling in the Gulf of Mexico is ongoing, as traditional methods still face challenges in efficiency, accuracy~~and on~~ , and capturing the complex dynamics of the Gulf's circulation, ~~specially~~ especially at submesoscale.

In this study, we introduce NeSPReSO (Neural Synthetic Profiles from Remote Sensing and Observations), a method to effectively estimate subsurface temperature and salinity profiles using satellite-derived absolute dynamic topography (ADT), SST, and SSS by leveraging in-situ Argo data ~~.~~ ~~The use of a ML model can offer multiple advantages compared to traditional methods: reduction of the computational cost of an operational model, accounting for nonlinearities, and inclusion of additional fields, such as SST, SSS, time, and location~~and Principal Component Analysis (PCA). Unlike previous methods, NeSPReSO focuses specifically on the Gulf of Mexico, utilizing a neural network architecture optimized for this region's oceanographic features. Our approach advances the field by combining PCA to reduce the dimensionality of the T/S profiles while capturing most of their variability, and a neural network that maps surface observations to these principal components. This methodology allows for efficient computation while capturing the complex, non-linear relationships between surface and subsurface ocean fields, thereby improving upon traditional methods and previous ML approaches in terms of accuracy and computational cost.

This study aims to address the following questions: How effectively can ~~machine learning~~ ML techniques, specifically neural networks (NN), be utilized to synthesize temperature and salinity profiles in the Gulf of Mexico? Can NeSPReSO provide an improvement over ~~state of the art~~ state-of-the-art methods? How do these synthetic profiles compare against independent measurements? ~~Application~~ Applications of this study include ~~the implementation of the machine-learning-based approach developed here to assimilate synthetic subsurface ,~~ investigating the effects of assimilating the synthetic subsurface temperature and salinity profiles into hindcast and forecast numerical models in the Gulf of Mexico ~~for science and operational use~~to determine whether they improve forecast accuracy. Additionally, we plan to provide a system through which the scientific community can request synthetic profiles for

specific locations and time periods (depending on satellite data availability) to foster further research and applications.

2. Data

~~This study integrates diverse data sources, encompassing~~ Our ML approach builds upon in situ observations and satellite-derived measurements~~, from which our ML approach builds upon~~. The following subsections details the specifics of each dataset, specifically Argo float, glider, and satellite datasets, as well as the ISOP statistics used as benchmark.

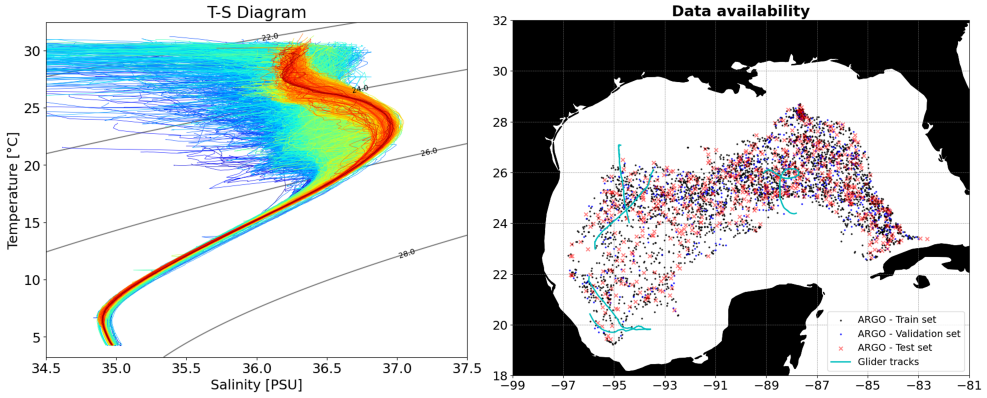
2.1. Argo Data

The main dataset for this study is a total of 4,145 temperature (T) and salinity (S) profiles acquired between 2015 and 2022 in the GoM region, and includes geographical coordinates, date, and time, as well as the estimated local steric height referenced to 1,950 dbar (SH1950) for each profile. The distribution of these profiles is shown in Figure 1. ~~Each profile provides~~ T and S measurements were taken at one-meter intervals from the ~~ocean surface up~~ surface to a depth of 2,000 meters~~. Both~~, capturing both major upper-ocean water masses present in the GoM~~are sampled~~: the warm and salty North Atlantic Subtropical Underwater (NASUW), ~~characteristic of the LC~~ typical of the Loop Current (SH1950 $\geq$ 0.17 m), and the fresher Gulf Common Water (GCW), ~~characteristic of~~ representative of the Gulf waters (SH1950 $<$ 0.17 m) (e.g., Hiron et al. (2022)).

The dataset, described in detail by Meunier et al. (2022, 2023, 2024), includes a mixture of real-time and delayed mode profiles, re-processed without using the standard quality control (QC) flags. Outliers, defined as values outside four standard deviations, were removed, as well as profiles showing biased salinity at depth. Although these profiles could potentially be recovered with further processing, they were excluded from this analysis to maintain data consistency.

~~The datasetincludes geographical coordinates, date, and time, as well as the estimated local steric height referenced to 1, 950 dbar (SH1950) for each profile.~~

ISOP statistics are limited to the 0 to 1,800-meter range. Given that our Argo database has missing data beyond 1,800 meters, we restricted our dataset for model training, testing, and validation to this range.



~~Temperature-Salinity diagram (left) and Spatial distribution (right) of the Argo profiles used in this study. The core of the Gulf Common water (GCW), North Atlantic Subtropical Underwater (NASUW) and Sub-Antarctic Intermediate water (SAAIW) are marked for reference.~~

Figure 1: Temperature-Salinity (T-S) diagram (left) and spatial distribution (right) of glider tracks and Argo profiles used in this study. The T-S diagram identifies key water masses, including Gulf Common Water (GCW), North Atlantic Subtropical Underwater (NASUW), and Sub-Antarctic Intermediate Water (SAAIW). The spatial distribution uses markers/colors to represent dataset categories (train, validation, and test).

2.2. Glider dataset

This dataset comprises T and S profiles from ~~four missions conducted between August 2016 and October 2018. These missions targeted~~ three missions (0006, 0010, and 0012) conducted between June 2017 and October 2018, targeting various mesoscale structures within the Gulf of Mexico ~~,~~ ~~including two crossings each of the LCE Poseidon, the~~ by the glider oceanographic monitoring group (GMOG) from Cicese. These missions, executed using Seagliders equipped with a Seabird free-flow CT-sail, aimed to capture the vertical thermohaline variability associated with these mesoscale features. Data were collected at an averaged vertical resolution of 1 m and horizontal resolution of 3 km.

Missions 0006 and 0012 sampled old and young LCEs, respectively, and mission 0010 targeted a cyclonic eddy in Campeche Bay~~, and one crossing of a notably intense LCE.~~

~~Glider data was initially recorded at 0.5m depth intervalsand up to 1, 000m. It underwent vertical binning at 5m intervals during~~ . During post-processing~~, and a~~ , data was vertically binned at 5 m intervals, and temperature

adjusted for thermal lag, while thermal-inertia effects on conductivity were corrected following the methodology of Lueck and Picklo (1990). A fourth-order low-pass Butterworth filter with a cut-off frequency of $\frac{1}{48}h^{-1}$ was applied to smooth out high-frequency, near-inertial gravity waves. ~~Segments of missing data~~ Missing segments were linearly interpolated to maintain the integrity of the profiles.

The gliders sampled contrasting thermohaline structures ~~which are pivotal in~~ critical for assessing the reconstruction algorithm's proficiency. ~~The dataset reveals notable~~ Significant differences in salinity ~~and temperature anomalies between the observed eddies, crucial for validating the synthetic profile reconstructions~~($\Delta S = 0.2$) and temperature ($\Delta T = 2°C$) anomalies were observed between the eddies, with variations in the depth of the 26°C isotherm between young and old LCEs indicative of the effect of eddy age on thermohaline structure. However, ~~significant~~ large discrepancies are anticipated at the ~~eddies' peripheries due to the influence of subesoscale~~ peripheries of the eddies due to submesoscale processes like density-compensated T and S layering and intrusions, which are not captured by the satellite fields, challenging the model's predictive capability in these areas.

2.3. Satellite data

Satellite-derived ~~ADT, SST, and SSS data~~ Absolute Dynamic Topography (ADT), sea surface temperature (SST) and salinity (SSS) were sourced from CMEMS, OISST, and SMAP, respectively. The Copernicus Marine Environment Monitoring Service (CMEMS) archives, validates, and interprets oceanographic satellite data. We utilized ~~Absolute Dynamic Topography (ADT)~~ADT, available since 1993, serving as a proxy for SSH. CMEMS provides an ADT gridded product with a daily resolution and a horizontal grid-spacing of approximately $\frac{1}{4}$ degrees (Copernicus Marine Service, 2024).

Optimum Interpolation Sea Surface Temperature (OISST) is a long-term climate data record that incorporates observations from different sources to provide a high-resolution analysis of sea surface temperatures. It uses an optimal interpolation technique to combine data from satellites, ships, buoys, and other sources to create a consistent and accurate record of sea surface temperatures. Analysed SST is available since 1981 on a daily basis, with a resolution of approximately $\frac{1}{4}$ degrees (Good et al., 2020).

Finally, SMAP, or "Soil Moisture Active Passive", is a NASA satellite mission that uses active and passive microwave sensors to provide high-resolution measurements of soil moisture, freeze/thaw state, and ocean surface salinity.

SMAP SSS has been available since 2015 on a daily basis and has a resolution of 40 km (Meissner et al., 2018).

The ADT, SST, and SSS fields are interpolated ~~into~~ to each location of the Argo and glider databases using bicubic interpolation, and together with spatial and temporal information, serve as input to the proposed neural network as described in Section 3.2. Following Leben (2005) and Hiron et al. (2020), the daily mean of ADT over the GoM deep waters (> 200 m) is removed from the ADT field for each day. This removes the variations in ADT associated with thermal expansion and contraction of the upper ocean due to seasonal variability.

2.4. ISOP statistics

ISOP projects surface ocean data downward, generating T and S profiles across the global ocean using surface observations and a mixed-layer depth (MLD) estimate. Optionally, a prior forecast of T and S profiles can be used. The creation of these synthetic profiles plays an important step in the Navy's operational forecasting and is seamlessly integrated into their data assimilation workflows. ISOP divides the ocean's depth into 78 fixed levels, extending from the surface to 6600 meters. The process begins with the compilation of a T and S covariance matrix and climatology database from a comprehensive set of in-situ observations, followed by the application of a multilayered approach that considers three different dynamics zones within the ocean subsurface. These regions include the *mixed layer*, extending from the surface to the MLD; the *thermocline layer*, reaching from the MLD down to 1000 meters; and the *deep ocean layer*, below 1000 meters.

For the *mixed layer*, there are two options. One option adjusts the initial estimated profile to align with the surface potential density at 4 meters depth and ensures consistency with the potential density and its gradient at the MLD within the *thermocline layer*. The second option for the *mixed layer* shifts the prior forecast profile (if provided) to match the input SST value. The *thermocline layer* prediction employs a variational method, leveraging climatological T and S values and the first vertical Empirical Orthogonal Functions (EOFs), or modes, extracted from historical data to constrain the forecast. Detailed descriptions of the each term involved in this variational approach is available in reference Helber et al. (2013). Finally, the prediction within the *deep ocean layer* involves modifying a decay function based on climatological data and the T and S readings from the *thermocline layer* at 1000 meters depth. This function accounts for the variance between

climatological values and the 1000-meter predictions, ensuring a coherent transition into the deep ocean predictions. The inputs for ISOP's predictive models include SST and sea surface height anomaly (SSHA), along with uncertainty estimates, an MLD estimation, and an (optional) T and S profile can be obtained from either climatological data or model outputs. In this work, the synthetics used climatological data for estimating the initial MLD and T and S profiles, along with Argo-derived SST and SSH.

The ISOP data used in this work was generated by the US Navy and corresponds to the entire Argo dataset (4,145 profiles). The provided data included only the average vertical statistics and binned spatial statistics of the ISOP synthetics relative to the Argo profiles (no individual profiles were provided). These statistics were used as benchmark for the other methods.

3. Methods

In this section, we detail our methodology for training and validating a multilayer perceptron (MLP) to predict ~~subsurfaceT~~ subsurface T and S profiles using surface data. The model is designed to learn the nonlinear functions that associates the ocean surface, through satellite observations, with subsurface information from a comprehensive dataset of Argo profiles. NeSPReSO uses ~~Principal Component Analysis (PCA)~~ PCA to focus the model on the main variability within the subsurface profiles, while also reducing the data's dimensionality and improving the efficiency of computation and training. Lastly, we assess the model's performance using unseen Argo profiles (15% of the dataset, randomly selected) and compare it with ~~the~~ MLR, GEM and ISOP methods. ~~Four~~ The four unseen glider transects in the GoM were also reconstructed using our method, and compared with the original glider data.

The Argo float dataset, consisting of T and S profiles, is inherently high-dimensional, containing 1801 measurements (from 0 to 1800 meters at 1-meter intervals) for each parameter. In order to obtain an efficient model that captures the overall shape of the profiles, we applied PCA to the data sets of the T and S profiles separately. By doing so, we can express each profile with a significantly reduced number of variables while retaining over 99% of the original data variability. Utilizing this transformation of data, we train the neural network to estimate the 30 most significant principal component scores (PCS) for each profile in the Argo dataset used for training, which are used to reconstruct the profiles using the inverse PCA.

Combining PCA with neural networks is an effective strategy for handling high-dimensional output spaces, as it reduces computational complexity and can improve prediction accuracy (Howley et al., 2006; Sun et al., 2023). PCA captures the most significant features in the data, and the neural network learns to predict these features from the inputs. This methodology has been successfully applied in various fields, including meteorology and oceanography (Preisendorfer and Mobley, 2023), finance (Sarıkoç and Celik, 2024), and engineering (Sun et al., 2023).

Figure 2 shows a general diagram of our methodology and the main components of the proposed neural network.

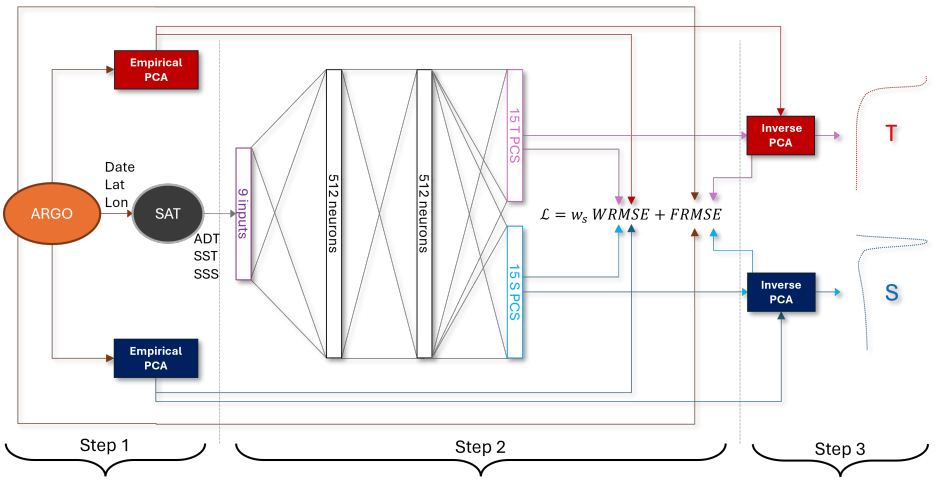


Figure 2: General diagram of NeSPReSO. Step 1 computes the empirical PCA of the Argo database. Step 2 trains a dense neural network from interpolated SST, SSH and SSS satellite data, location and date to predict the PCS. Step 3 reconstruct the profiles using the predicted PCS and inverse PCA.

3.1. Principal Component Analysis

Principal Component Analysis (PCA) is employed in various fields for dimensionality reduction of large datasets while preserving most of the original data variability. This method identifies orthogonal axes, known as principal components (PC), each representing a direction in which the data's variance is maximized.

Given a centered data matrix ~~$\mathbf{Y}$ of size $n \times p$, where n~~ $\mathbf{Y}$ of size $n \times p$, where n is the number of observations (profiles) and ~~p~~ p is the number of variables (measurements).

A covariance matrix ~~S~~ S is computed as:

$$\mathbf{S} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}, \tag{1}$$

which captures the variances (in the diagonal) and the covariances (off-diagonals).

The next step involves solving the eigenvalue problem for ~~S~~S:

$$\mathbf{SV} = \mathbf{DV}, \tag{2}$$

where ~~V and D~~ V and D are the eigenvector matrix and eigenvalue diagonal matrix of ~~S~~S, respectively. These eigenvectors define the directions of maximum variance in the data, and the eigenvalues indicate the magnitude of variance in these directions.

The eigenvectors and eigenvalues are arranged in descending order based on the magnitude of the eigenvalues. The first eigenvector, associated with the largest eigenvalue, becomes the first principal component (PC), and so forth. The eigenvector matrix ~~V~~V, which is the concatenation of all $\mathbf{v_i}$ eigenvectors, is used to project the centered data matrix ~~Y~~ Y into the principal component space:

$$\mathbf{Z} = \mathbf{YV}, \tag{3}$$

where ~~Z~~ Z is a matrix of principal component scores (PCS), each column representing a principal component. To reduce dimensionality, ~~V~~ V can be truncated, keeping only the eigenvectors corresponding to the largest eigenvalues.

The PCA transformation is linear and reversible. The inverse transformation, which approximates the original data from its reduced principal component representation, is given by:

$$\hat{\mathbf{Y}} = \mathbf{ZV}^T, \tag{4}$$

where ~~$\hat{\mathbf{Y}}$~~ $\hat{\mathbf{Y}}$ is the reconstructed data. Note that if ~~V~~ V is truncated, this reconstruction is an approximation with some loss of information.

We applied PCA to the T and S datasets, reducing the dimensionality of the data (from 1801 to 15) by transforming the raw measurements ($\mathbf{Y}$) into PCS ($\mathbf{Z}$), while retaining most of the variance: 99.8% for temperature and 99.4% for salinity. Figure 3 illustrates the first 500 meters of a temperature and salinity profile and its reconstruction using 15 PCS.

Our proposed model is then trained to generate these 30 PCS for each Argo location in our training set. ~~In the next section we describe our NN~~ Next we describe NeSPReSO's architecture and training.

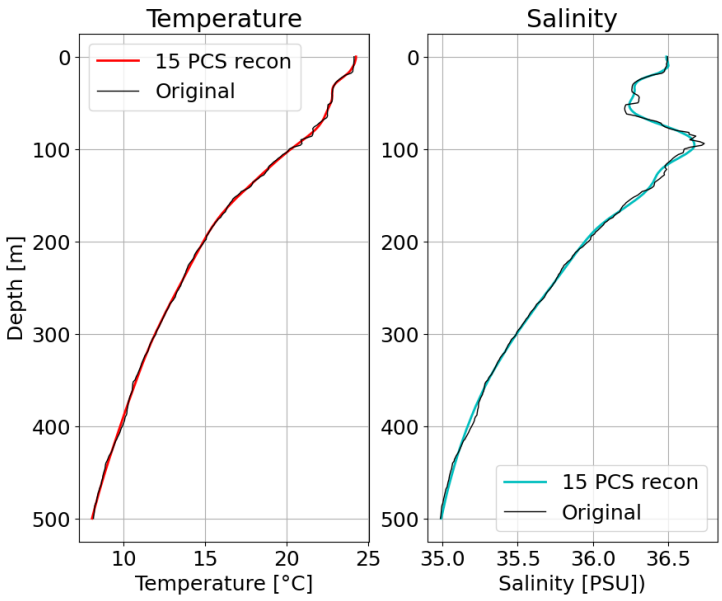


Figure 3: Example of reconstruction of temperature and salinity profiles using 15 PCS. The profile were truncated at 500 meters to emphasize the differences, which occur mostly in the upper ocean.

3.2. ~~NN Architecture~~NeSPReSO

Let ~~$X \subset \mathbb{R}^{d_X}$ be~~ $X \subset \mathbb{R}^{d_X}$ denote our input space~~(possible surface measurements) and $Y \subset \mathbb{R}^{d_Y}$~~, representing spatial and temporal information along with surface measurements (e.g., sea surface temperature, salinity, and height), and let $Y \subset \mathbb{R}^{d_Y}$ be the output space ~~(possible vertical profiles). Our ultimate goal is to find a mapping operator $\Phi : X \to Y$ that for all measurement vectors $x \in X$, there exists a corresponding T and S profile $y \in Y$ such that $y = \Phi(x)$~~consisting of the corresponding vertical profiles of temperature and salinity that we aim to predict. Our objective is to construct a mapping $\Phi : X \to Y$ such that for each input vector $x \in X$, the predicted profile $y = \Phi(x)$ approximates the true profile $y \in Y$.

~~Suppose the output space Y can be encoded into a space $Z \subset \mathbb{R}^{d_Z}$, where $d_Z \leq d_Y$~~Due to the high dimensionality of the vertical profiles, directly predicting y with a neural network can be computationally intensive, inaccurate, and prone to overfitting. To address this, we employ Principal Component Analysis (PCA) for dimensionality reduction, focusing on modeling the most significant features of the profiles (Jolliffe and Cadima, 2016; Preisendorfer and Mobley, 2023)

. Formally, we encode the output space Y into a lower-dimensional space $Z \subset \mathbb{R}^{d_Z}$, where $d_Z \ll d_Y$, using an encoder ~~E_Y, and reconstructed~~ E_Y such that $z = E_Y(y)$, and reconstruct the profiles with a decoder ~~D_Y, such that $y \approx E_Y(z)$ when $z = D_Y(y)$, for all $z \in Z$~~D_Y such that $y \approx D_Y(z)$.

~~Given a collection of inputs from X with corresponding profiles from Y, applying empirical PCA on these profiles yields the principal components (encodings) z and defines a decoder operator $D_{PCA}(z) = z\mathbf{V}^T$, where $\mathbf{V}$ is the eigenvector matrix calculated by the empirical PCA . In this framework~~Applying PCA to the profiles in Y yields the PCS z and defines the decoder operator $D_{PCA}(z) = z\mathbf{V}^{\mathrm{T}}$, where $\mathbf{V}$ is the matrix of eigenvectors from the PCA decomposition. Here, the encoder ~~$\xi : X \to Z$ emerges, a transformation that compresses~~ E_Y corresponds to the PCA transformation mapping profiles y to their PCS z, and the decoder D_Y corresponds to the inverse PCA transformation reconstructing y from z.

To predict z from the surface measurements x, we design a neural network $\zeta : X \to Z$ that approximates the mapping from the input space ~~X into the reduced PCA space$Z$, capturing the essential features of the available data.~~ to the PCA space. This approach leverages the ability of neural networks to model complex nonlinear relationships between inputs and outputs. By training the neural network to predict z, we can reconstruct the full profiles using the inverse PCA transformation. Combining PCA with neural networks is a common practice in machine learning for handling high-dimensional outputs (Howley et al., 2006; Sun et al., 2023), as PCA reduces the output dimmensionality and the neural network captures the nonlinear relationships between inputs and principal components.

~~We approximate this encoding process ξ with a neural network ζ. We have two possible minimization approaches for~~ In designing the loss function ~~: we can evaluate the NN outputs $\hat{z}$ directly against the known PCS z, or the differences~~ for training the neural network ζ, we consider the accuracy of the reconstructed profiles. Specifically, we minimize the difference between the reconstructed ~~profile $D_{PCA}(\zeta(x))$ and the actual data y.~~

PCS $\hat{z}$ and the true PCS z, and difference between the reconstructed profiles $\hat{y} = D_{PCA}(\hat{z})$ and the true profiles y. Our approximation process can be formalized as:

$$\min_{\zeta} \mathcal{L} = \underbrace{w_s \, WRMSE \, \frac{1}{nL_W} \sum_{i=1}^{n} \sum_{j=1}^{d_z} \frac{v_j}{\sigma_z^2} \left(\hat{z}_{ij} - z_{ij}\right)^2}_{WMSE} + \underbrace{FRMSE \, \frac{1}{nL_F} \left(\sum_{i=1}^{n} \left(\frac{1}{\sigma_T^2} \sum_{k=1}^{d_Y} \left(\hat{Y}_{ik}^T - Y_{ik}^T\right)^2 + \frac{1}{\sigma_S^2} \sum_{k=1}^{d_Y} \left(\hat{Y}_{ik}^S - Y_{ik}^S\right)^2 \right) \right)}_{FMSE} \tag{5}$$

~~where $\mathcal{L}$ denotes the loss function~~where $\mathcal{L}$ denotes the total loss function, n is the number of profiles (indexed by i), d_z is the number of principal components used (indexed by j), and d_Y is the number of depth levels in each profile (indexed by k). The first term is the ~~weighted root mean square error (WRMSE~~weighed mean squared error (WMSE) of the PCS, ~~with each j-th component scaled by a weight proportional to the corresponding captured variance~~ weighted by the variance captured by each component v_j, where $\hat{z}_{ij}$ and z_{ij} represent the predicted and true PCS for sample i and component j, respectively. The second term ~~is a functional root mean square error (FRMSE~~represents the functional mean squared error (FMSE), which ~~compares the functional output $\hat{y} = D_{PCA}(\zeta(x))$ against the target y.~~ is computed for both temperature and salinity profiles. Specifically, $\hat{Y}_{ik}^T$ and Y_{ik}^T denote the predicted (after inverse PCA transformation) and true temperature values, respectively, at depth k for sample i. Similarly, $\hat{Y}_{ik}^S$ and Y_{ik}^S represent the predicted and true salinity values.

It's important to note that in our model ~~$\mathcal{L}$~~ $\mathcal{L}$ accounts for both temperature and salinity predictions simultaneously~~. This is not an issue for WRMSE since the term v_j already performs the scaling, however FRMSE requires a normalization term w_p in order to penalize temperatureand salinity differences properly. A scaling factor w_s is used to account for the differences in scale between WRMSE and FRMSE as well~~, which have different scales and units. To ensure that the contributions of these parameters are appropriately scaled in this multi-task model, each mean squared error term is divided by the variance of the respective parameter: σ_z^2 for the PCS, σ_T^2 for temperature, and σ_S^2 for salinity (Zhang and Yang, 2017).

Additionally, training the model using WMSE or FMSE individually results in different loss values, with $L_W \approx 0.0255$ for WMSE and $L_F \approx 2.8294$ for FMSE. These values are used to normalize each term when combining them in the final loss function.

The neural network used in this study consists of a simple multilayer perceptron, suitable for regression tasks involving continuous outputs, with an input layer that receives satellite-derived ~~SSH~~ADT, SST, and SSS bicubicly interpolated to the location of each Argo profile. It also receives

spatial information coming from the latitude and longitude. ~~We normalize the temporal and spatial inputs to 1 ($\frac{lat}{180}$, $\frac{lon}{360}$ and $\frac{day}{365}$), and~~ Recognizing that latitude and longitude represent angular measurements with cyclical properties, we compute the sine and cosine harmonics for each normalized temporal and spatial inputs ($2\pi\frac{lat}{180}$, $2\pi\frac{lon}{360}$ and $2\pi\frac{day}{365}$), helping the network to ~~"understand" the circular~~ capture the cyclical nature of these parameters, which has been shown to improve model performance in previous studies (Thottakkara et al., 2016). The output layer produces the predicted PCS, which are then used to reconstruct the full temperature and salinity profiles using the inverse PCA transformation.

We use ~~a simple multilayer perceptron, with~~ 2 fully connected hidden layers with 512 neurons each, employing the Rectified Linear Unit (ReLU) activation function ~~, and~~ to reduce computational complexity and mitigate vanishing gradients (Dubey et al., 2022; Nguyen et al., 2021). To prevent over-fitting, we apply a dropout rate of 20%~~for regularization. The output layer of the NN is designed to approximate the PCS, enabling the later reconstruction of temperature and salinity profiles using the inverse PCA transformation~~, randomly disabling neurons during training, which encourages the network to learn more robust features (Zhang et al., 2024). Additional training parameters include using a batch size of 300, a maximum number of 8000 epochs and an early stopping mechanism of 500 epochs, if the loss value in the validation set is not improved.

The training of the neural network involves an iterative process where the model learns to approximate the ~~principal component scores (PCS)~~ PCS through exposure to different subsets of the data. The model is trained using 70% of the profiles (2,895 in total), while its performance is continuously monitored against a separate validation set comprising 621 (15%) profiles, which effectively determines when the training should stop. Training the model on this setting took 8 minutes using a single GPU. Evaluation of the model's accuracy is conducted on the remaining 15% of the data (621 profiles), the test set, to assess its predictive capabilities. ~~Additional training parameters include using a batch size of 300, a maximum number of 8000 epochs and an early stopping mechanism of 500 epochs, if the loss value in the validation set is not improved.~~

NeSPReSO is compared against ~~two~~ standard models for creating synthetic profiles: Multiple Linear Regression, Gravest Empirical Modes and Improved Synthetic Ocean Profile.

3.3. *Multiple Linear Regression Approach*

In addition to the neural network architecture, we explore a MLR model as a baseline method for predicting the PCS from surface measurements (Carnes et al., 1994). The MLR serves to assess the effectiveness of the neural network by comparing its performance with a simpler, linear approach.

Let us consider the same input space $X \subset \mathbb{R}^{d_X}$, output space $Y \subset \mathbb{R}^{d_Y}$ and the reduced-dimensional space $Z \subset \mathbb{R}^{d_Z}$, where $d_Z \ll d_Y$, along with the encoder E_Y and decoder D_Y mappings. The MLR model aims to establish a linear relationship between the input variables in X and the PCS in Z. Specifically, we model each principal component score z_j as a linear combination of the input features:

$$\hat{z}_j = \beta_j + \sum_{i=1}^{d_X} \beta_{ij} x_i, \tag{6}$$

where $\hat{z}_j$ is the predicted PCS for component j, β_j is the intercept term, β_{ij} are the regression coefficients, and x_i represents the input features from X. The regression coefficients β are then estimated by solving the least squares problem:

$$\beta = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}, \tag{7}$$

where $\mathbf{Z}$ is the matrix of true PCS obtained from PCA, and $\mathbf{X}$ is the expanded feature matrix. The inverse operation $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ denotes the pseudoinverse when $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible. This estimation provides the exact least squares solution for the regression coefficients.

The MLR model predicts the PCS by applying the estimated coefficients to new input data:

$$\hat{\mathbf{Z}}_{MLR} = \mathbf{X}_{\mathrm{new}}\boldsymbol{\beta}, \tag{8}$$

where $\mathbf{X}_{\mathrm{new}}$ contains the polynomial features of the new input samples. The predicted PCS $\hat{\mathbf{Z}}_{MLR}$ are then used with the decoder D_Y to reconstruct the full temperature and salinity profiles:

$$\hat{Y}_{MLR} = D_Y(\hat{\mathbf{Z}}_{MLR}) = \hat{\mathbf{Z}}_{MLR}\mathbf{V}^{\mathrm{T}}, \tag{9}$$

where $\mathbf{V}$ is the matrix of eigenvectors from the PCA decomposition.

In our implementation, we include the same inputs as in our NN approach: spatial and temporal harmonics of latitude, longitude, day of the year, and satellite SST, SSH and ADT. The MLR model is trained using the combined training and validation datasets, comprising 3,516 profiles (85% of the total data), to ensure sufficient data for estimating the regression coefficients accurately. Fitting the model took 180 milliseconds on a single GPU.

The remaining 15% of the data (621 profiles) is used as a test set to evaluate the model's predictive performance. By comparing the MLR results with those of the neural network, we can assess the benefits of incorporating nonlinear activation functions and deeper architectures in capturing complex relationships within the data, and by comparing with GEM, we can assess the advantages of operating in a reduced dimensional space.

It's important to note that we initially experimented with polynomial expansions up to degree 3 to capture potential nonlinear relationships between the input variables and the PCS. However, these higher-degree models exhibited significant issues:

- Computational Challenges: The inclusion of polynomial terms up to degree 3 dramatically increased the dimensionality of the feature matrix. With a large number of samples and input variables, the feature matrix became extremely large. This led to high memory consumption ($\approx$ 80GB) and computational inefficiency and instabilities during model fit.

- Numerical Instability: The large size of the matrices exacerbated numerical issues, such as difficulty in inverting matrices during the estimation of regression coefficients. This instability adversely affected the model's ability to learn accurate relationships.

- Overfitting: The expanded feature space increased the risk of overfitting, where the model captured noise rather than underlying patterns, resulting in poor generalization to unseen data.

- Multicollinearity: Higher-degree polynomial terms introduced strong correlations among predictor variables, destabilizing coefficient estimates and reducing the reliability of the model.

As a result of these challenges, the higher-degree polynomial models were unstable, producing predictions that were too inaccurate for practical application. Therefore, we opted to use the degree 1 MLR model, which captures linear relationships between the input variables and the PCS.

3.4. Gravest Empirical Modes

~~The Gravest Empirical Modes (GEM)~~ The GEM method is a technique extensively utilized in oceanography for the generation of synthetic temperature and salinity profiles. The GEM method is based on the establishment of an empirical relationship between dynamic height and other oceanographic parameters, capturing the essential spatiotemporal patterns of oceanic temperature and salinity, making it a valuable tool for studying and simulating these parameters. This method has been applied to various oceanic regions, contributing to a better understanding of ocean dynamics and climate processes ~~(Watts et al. (2001); Liu et al. (2021); Meunier et al. (2022))~~(Watts et al., 2001; Liu et al., 2021; Meunier et al., 2022).

The implementation of the GEM method ~~by month~~ is described as follows:

A. The steric height is computed for each in situ profile of temperature and salinity.
B. All profiles are sorted according to their steric height, and grouped by month.
C. A regular pressure grid is defined (0–1800 dbar) with a vertical grid-step of 1 dbar. For each reference pressure value and for each month, a cubic smoothing spline is fitted to the functions ~~$T(\zeta)|_{p,m}$ and $S(\zeta)|_{p,m}$, where T and S~~ $T(\zeta)|_{p,m}$ and $S(\zeta)|_{p,m}$, where T and S are temperature and salinity, ~~ζ~~ ζ is ADT, ~~p~~ p is the pressure at which the variables are evaluated, and m is the month.

3.5. ~~Improved Synthetic Ocean Profile (ISOP)~~

~~The Improved Synthetic Ocean Profile (ISOP) method projects surface ocean data downward, generating T and S profiles across the global ocean using surface observations and a mixed-layer depth (MLD) estimate. Optionally, a prior forecast of T and S profiles can be used. The creation of these synthetic profiles plays an important step in the Navy's operational forecasting and is seamlessly integrated into their data assimilation workflows. ISOP divides the ocean's depth into 78 fixed levels, extending from the surface to 6600 meters. The process begins with the compilation of a T and S~~

~~covariance matrix and climatology database from a comprehensive set of in-situ observations, followed by the application of a multilayered approach that considers three different dynamics zones within the ocean subsurface. These regions include the *mixed layer*, extending from the surface to the MLD; the *thermocline layer*, reaching from the MLD down to 1000 meters; and the *deep ocean layer*, below 1000 meters.~~

~~For the *mixed layer*, there are two options. One option adjusts the initial estimated profile to align with the surface potential density at 4 meters depth and ensures consistency with the potential density and its gradient at the MLD within the *thermocline layer*. The second option for the *mixed layer* shifts the prior forecast profile (if provided) to match the input SST value. The *thermocline layer* prediction employs a variational method, leveraging climatological T and S values and the first vertical Empirical Orthogonal Functions (EOFs), or modes, extracted from historical data to constrain the forecast. Detailed descriptions of the each term involved in this variational approach is available in reference Helber et al. (2013). Finally, the prediction within the *deep ocean layer* involves modifying a decay function based on climatological data and the T and S readings from the *thermocline layer* at 1000 meters depth. This function accounts for the variance between climatological values and the 1000-meter predictions, ensuring a coherent transition into the deep ocean predictions.~~

~~As previously mentioned, the inputs for ISOP's predictive models include SST and SSH, along with uncertainty estimates, an MLD estimation, and an (optional) T and S profile obtained from either climatological data or model outputs. Later comparisons in this work use climatological data for estimating the initial MLD and T and S profiles, along with Argo-derived SST and SSH. Please note we did not have access to the ISOP profiles, only the statistics.~~

The process of fitting GEM to the dataset took 3 seconds on CPU.

4. Results

In this section we analyze the performance of NeSPReSO with respect to ~~the~~ 621 Argo profiles in our test dataset (15% of the dataset, randomly selected, not used in training), and compare its performance against GEM, MLR and ISOP methods. We also generate ~~synthetic profiles~~ NeSPReSO synthetics to reconstruct four glider sections in the GoM. ~~NeSPReSO~~

The average processing time per profile on CPU is around 60µs for NeSPReSO, 20µs for MLR and 11600µs for our GEM implementation, when generating synthetics for our test set. However, it's important to note that in an operational setting, where profiles are generated on the fly, the time to extract the satellite information from stored data is the limiting factor for generating synthetics, regardless of the method used (0.5s per day of interest, regardless of the number of profiles).

NeSPReSO and MLR synthetics were generated using satellite surface information (ADT, SST and SSS) interpolated to the locations of the measurements, location, and day of the year, while GEM synthetics used month and ADT. ISOP utilized climatological MLD and profile-derived SSH and SST, ~~which were~~ with only statistical summaries of the ISOP synthetics being available, rather than individual profiles. This limitation, along with the fact that ISOP synthetics was not derived from satellite ~~data as with~~ sources like the other methods~~. This distinction potentially skews~~ , may skew the comparison in the upper ocean~~between ISOP and the other techniques~~.

4.1. Test set

We use root mean square error (RMSE) and bias as analysis metrics to evaluate the performance of our model relative to observations. RMSE, measuring precision and accuracy, indicates the model's prediction consistency and closeness to observed values. RMSE penalizes larger deviations and reflects the average prediction error, with lower RMSE indicating more reliable predictions. Bias measures the average deviation from observed values, showing if the model consistently overestimates or underestimates the variable under consideration. Both statistics are given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}, \tag{10}$$

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right), \tag{11}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and N is the number of observations. For calculations at each depth level, N represents the number of profiles at that depth. When computing RMSE and bias over a depth range, the statistics are averaged over all depths within that range.

The Pearson correlation coefficient (R^2) quantifies the degree of linear correlation between the predicted and observed values, with values closer to 1 indicating a stronger correlation. It is calculated as:

$$R^2 = \left(\frac{\sum_{i=1}^{N} (y_i - \bar{y}) (\hat{y}i - \bar{\hat{y}})}{\sqrt{\sum i = 1^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2, \tag{12}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the mean values of the observed and predicted data, respectively. The R^2 metric assesses the proportion of variance in the observed data that is predictable from the predicted data. Since we don't have access to individual ISOP synthetics, we could not calculate R^2 for ISOP.

The statistics of the profiles in the test set are shown on table 1, calculated using predictions at the same depths as ISOP, for fairness. For temperature, the RMSE values indicate that NeSPReSO consistently outperforms the GEM predictions across all depth ranges, MRL below 20 meters and ISOP below 100 meters. ~~It~~ However, it is difficult to draw comparisons with ISOP near the surface, given that it uses Argo SST, but we observe a more accurate estimation of temperature profiles compared to the GEM method, which we attribute to the use of satellite SST. Bias values for temperature are comparable between all methods, implying that the methods exhibit a similar direction and magnitude of systematic error in temperature estimation.

~~In salinityestimation~~For salinity, NeSPReSO also demonstrates lower RMSE and bias values than the other methods for most of the depth ranges, indicating superior performance in salinity predictions.

The Pearson correlation coefficient (R^2) values for both T and S predictions are higher for NeSPReSO compared to GEM across all depths, and particularly pronounced in the upper 100 meters. NeSPReSo also overperforms MLR in most cases, except for T on the range from 0 to 20 meters. This improvement in R^2 signifies a stronger ~~linear~~ correlation between predictions and observations, ~~highlighting NeSPReSO's enhanced accuracy in characterizing~~ meaning a better characterization of the upper-ocean.

Table 1: Statistics (RMSE, Bias, and R^2) by depth range. Best results in bold.

Depth range		0-20	20-100	100-200	200-500	500-1000	1000-1800	0-1000
T RMSE (°C)	NeSPReSO	0.439 0.430	0.846 **0.816**	**0.832** **0.802**	**0.608** **0.587**	**0.313** **0.301**	**0.084** **0.083**	0.706 0.682
	GEM	1.461 1.468	1.416 1.419	1.095 1.094	0.860 0.854	0.396 0.394	0.127 0.125	1.194 1.195
	MLR	0.380	1.031	0.944	0.699	0.357	0.087	0.823
	ISOP	**0.140**	**0.835** 0.835	0.917	0.756	0.360	0.111	**0.673**
T BIAS (°C)	NeSPReSO	0.042 0.047	**-0.058** **-0.038**	**-0.040** **0.015**	**-0.034** 0.016	-0.027 **0.005**	**0.001** 0.003	**-0.030** **0.00**
	GEM	-0.042 -0.043	-0.151 -0.153	-0.058 -0.059	-0.037 -0.036	**0.006** 0.006	0.006	-0.076 -0.07
	MLR	**-0.011**	-0.041	0.016	**-0.001**	-0.010	**0.000**	-0.014
	ISOP	**0.022** 0.022	0.186	0.203	0.137	-0.057	-0.074	0.127
T R^2	NeSPReSO	**0.982** **0.953** 0.983	**0.969** **0.956**	**0.985** **0.971**	**0.986**	**0.972** **0.987**	**0.973**	**0.995**
	GEM	0.775 0.773	0.871 0.870	0.949	0.970	0.977 0.978	0.940 0.941	0.986
	MLR	**0.986**	0.929	0.960	0.980	0.981	0.970	0.993
S RMSE (PSU)	NeSPReSO	**0.334** **0.280**	**0.151** **0.139**	**0.119** **0.116**	**0.092** **0.088**	**0.034** **0.032**	**0.009**	**0.174** **0.15**
	GEM	0.478	0.193	0.165 0.163	0.123 0.122	0.046	0.009 **0.009**	0.241
	MLR	0.299	0.154	0.155	0.112	0.044	**0.009**	0.173
	ISOP	0.604	0.229	0.160	0.147	0.049	0.015	0.240
S BIAS (PSU)	NeSPReSO	**-0.018** **0.012**	**-0.007** **-0.002**	**0.000** 0.005	-0.009 0.003	**0.000** **-0.001**	**0.000**	**-0.007** **0.00**
	GEM	-0.035 -0.036	-0.010	-0.014	**-0.005** -0.005	0.002	0.000 **0.000**	-0.013
	MLR	-0.021	-0.007	**0.002**	**0.001**	**-0.001**	**0.000**	-0.005
	ISOP	-0.092	-0.086	-0.033	0.023	-0.009	-0.010	-0.048
S R^2	NN NeSPReSO	**0.748** **0.829**	**0.674** **0.729**	**0.879** **0.887**	**0.984** **0.985**	**0.976** **0.977**	**0.858** **0.861**	**0.951** **0.962**
	GEM	0.337	0.409 0.411	0.786 0.789	0.971	0.957 0.958	0.833	0.905
	MLR	0.803	0.654	0.786	0.975	0.957	0.857	0.952

Figure 4 presents the average T and S RMSE and bias ~~of the three methods across varying depths for temperature and salinity~~per depth for all methods. In general, NeSPReSO yields better approximations compared to the other methods, as indicated by the lower RMSE and bias values overall. The ~~method also exhibits bias comparable to GEM and lower than ISOP. The~~ improved prediction of upper-ocean temperature and salinity profiles in our model compared to GEM is likely due to the use of satellite SST and SSS, which offer additional information about the upper thermal and haline structures that might not be captured in the ADT fields, such as low salinity due to river outflow.

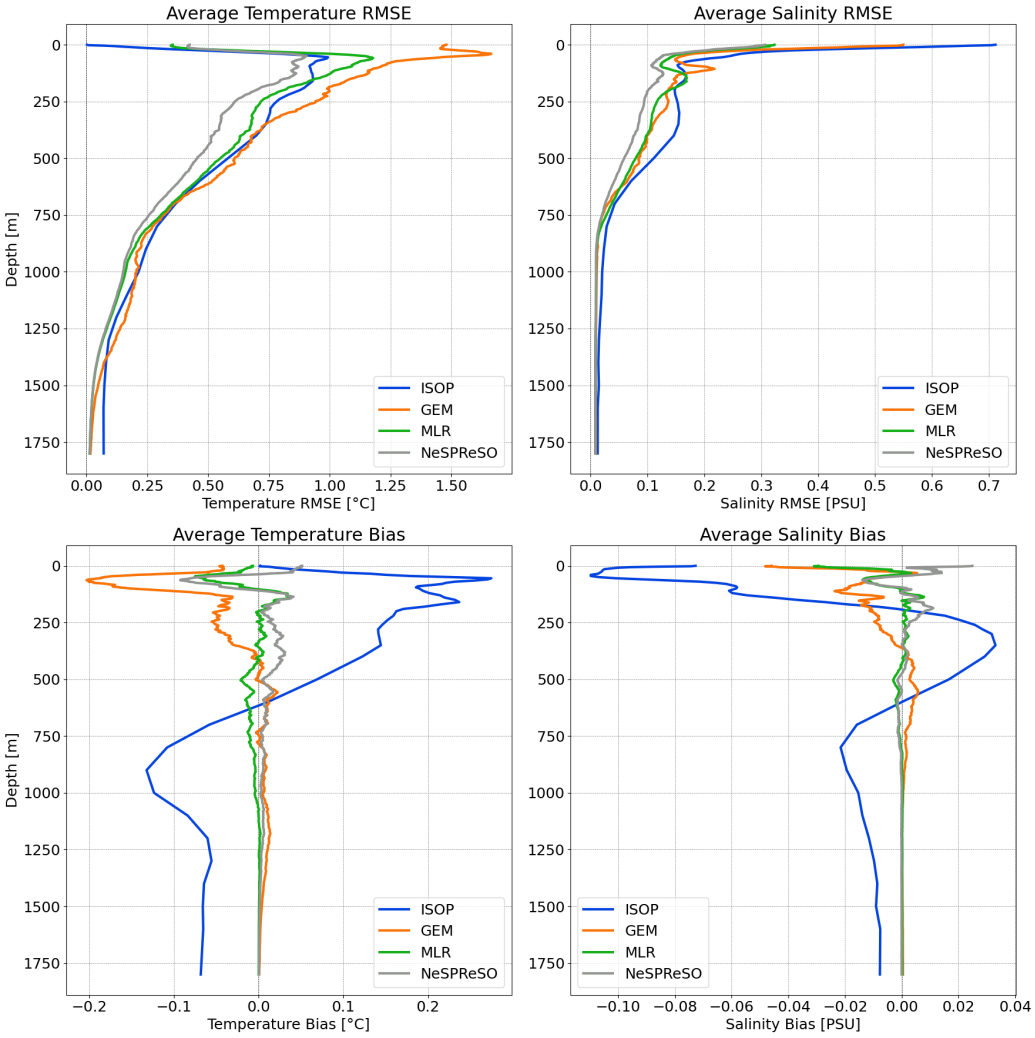


Figure 4: Average RMSE for temperature and salinity predictions (top), and average bias (bottom) as a function of depth.

The synthetic profiles were aggregated spatially into 1-degree latitude by 1-degree longitude grid cells to assess the methods' performance in predicting T and S across the area of study. Figures referenced as 5 through 8 present the spatial distribution of RMSE and bias for T and S. The statistics were calculated using predictions at the same depths as ISOP for a fair comparison.

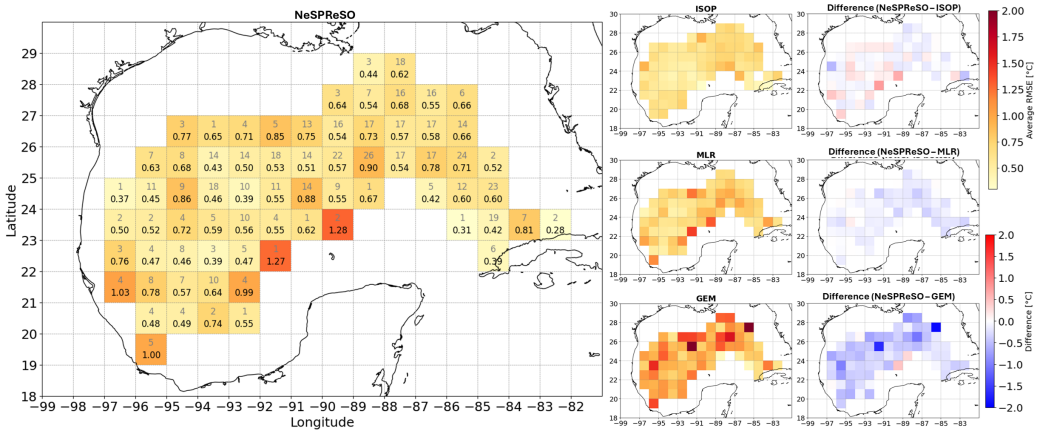


Figure 5: Distribution of average temperature RMSE for predictions down to 1,800m for NeSPReSO (left), ~~GEM (top), ISOP (bottom), and differences compared to NeSPReSO (right). The~~ with the number of profiles in each bin is displayed in gray, and RMSE values in black. Statistics for ISOP (top), MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO are shown on the right column (blues indicate NeSPReSO performs better, and reds indicate NeSPReSO performs worse).

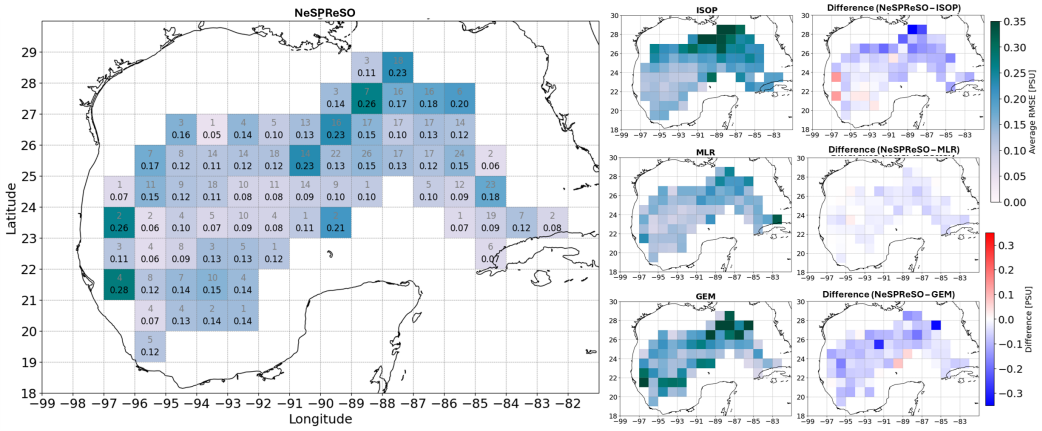


Figure 6: Distribution of average salinity RMSE for predictions down to 1,800m for NeSPReSO (left), ~~GEM (top), ISOP (bottom), and differences compared to NeSPReSO (right). The~~ with the number of profiles in each bin is displayed in gray, and RMSE values in black. Statistics for ISOP (top), MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO are shown on the right column (blues indicate NeSPReSO performs better, and reds indicate NeSPReSO performs worse).

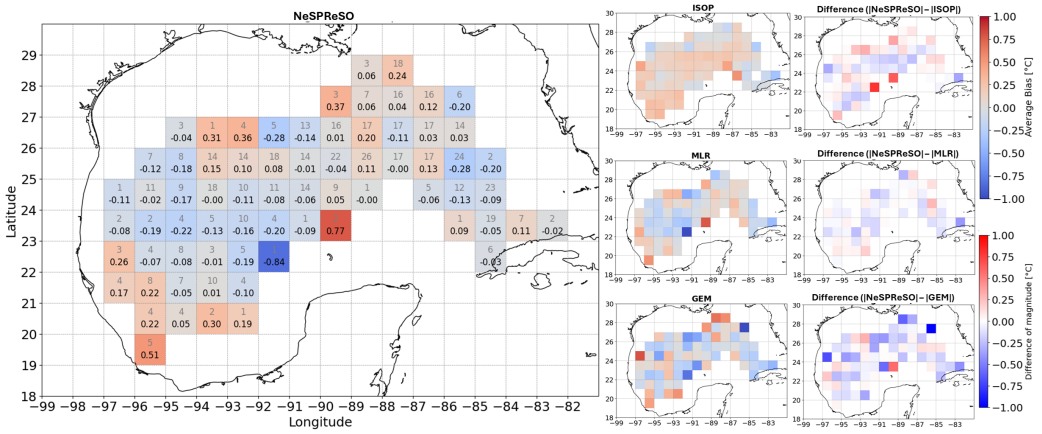


Figure 7: Distribution of average temperature bias for predictions down to 1,800m for NeSPReSO (left) with the number of profiles in each bin is displayed in gray, ~~GEM~~ and bias values in black. Statistics for ISOP (top), ~~ISOP~~ MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO ~~(~~are shown on the right ~~). The number of profiles in each bin is displayed in gray~~column (blues indicate NeSPReSO performs better, and ~~bias values in black~~reds indicate NeSPReSO performs worse).

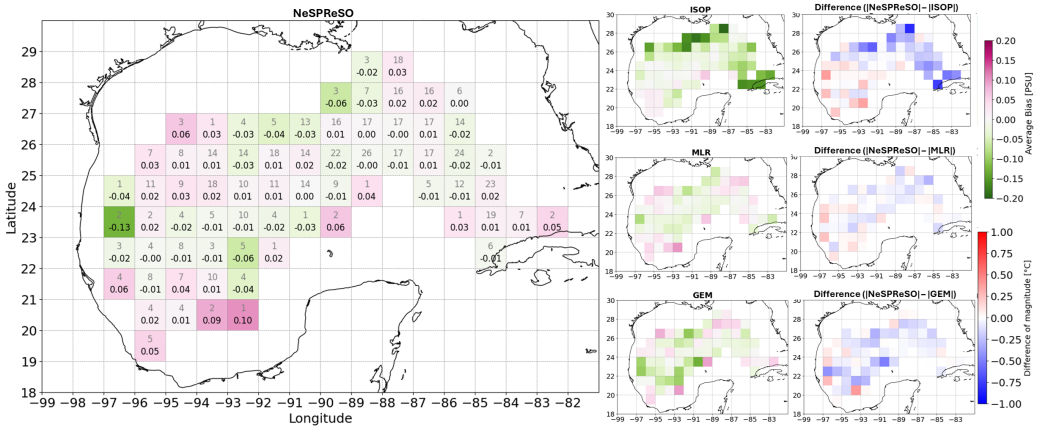


Figure 8: Distribution of average salinity bias distribution for predictions down to 1,800m for NeSPReSO (left), ~~GEM (top), ISOP (bottom), and differences compared to NeSPReSO (right). The~~ with the number of profiles in each bin is displayed in gray, and bias values in black. Statistics for ISOP (top), MLR (center), and GEM (bottom) are shown in the center column, and their respective differences in magnitude compared to NeSPReSO are shown on the right column (blues indicate NeSPReSO performs better, and reds indicate NeSPReSO performs worse).

The results indicate a robust performance of NeSPReSO in real-world scenarios and applications, as NeSPReSO has lower overall RMSE for both T and S predictions across the entire GoM region, with a few exceptions. NeSPReSO shows a spatial distribution of bias predominantly of low magnitude and somewhat homogeneous (no apparent predominant bias). MRL has a very similar spatial distributions as NeSPReSO, with slightly higher magnitudes. GEM also demonstrates a relatively homogeneous distribution, but ~~of~~ with even higher magnitude on average. Meanwhile, ISOP exhibits a clear warmer and low magnitude trend for T and fresher for S, with greater magnitudes in the eastern portion of the GoM. Notably, in regions adjacent to the Mississippi River, ISOP demonstrates increased errors.

4.2. Glider tracks

This section presents a comparative analysis of processed glider tracks against the reconstructions from NeSPReSO, offering a direct assessment of the model's performance by replicating independent observations.

Figures 9 to 12 illustrate four different processed glider crossings with the corresponding synthetic reconstructions and the differences. Overall, the displacement of isothermals and isohalines are in agreement with the observations, and the reconstructed fields are smoother, as expected.

Table 2 shows the RMSE, bias, and ~~the coefficient of determination (~~R^2 ~~)~~ for each LCE crossing~~, which quantifies the variance captured by the model~~. The T and S RMSE closely aligns with those derived from the test set ([0-1000] range on Table 1). The bias for T and S exhibits a larger magnitude relative to the test set across each crossing, with variations between positive and negative biases. One possible explanation for these variations is related to the temporal and spatial resolution of satellite observations, particularly of ADT. These factors may contribute to a consistent directional bias in the model's predictions.

The R^2 values range from ~~0.993 to 0.999~~ 0.996 to 0.998 for T predictions, and from 0.988 to ~~0.992~~ 0.994 for S predictions, meaning NeSPReSO consistently captures around 99% of the T and S variances.

Crossing	T RMSE	T Bias	T R^2	S RMSE	S Bias
~~Poseidon~~ Mission 0006, crossing #1	~~0.586~~ 0.546	~~0.031~~ 0.070	~~0.996~~ 0.997	~~0.118~~ 0.096	~~-0.011~~ -0.006
~~Poseidon~~ Mission 0006, crossing #2	~~0.553~~ 0.516	~~-0.207~~ -0.119	0.998	~~0.111~~ 0.094	~~-0.035~~ -0.025
~~Campeche~~ Mission 0010	~~0.524~~ 0.544	~~0.079~~ 0.121	0.996	~~0.069~~ 0.072	~~0.017~~ 0.020
~~Intense LCE~~ Mission 0012	~~0.730~~ 0.586	~~-0.133~~ 0.003	~~0.996~~ 0.997	~~0.105~~ 0.086	~~-0.047~~ -0.035

Table 2: RMSE, bias and ~~R^2~~ R^2 between observations and synthetics across mesoscale eddy crossings.

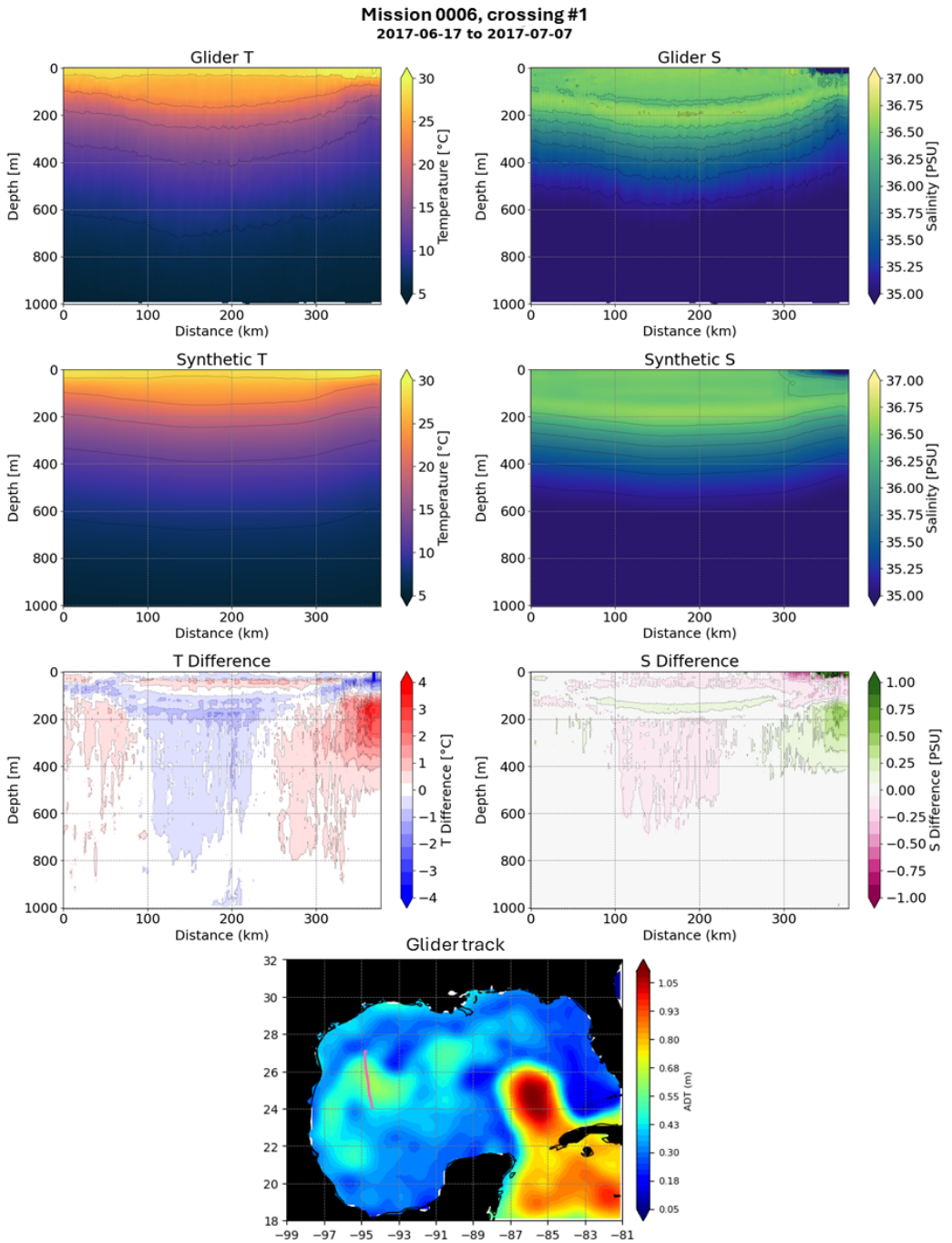


Figure 9: Temperature and salinity sections of ~~the Poseidon LCE.~~ mission 0006, crossing #1. First column: Temperature. Second column: Salinity. ~~Top~~ First row: processed data from glider. ~~Middle~~ Second row: synthetic profiles using NeSPReSO. ~~Bottom~~ Third row: differences. Last row: ADT field and position of the glider track.

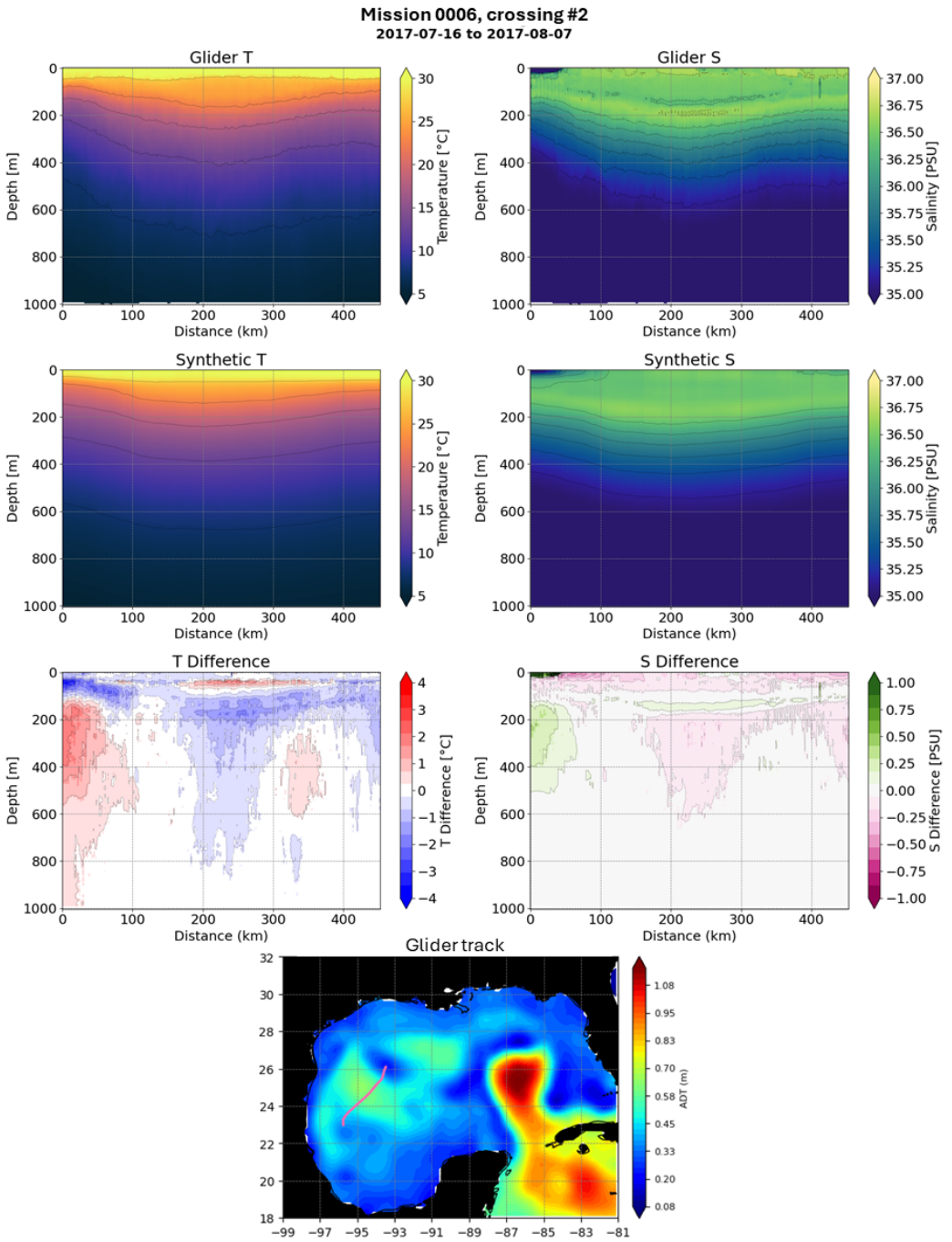


Figure 10: ~~Another~~ Temperature and salinity sections of ~~the Poseidon LCE.~~ mission 0006, crossing #2. First column: Temperature. Second column: Salinity. ~~Top~~ First row: processed data from glider. ~~Middle~~ Second row: synthetic profiles using NeSPReSO. ~~Bottom~~ Third row: differences. Last row: ADT field and position of the glider track.

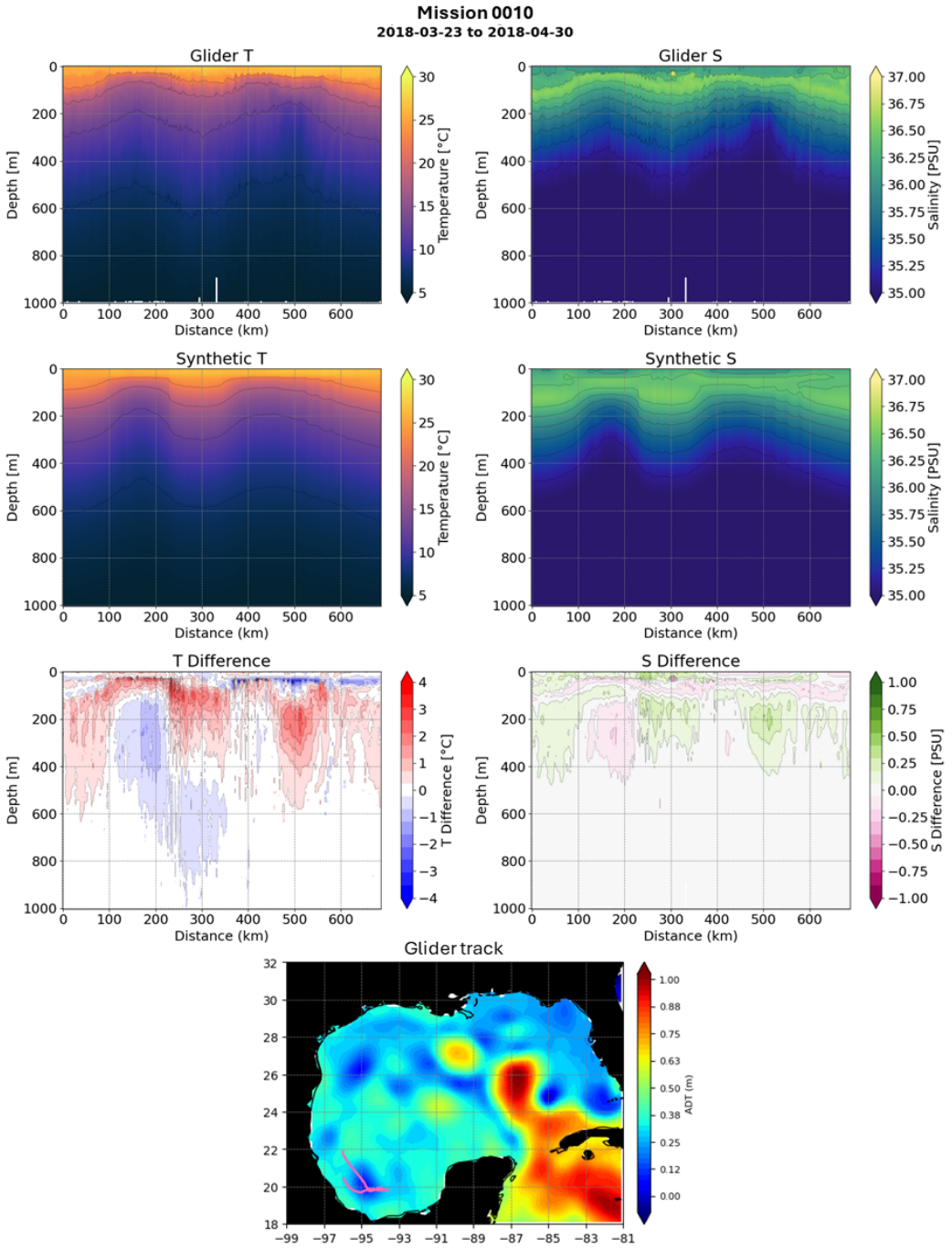


Figure 11: Temperature and salinity sections of ~~the cyclonic eddy in Campeche Bay.~~ mission 0010. First column: Temperature. Second column: Salinity. ~~Top~~ First row: processed data from glider. ~~Middle~~ Second row: synthetic profiles using NeSPReSO. ~~Bottom~~ Third row: differences. Last row: ADT field and position of the glider track.

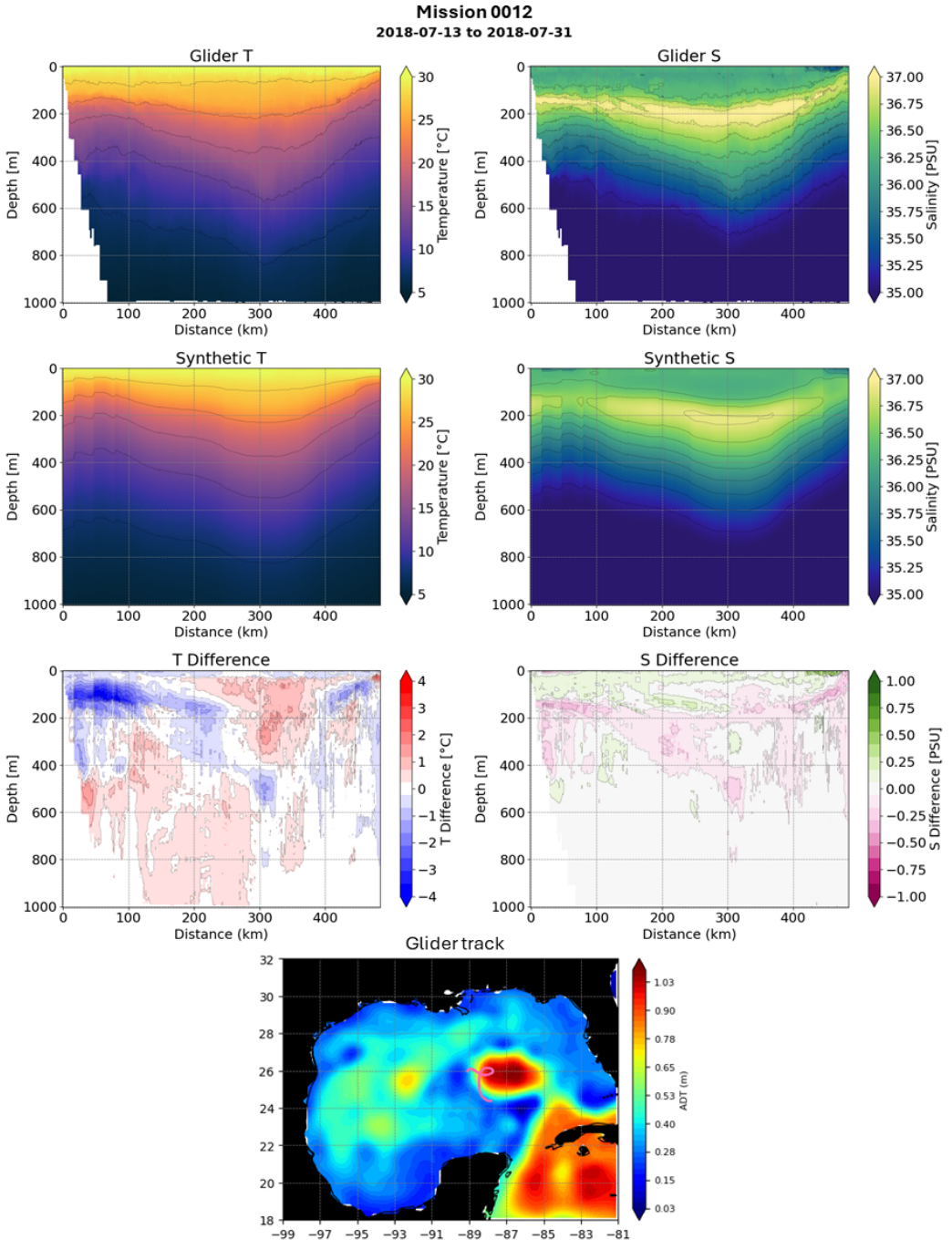


Figure 12: Temperature and salinity sections of ~~an intense LCE.~~ mission 0012. First column: Temperature. Second column: Salinity. ~~Top~~ First row: processed data from glider. ~~Middle~~ Second row: synthetic profiles using NeSPReSO. ~~Bottom~~ Third row: differences. Last row: ADT field and position of the glider track.

5. Conclusions

This study underscores the efficacy of machine learning in producing synthetic temperature and salinity profiles for oceanographic data. By integrating Principal Component Analysis (PCA) with neural network models, we successfully generated subsurface profiles from surface data, surpassing traditional methods like MLR, GEM and ISOP in accuracy and reliability.

Our results indicate that the neural network model consistently outperforms ~~both GEM and ISOP method~~ other investigated methods in terms of average RMSE~~and bias~~, bias, and R^2, suggesting a more accurate representation of the temperature and salinity profiles in the Gulf of Mexico. This improvement is notable given the complex, nonlinear relationships between surface and subsurface properties of the ocean, which machine learning models are particularly adept at capturing.

These results raises several questions that warrant further investigation. For instance, how will NeSPReSO perform in different oceanic regions with distinct hydrodynamic and thermohaline characteristics, and what adaptations might be required for different regional applications? Also, how can NeSPReSO be adapted and trained to effectively generate accurate temperature and salinity profiles in oceanic regions with depths shallower than the model's current maximum depth range?

Future work should focus on addressing these questions, perhaps exploring other machine learning techniques or hybrid models that combine the strengths of various approaches. With the UGOS3 autonomous profiling floats fleet projected to accumulate approximately 1500 profiles annually, the expanding dataset will significantly enhance the model's training and refinement. This expansion is crucial for extending the model's applicability across different oceanic areas, enriching our comprehension of its potential and constraints.

In conclusion, this work lays a precedent for using advanced machine learning methods in oceanographic data synthesis, offering a promising direction for future research in this field. The ability to accurately predict subsurface oceanographic profiles using surface data not only aids in understanding ocean dynamics but also has practical implications in weather forecasting, climate modeling, and resource exploration.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT (3.5~~and~~ , 4, and 4o) in order to improve grammar, clarity and coherence. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Behringer, D.W., Molinari, R.L., Festa, J.F., 1977. The variability of anticyclonic current patterns in the gulf of mexico. Journal of Geophysical Research (1896-1977) 82, 5469–5476. doi:10.1029/JC082i034p05469.

Carnes, M., Teague, W., Mitchell, J., 1994. Inference of subsurface thermohaline structure from fields measurable by satellite. Journal of Atmospheric and Oceanic Technology 11, 551–566. URL: https://journals.ametsoc.org/view/journals/atot/11/2/1520-0426_1994_011_0551_iostsf_2_0_co_2.xml.

Chen, Z., Wang, P., Bao, S., Zhang, W., 2022. Rapid reconstruction of temperature and salinity fields based on machine learning and the assimilation application. Frontiers in Marine Science 9, 985048. doi:10.3389/fmars.2022.985048.

Copernicus Marine Service, 2024. Global ocean gridded l4 sea surface heights and derived variables nrt. URL: https://doi.org/10.48670/moi-00149.

Dubey, S.R., Singh, S.K., Chaudhuri, B.B., 2022. Activation functions in deep learning: A comprehensive survey and benchmark URL: http://arxiv.org/abs/2109.14545.

Forristall, G.Z., Schaudt, K.J., Cooper, C.K., 1992. Evolution and kinematics of a loop current eddy in the gulf of mexico during 1985. Journal of Geophysical Research: Oceans 97, 2173–2184. doi:https://doi.org/10.1029/91JC02905.

Fox, D., Teague, W., Barron, C., Carnes, M., Lee, C., 2002. The modular ocean data assimilation system (modas). Journal of Atmospheric and Oceanic Technology 19, 240–252.

Fu, Z., Hu, L., Chen, Z., Zhang, F., Shi, Z., Hu, B., Du, Z., Liu, R., 2020. Estimating spatial and temporal variation in ocean surface pco2 in the gulf of mexico using remote sensing and machine learning techniques. Science of The Total Environment 745, 140965. doi:https://doi.org/10.1016/j.scitotenv.2020.140965.

Good, S., Fiedler, E., Mao, C., Martin, M.J., Maycock, A., Reid, R., Roberts-Jones, J., Searle, T., Waters, J., While, J., Worsfold, M., 2020. The current configuration of the ostia system for operational production of foundation sea surface temperature and ice concentration analyses. Remote Sensing 12, 720. doi:10.3390/rs12040720.

Helber, R.W., Smith, S.R., Jacobs, G.A., Barron, C.N., Carrier, M.J., Yaremchuk, M., Rowley, C.D., Ngodock, H.E., Bartels, B.P., Pasmans, I., et al., 2022. Velocity Assimilation with Improved Synthetic Ocean Profiles (ISOP2): Validation Test Report.

Helber, R.W., Townsend, T.L., Barron, C.N., Dastugue, J.M., Carnes, M.R., 2013. Validation test report for the Improved Synthetic Ocean Profile (ISOP) system, Part I: Synthetic profile methods and algorithm. Naval Res. Lab. Rep. NRL/MR/7320-13-9364 .

Hiron, L., de la Cruz, B.J., Shay, L.K., 2020. Evidence of loop current frontal eddy intensification through local linear and nonlinear interactions with the loop current. Journal of Geophysical Research: Oceans 125, e2019JC015533. doi:https://doi.org/10.1029/2019JC015533. e2019JC015533 10.1029/2019JC015533.

Hiron, L., Miron, P., Shay, L.K., Johns, W.E., Chassignet, E.P., Bozec, A., 2022. Lagrangian coherence and source of water of loop current frontal eddies in the gulf of mexico. Progress in Oceanography 208, 102876. doi:https://doi.org/10.1016/j.pocean.2022.102876.

Hiron, L., Nolan, D.S., Shay, L.K., 2021. Study of ageostrophy during strong, nonlinear eddy-front interaction in the gulf of mexico. Journal of Physical Oceanography 51, 745 – 755. doi:10.1175/JPO-D-20-0182.1.

Howley, T., Madden, M.G., O'Connell, M.L., Ryder, A.G., 2006. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data, in: Macintosh, A., Ellis, R., Allen, T. (Eds.), Applications and Innovations in Intelligent Systems XIII. Springer, London, pp. 209–222. doi:10.1007/1-84628-224-1_16.

Jaimes, B., Shay, L.K., Brewster, J.K., 2016. Observed air-sea interactions in tropical cyclone isaac over loop current mesoscale eddy features. Dynamics of Atmospheres and Oceans 76, 306–324. doi:https://doi.org/10.1016/j.dynatmoce.2016.03.001.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374. doi:10.1098/rsta.2015.0202.

Koch, S., Barker, J., Vermersch, J., 1991. The Gulf of Mexico Loop Current and Deepwater Drilling. Journal of Petroleum Technology 43, 1046–1119. doi:10.2118/20434-PA.

Leben, R.R., 2005. Altimeter-Derived Loop Current Metrics. American Geophysical Union (AGU). pp. 181–201. doi:https://doi.org/10.1029/161GM15.

Liu, H., Zhou, H., Yang, W., Liu, X., Li, Y., Yang, Y., Chen, X., Li, X., 2021. A three-dimensional gravest empirical mode determined from hydrographic observations in the western equatorial pacific ocean. Journal of Marine Systems 214, 103487. doi:https://doi.org/10.1016/j.jmarsys.2020.103487.

Lueck, R., Picklo, J., 1990. Thermal inertia of conductivity cells: Observations with a sea-bird cell. Journal of Atmospheric and Ocean Technology 7, 756–768.

Mafi, S., Amirinia, G., 2017. Forecasting hurricane wave height in gulf of mexico using soft computing methods. Ocean Engineering 146, 352–362. doi:https://doi.org/10.1016/j.oceaneng.2017.10.003.

Mao, K., Liu, C., Zhang, S., Gao, F., 2023. Reconstructing ocean subsurface temperature and salinity from sea surface information based on dual path convolutional neural networks. Journal of Marine Science and Engineering 11, 1030. doi:10.3390/jmse11051030.

Meissner, T., Wentz, F.J., Le Vine, D.M., 2018. The salinity retrieval algorithms for the nasa aquarius version 5 and smap version 3 releases. Remote Sensing 10, 1121. doi:10.3390/rs10071121.

Meng, L., Yan, C., Zhuang, W., Zhang, W., Yan, X.H., 2021. Reconstruction of three dimensional temperature and salinity fields from satellite observations. Journal of Geophysical Research: Oceans 126, e2021JC017605.

Meunier, T., Bower, A., Pérez-Brunius, P., Graef, F., Mahadevan, A., 2024. The Energy Decay of Warm-Core Eddies in the Gulf of Mexico. Geophysical Research Letters 51. doi:10.1029/2023GL106246.

Meunier, T., Le Boyer, A., Molodtsov, S., Bower, A., Furey, H., Robbins, P., 2023. Internal wave activity in the deep Gulf of Mexico. Frontiers in Marine Science 10. doi:10.3389/fmars.2023.1285303.

Meunier, T., Pérez-Brunius, P., Bower, A., 2022. Reconstructing the three-dimensional structure of loop current rings from satellite altimetry and in situ data using the gravest empirical modes method. Remote Sensing 14, 4174.

National Academies of Sciences, Engineering, and Medicine, 2018. Understanding and Predicting the Gulf of Mexico Loop Current: Critical Gaps and Recommendations. The National Academies Press, Washington, DC. doi:10.17226/24823.

Nguyen, A., Pham, K., Ngo, D., Ngo, T., Pham, L., 2021. An analysis of state-of-the-art activation functions for supervised deep neural network URL: http://arxiv.org/abs/2104.02523.

Pauthenet, E., Bachelot, L., Balem, K., Maze, G., Tréguier, A.M., Roquet, F., Fablet, R., Tandeo, P., 2022. Four-dimensional temperature, salinity and mixed-layer depth in the gulf stream, reconstructed from remote-sensing and in situ observations with neural networks. Ocean Science 18, 1221–1244. doi:10.5194/os-18-1221-2022.

Preisendorfer, R.W., Mobley, C.D., 2023. Principal component analysis in meteorology and oceanography. SERBIULA (sistema Librum 2.0). Posthumously compiled and edited by Curtis D. Mobley.

Roemmich, D., Gilson, J., 2009. The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the argo program. Progress in oceanography 82, 81–100.

Sarıkoç, M., Celik, M., 2024. Pca-ica-lstm: A hybrid deep learning model based on dimension reduction methods to predict s&p 500 index price. Computational Economics doi:10.1007/s10614-024-10629-x.

Shay, L.K., 2010. Air-Sea Interactions in Tropical Cyclones. pp. 93–131. doi:10.1142/9789814293488_0003.

Shay, L.K., Uhlhorn, E.W., 2008. Loop current response to hurricanes isidore and lili. Monthly Weather Review 136, 3248 – 3274. doi:10.1175/2007MWR2169.1.

Sturges, W., Leben, R., 2000. Frequency of ring separations from the loop current in the gulf of mexico: A revised estimate. Journal of Physical Oceanography 30, 1814–1819. doi:10.1175/1520-0485(2000)030<1814:FORSFT>2.0.CO;2.

Sturges, W., Lugo-Fernandez, A., Shargel, M.D., 2005. Introduction to Circulation in the Gulf of Mexico. American Geophysical Union (AGU). doi:https://doi.org/10.1029/161GM02.

Sun, C., Watts, D.R., 2001. A circumpolar gravest empirical mode for the southern ocean hydrography. Journal of Geophysical Research: Oceans 106, 2833–2855.

Sun, Y., Zhou, S., Meng, S., Wang, M., Mu, H., 2023. Principal component analysis–artificial neural network-based model for predicting the static strength of seasonally frozen soils. Scientific Reports 13. doi:10.1038/s41598-023-43462-7.

Suthers, I.M., Schaeffer, A., Archer, M., Roughan, M., Griffin, D.A., Chapman, C.C., Sloyan, B.M., Everett, J.D., 2023. Frontal eddies provide an oceanographic triad for favorable larval fish habitat. Limnology and Oceanography 68, 1019–1036. doi:https://doi.org/10.1002/lno.12326.

Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B.B., Rashidi, P., Pardalos, P., Momcilovic, P., Bihorac, A., 2016. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. PLOS ONE 11. doi:10.1371/journal.pone.0155705.

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

Townsend, T., Barron, C., Helber, R., 2015. Ocean prediction with improved synthetic ocean profiles (isop). 2015 NRL Review .

Vukovich, F.M., 1988. Loop current boundary variations. Journal of Geophysical Research: Oceans 93, 15585–15591. doi:10.1029/JC093iC12p15585.

Wang, J.L., Zhuang, H., Chérubin, L.M., Ibrahim, A.K., Muhamed Ali, A., 2019. Medium-term forecasting of loop current eddy cameron and eddy darwin formation in the gulf of mexico with a divide-and-conquer machine learning approach. Journal of Geophysical Research: Oceans 124, 5586–5606. doi:https://doi.org/10.1029/2019JC015172.

Watts, D.R., Sun, C., Rintoul, S., 2001. A two-dimensional gravest empirical mode determined from hydrographic observations in the subantarctic front. Journal of Physical Oceanography 31, 2186–2209.

Zeng, X., Li, Y., He, R., 2015. Predictability of the loop current variation and eddy shedding process in the gulf of mexico using an artificial neural network approach. Journal of Atmospheric and Oceanic Technology 32, 1098 – 1111. doi:10.1175/JTECH-D-14-00176.1.

Zhang, A., Lipton, Z.C., Li, M., Smola, A.J., 2024. Multilayer perceptrons, in: Dive into Deep Learning. URL: https://d2l.ai/chapter_multilayer-perceptrons/mlp.html. retrieved October 8, 2024.

Zhang, Y., Yang, Q., 2017. A survey on multi-task learning. arXiv preprint arXiv:1707.08114 .

Declaration of interests

☐The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Robert W. Helber was part of the editorial for Deep-Sea Research 1 and 2. Subrahmanyam Bulusu was an associate editor for IEEE Geoscience Remote Sensing Letters. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Jose R. Miranda: Conceptualization; Methodology; Software; Formal analysis; Investigation; Resources; Visualization; Writing – original draft; Writing – review & editing.

Olmo Zavala-Romero: Conceptualization; Investigation; Project administration; Supervision; Writing – review & editing.

Luna Hiron: Investigation; Writing – review & editing.

Eric P. Chassignet: Supervision; Funding acquisition; Investigation; Writing – review & editing.

Bulusu Subrahmanyam: Investigation; Writing – review & editing.

Thomas Meunier: Investigation; Data curation; Writing – review & editing.

Robert W. Helber: Investigation; Data curation; Writing – review & editing.

Enric Pallas-Sanz: Investigation; Data curation; Writing – review & editing.

Miguel Tenreiro: Data curation.