A Deep Learning Approach to Estimate Ocean Salinity with Data Sampled with Expendable Bathythermographs

Edmo J.D. Campos[a,*], Cesar B. Rocha[a], Marlos Goes[b,c], Shenfu Dong[b], Hosmay Lopez[b], Gustavo J. Goni[b]

[a]*Instituto Oceanografico, Universidade de Sao Paulo (IOUSP)*
Praca do Oceanografico 191, Cidade Universitaria
Sao Paulo, 05508-120, SP, Brazil

[b]*Atlantic Oceanographic and Meteorological Laboratory (AOML*
National Oceanographic and Atmospheric Administration (NOAA)
4301 Rickenbacker Causeway, Miami, 33149, FL, USA

[c]*Cooperative Institute for Marine and Atmospheric Studies (CIMAS)*
University of Miami
4600 Rickenbacker Causeway, Miami, 33149, FL, USA

[*]Corresponding Author
Email address: edmo@usp.bf (Edmo J.D. Campos)

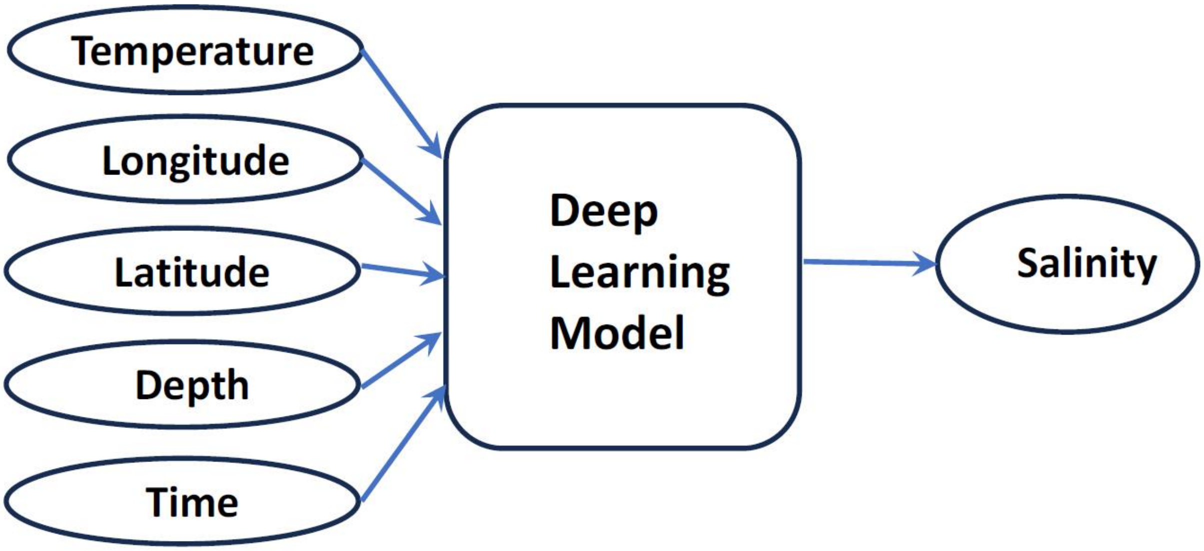
Preprint submitted to

Applied Ocean Research

Highlights

- Deep learning is a powerful tool for estimating the salinity fields associated with temperature sampled by bathythermographs.
- Additional information such as the latitude, longitude, depth, and month of the year of the sampling increases substantially the accuracy of the predicted salinity.

Graphical Abstract



Abstract

Expendable bathythermographs (XBTs) have made possible the build-up of an impressive dataset of temperature in the upper one thousand meters of the ocean. In the absence of direct measurements of salinity, the salinity profile associated with the temperature may be estimated by regression from temperature data or by objective analysis and data assimilation techniques. With the advent of the Argo program, the number of in situ salinity sampling has increased considerably. However, it is still far from ideal and XBTs continue to provide invaluable contribution to oceanography. Here, considering the large amount of data available, and motivated by the increasing use of machine learning to solve complex problems, a deep learning model based on a feed-forward neural network was used to estimate salinity directly from the temperature measurements. The model was independently trained and evaluated with two different datasets: (1) the sampled temperature and the associated salinity in XBT datasets, estimated with traditional methodologies; and (2) data sampled by conductivity-temperature-depth (CTD) profilers. The fitted deep learning model was then used to estimate the salinity from independent XBT datasets. The results show that the model's accuracy increases substantially when longitude, latitude, depth, and month of the samplings are considered. Compared with four traditional methods, the deep learning performed much better, particularly near the ocean's surface. In general, the results are highly accurate. However, when the CTD-trained model is used to predict salinity with XBT temperature input, the estimates are less accurate when compared with the salinity in the original XBT data. This seems to be caused by poorer estimates of salinity obtained by other methods in the original XBT datasets.

Keywords:

XBT, salinity, neural network, machine learning, deep learning

1. Introduction

The expendable bathythermograph (XBT) is a small, torpedo-shaped instrument that is dropped into the water from a ship. As it falls through the water, it measures the temperature and depth, and transmits the data back to a recorder in the vessel. XBTs have been used for several decades and have been proven to be an invaluable tool in the measurement of temperature in the upper layers of the ocean. The ease of use and low cost of XBT-based surveys have made possible the build-up of a large dataset used for inferring ocean heat content and dynamic properties, such as dynamic height, geostrophic velocity, and sound speed (Abraham et al., 2013; Cheng et al., 2016; Goni et al., 2019). However, for the proper estimation of such properties, the corresponding salinity profiles are essential.

Traditional in situ observations of oceanic variables are highly expensive and scarce, and sometimes logistically difficult to obtain. With satellite remote sensing, the knowledge of surface variables was greatly enhanced. More recently, the development of the Argo floats, autonomous platforms for profiling properties from the surface to deeper regions, provided a significant increase in the capability to observe the ocean. However, these floats are still unable to provide measurements with adequate spatiotemporal coverage for global studies without highly complex and expensive international efforts. Studies show that using Argo and XBT data jointly, rather than separately, improves the salinity estimates of boundary currents and meridional heat transport (Goes et al., 2018).

Contrary to temperature, the sampling of the water-column salinity is not an easy task. It commonly requires the use of more complex instruments, such as the conductivity-temperature-depth (CTD) profiler, during expensive oceanographic surveys. This was particularly true before the Argo period. Thus, due to the inherent difficulty of in situ measurements, salinity data are much sparse compared to temperature data. Different methodologies were developed to infer salinity from temperature data, taking advantage of the close relationship that holds between temperature and salinity in most of the ocean waters (e.g., Stommel, 1947; Emery and Dewar, 1982; Hansen and Thacker, 1999; Thacker, 2008; Thacker and Sindlinger, 2007; Goes et al., 2018). In addition to the empirical relationship between salinity and temperature, some approaches have included other predictors, such as the spatial dependence of the temperature-salinity (TS) relationship. For instance, Thacker (2008) showed that the inclusion of longitude and latitude improves the results based only on the TS relationship, with a quadratic dependence on temperature and linear on longitude and latitude. The salinity field can also be estimated by objective analysis of coarse in-situ observations (Chang et al., 2014), or by data assimilation techniques combining observations and numerical models (Dorfschäfer et al., 2020).

In the recent years, with the increasing power of computers and the availability of immense amounts of data, artificial intelligence (AI) has been increasingly used to solve complex problems in all fields of science. In oceanography, AI can be applied to a variety of tasks, including remote sensing, ocean modeling, coastal hydrodynamics, wave predictions, and data inference and analysis (Abbas et al., 2024; Xu et al., 2023; Berlinghieri et al., 2023; Haddad et al., 2022; Zhang et al., 2022; Chen and Xie, 2020; Li and Zhan, 2019; Chakraborty and Gangopadhyay, 2018; Guo and Huang, 2020; Gao and Li, 2018). Combining surface data with the scarcer in situ observations, AI can be used in a variety of applications in oceanographic studies. For instance, using a model constructed as a stack of long short-term memory (LSTM) neural network, Nardelli (2020) obtained vertical profiles from remotely sensed surface properties, which outperformed reconstructions from simpler statistical algorithms. Tian et al. (2022) applied a feed-forward neural network model to reconstruct the subsurface salinity distribution by merging in situ observations with satellite-sensed altimetry, sea surface temperature (SST), sea surface wind and coarser resolution gridded salinity products. Machine learning (ML) can be used in a variety of other ocean related applications such as prediction of wave patterns and significant wave height (Abbas et al., 2024), vulnerability of coastal bridges to natural hazards (Xu et al., 2023); to reconstruct of sea surface properties (Denvil-Sommer et al., 2019; Zhang et al., 2022; Berlinghieri et al., 2023); to fill in gaps in time series (Vieira et al., 2020; Haddad et al., 2022); or to infer information at small and fast turbulent scales, usually parameterized due to the limited observational sampling rates or model resolutions (Bolton and Zanna, 2019).

In general, AI is being increasingly used to take better advantage of all available observations. In particular, due to the sparsity of in situ salinity observations, ML models could be applied to provide more accurate estimates of the three-dimensional salinity profiles. In this regard, the work carried out by Tian et al. (2022) is a relevant example. They used a fully connected feed-forward neural network composed by an input layer, a stack of hidden layers, and an output layer. The input included the variables longitude, latitude, time, depth, surface properties, and gridded salinity. The output was the corresponding subsurface salinity field. The robust results, showing that the model could effectively transfer small-scale spatial variations in the input fields to the estimated salinity, confirm that ML approaches can provide an effective alternative for the estimation of the subsurface salinity, as a complement to the existing methods.

The successful results of Tian et al. (2022) motivated the development a machine learning model for the estimation of subsurface salinity corresponding to the temperature sampled by XBTs, using ancillary data such as temperature and salinity from CTD profilers and ARGO floats, and satellite sea surface temperature (SST) and sea surface height (SSH). As a first step, in this manuscript only XBT and CTD data are considered. The contribution of other data, such as SSH, SST and Argo floats, will be included in a future work.

The remaining of this work is subdivided in the following way. Section 2 contains a description of the data, including results of an exploratory data analysis, and a description of the methodology. Further details of the model's implementation and tune-up are given in Appendix A. The training and validation are described in Section 3. In Section 4 the model's predictions are discussed and, finally, Section 5 presents a concise summary of the results and the conclusions.

2. Data and Methods

Two independent datasets were used in this work: XBT data, containing sampled temperature and salinity estimated by different approaches, and CTD data, with both sampled temperature and salinity.

2.1. The XBT Data

Data from three XBT transects (AX08, AX18 and AX25), conducted by the Ship of Opportunity Program (SOOP), of the National Oceanographic and Atmospheric Administration (NOAA) Atlantic Oceanographic and Meteorological Laboratory (AOML), were downloaded from the SOOP repository at AOML. AX08 extends from Cape Town to New York and AX25 goes from Cape Town to Antarctica. AX18 forms a cluster of lines between Cape Town and eastern South America (Fig. 1). The downloaded datasets contain a total of 43564 archives with the vertical profiles of the temperature sampled in the upper 800 meters. The archives also contain salinity profiles, which according to the metadata, are monthly climatological salinity data from the 0.25-degree NOAA's World Ocean Database 2013 (WOD13), interpolated to the profile location. Additional information in the SOOP datasets is the longitude, latitude, depth, and the date of each XBT launching. The timeline for the three transects spans from the early 2000s to 2020. Although condensed around the nominal direction of the transects, they form a relatively dense coverage in the region between 30°S and 40°S, from South America to Africa, which will be the study area of the present work.

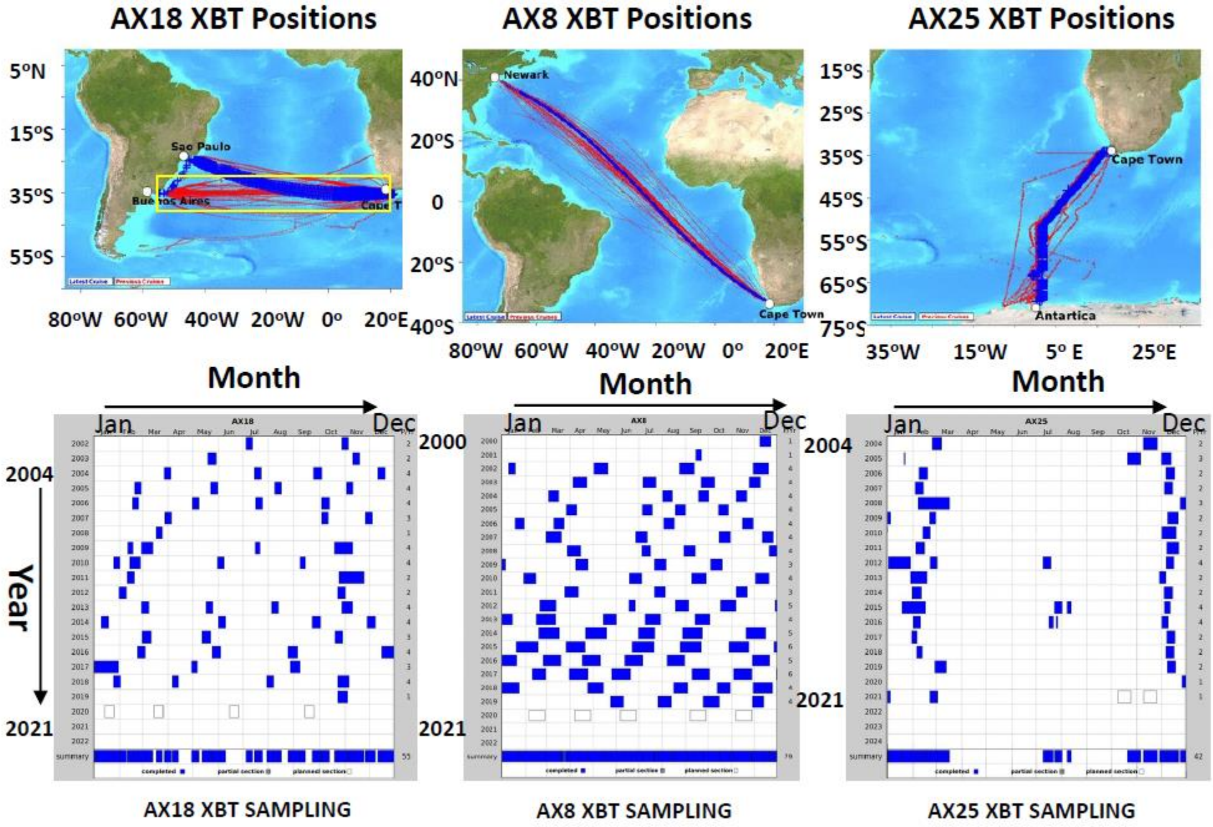


Figure 1: Location and time distribution of the AX08, AX18, and AX25 XBT lines.

2.2. The CTD Data

The CTD data used was obtained from the World Ocean Database 2018 (WOD18; (Boyer et al., 2018)), for the same region of the South Atlantic (30°S-40°S), from 1971 to 2021. The dataset, constituted by 16755 profiles, was downloaded from the NOAA's National Centers for

Environmental Information (NCEI). The dataset contains several ocean variables but only temperature and salinity were considered in this work.

2.3. Data Preparation and Exploratory Analysis

The datasets downloaded from AOML and NCEI are composed of individual files for each XBT or CTD profiles. The spatial dimensions in each file are longitude, latitude, and depth. Because of the relatively high rate in which the samples were recorded, the depth intervals are of the order of one meter or less, and different from one file to another. To reduce the number of depths and to compatibilize them with the standard oceanographic databases, the data at each station were binned to subsets of the Levitus 102 extended standard depths. For the XBT data, 41 vertical levels were selected, starting at 5 m down to 750 m. For the CTD data, 96 depths from 5 m to 5000 m were considered. This method introduced missing depths at stations shallower than 800 m for the XBT or shallower than 5000 m for the CTD stations. In a second step, the missing data in the binned datasets were removed. The method also detected the presence of some outliers or strange values, such as negative salinity. To avoid any undesirable values, stations located outside the study area, and data with salinity and temperature outside the ranges 31-38°C and 0°C-30°C, respectively, were discarded. In spite of the average salinity range for the Atlantic being 33-37, the minimum value 31 was chosen to allow points sampled in the low salinity waters near the western boundary (Region I, Fig 2). After that, the data files were merged into two single datasets, one for all XBT and another for all CTD stations. To maintain compatibility with the XBT data, only the 45 first levels (from 5 m to 750 m) of the CTD data will be considered from now on.

In continuation, the statistics of the two datasets were calculated (Tables 1 and 2), the geographic location of each station in the filtered data was plotted, and the temperature and salinity used to construct TS-diagrams (Fig. 2). The color scheme in Fig. 2, with a continuous blue-green-yellow pallet, reflects the longitudinal position of the stations, from west to east. The statistics show that there are approximately twice as many data points in the XBT than in the CTD dataset. The horizontal distributions of the stations are quite different, as seen in the lower panel of Fig. 2. Near the continental boundaries, regions I and III in the figure, the concentration of stations is similar in both XBT and CTD datasets. However, in the interior (region II), the number of XBT points is much larger. This is somewhat expected since the more expensive CTD surveys are more likely to be conducted in the vicinities of the continents.

With the color scheme adopted in the TS-diagrams, it is possible to distinguish the different water masses according to their longitudinal location. The major differences in the two diagrams are mainly due to the different concentrations of sampling points near South America. As shown in the upper panels of Fig. 2, there are much more CTD stations in the southwestern corner of the domain, to the west of 55°W, than XBT launching points. That region is dominated by the fresher and cooler waters of the northward flowing Malvinas Current, which are not captured by the XBTs. In the TS-diagram with CTD data it is also seen the presence of relatively warm and very fresh waters (salinity lower than 32), which are most likely to be waters from the Patagonian shelf and/or from the Rio de La Plata outflow. Based on the color scheme, it is also possible to see, in the upper-left panel of Fig. 2, waters with temperature above 25°C and salinity higher than 36.5. These are possibly Tropical Waters (TW) transported by the Brazil Current.

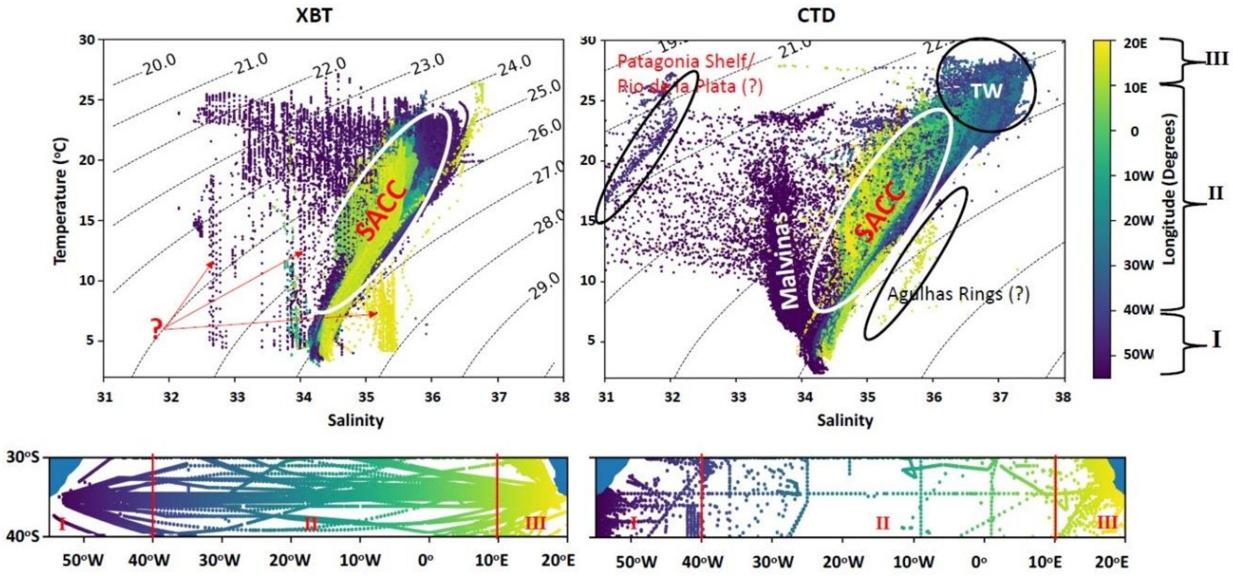


Figure 2: Top: TS-diagrams constructed with XBT and CTD in the latitudinal band 30°S-40°S from 55°W to 20°E. The question mark in the left panel indicates regions with possible wrong values of salinity. Bottom: spatial distribution of the XBT and CTD stations around SAMBA.

Table 1: Statistics of XBT data. Level is in meters, temperature is in degrees Celsius, and salinity is in salinity units.

	Count	Mean	Std	Minimum	Maximum
Level	487470	222.9	211.9	5.0	750.0
Temperature	487470	14.5	4.3	3.0	27.1
Salinity	487470	35.23	0.46	32.15	36.86

Table 2: Statistics of CTD data. Level is in meters, temperature is in degrees Celsius, and salinity is in salinity units.

	Count	Mean	Std	Minimum	Maximum
Level	215258	134.6	159.9	5.0	750.0
Temperature	215258	13.5	4.2	2.4	25.3
31.00	215258	35.03	0.58	31.00	37.33

Another important information that can be inferred from the XBT TS-diagram in Fig. 2 (upper-left) is the likelihood of errors in the climatological salinity estimates in the XBT dataset. There are regions with TS points distributed along vertical straight lines, in which the temperature varies several degrees for constant salinity. This seems to be more common in areas with mixed characteristics. They are probably spurious results of the methodology used to estimate the salinity from climatology in the XBT datasets, since using mean climatological salinity does not maintain the TS properties of the water, as also stated in Goes et al. (2018).

Near the eastern end of the domain (region III, yellow colors), it is possible to see in the CTD TS-diagram some anomalous waters, with a jump of almost 1 in salinity in the 27.0-28.0 density (σ_t) range . This is likely to be associated with Agulhas Rings, which transport into the South Atlantic waters with higher salinity from the Indian Ocean. This appears to be present in the XBT TS-diagram as well, in spite of the seemingly spurious nature of the data, with vertically oriented sequences of points.

2.4. Methodology: The Deep Learning Model

Deep learning is a subfield of machine learning in which the model learns representations from data by means of successive layers of increasingly meaningful representations. Here, a deep learning (DL) model based on a feed-forward neural network (FFNN) was developed using Tensorflow/Keras, a high-level neural network Python library (Chollet, 2021). FFNN is an artificial neural network in which the information flows forward through a succession of layers, in one single direction. It is composed of an input layer, intermediate hidden layers, and a final output layer. In this study, the FFNN was constructed with a structure similar to the models used by Tian et al. (2022) and Denvil-Sommer et al. (2019). See Appendix A for more information on the machine learning approach.

2.5. Hyperparameter Optimization

Among the many hyperparameters in a neural network, some of the most important are: the learning rate, the number of layers, the number of nodes or size of each layer, the activation functions, and the batch size. There are no formal rules on how to set up the proper combination of these hyperparameters. In general, an experienced modeler can define some good first guesses. However, it is important to subjectively redefine these choices by tweaking them until getting to more appropriate values. In this work we used a combination of repeated training and the use of a KerasTurner model building function, as recommended by Chollet (2021), based on the Hyperband algorithm (Li et al., 2018). The hyperparameter search suggested the number of 512 units for the first densely-connected layer and the value of 0.001 for the optimizer's (Keras.RMSprop) learning rate. Additionally, based on a number of runs, we set to five the number of hidden layers, with the following number of nodes: 512, 256, 128, 64, and 16, with RELU as the activation function.

The batch size and the number of iterations can affect the performance of deep neural networks. After testing several combinations, we decided to use a batch size of 1024 and 1000 epochs for the estimations analyzed in this article. These estimates can be reproduced with the data and scripts provided with the accompanying supplementary material.

3. Training, Validation and Fitting the Model

With the code free of syntax and logic errors, some exploratory experiments were carried out with both the XBT and CTD data to find a setup capable of making accurate estimates of the salinity field based on sampled temperature. For these experiments, the variables temperature, salinity, and level were centered to zero-mean and reduced to unit standard deviation by subtracting the mean value and dividing by the standard deviation. Because it is highly recommendable to normalize all the input data, the longitude and latitude were replaced by *sin*(Longitude ×π/180) and *sin*(Latitude ×π/180), respectively. The time was transformed into a date string and the number representing the month was selected (i.e: 1 for January 2 for February and so on). Then, the whole dataset was rearranged according to the month. With this time-transformation, only the intra-annual variability of the data was retained. Following, the month of the year in the time variable was also demeaned and standardized. After having all variables conveniently prepared, each dataset was then divided into a train set, containing 80% of the samples and a test set, with the remaining 20% of the record. The train was further divided into a partial train subset, with 80% of train, and a val set with the remaining 20% of train.

3.1. Choice of Optimizer and Validation Accuracy

To evaluate the effects of using different metrics in the model validation, we compiled the model with both the root mean squared error (MSE) and mean absolute error (MAE) as validation error. Also, to assess the effect of different optimizers, we tried RMSprop and Adam. After 100 iterations, the MAE and MSE were similar for the different optimizers: 0.105 and 0.028 for RMSprop, and 0.103 and 0.030 for Adam. However, for each optimizer, the final MAE was higher than MSE. Based on these results, shown in Fig. 3, we decided to use MAE for validation error and the RMSprop as the optimizer.

3.2. The Final Model Setup

Several combinations of parameters where tested, until a final setup was reached with the following configuration:

- five hidden layers with 512, 256, 128, 64 and 32 nodes;
- the rectified linear unit (RELU) as activation function;
- the root-mean-square propagation (RMSProp) as optimizer;
- loss estimation by the mean root squared error (MSE); and
- the mean absolute error (MAE) as the validation metric.

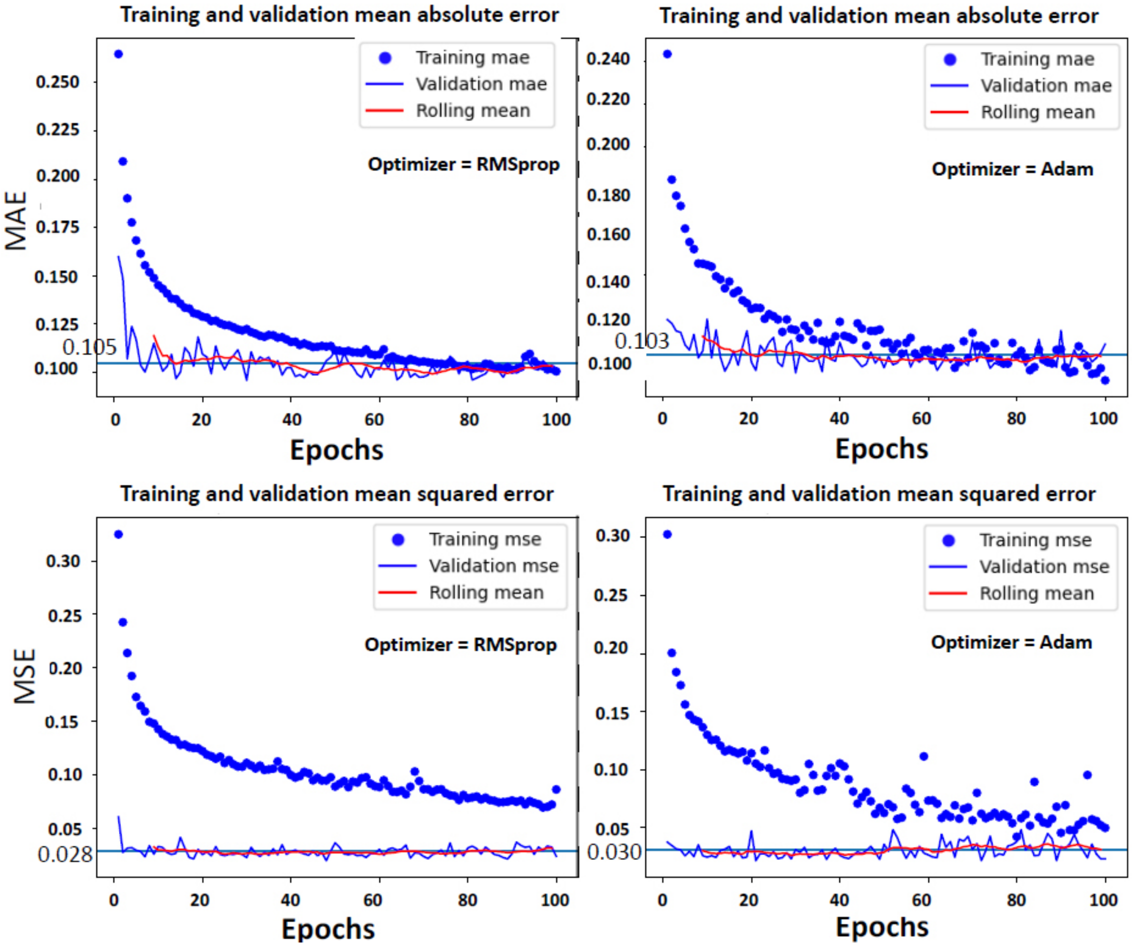


Figure 3: Training and validation MAE (top) and MSE (bottom) after 100 iterations using RMSprop (left) and Adam (right) optimizers.

3.3. Check for Overfitting

The canonical overfitting curves for the CTD and XBT data with final setup considering all variables (temperature, longitude latitude, level, and time) as input, after 1,000 iterations, are shown in Fig. 4. For both cases, the train loss decays in an exponential-like fashion, tending to a minimum value of less than 0.05. However, the validation loss is quite different between the two datasets. The val loss for the CTD data starts and remains less than 0.05 for the entire training time, with relatively small standard deviation. For the XBT data, the val loss starts at around 0.15, increases slightly and stays with an almost constant mean value below 0.20. The variability is much higher in the val loss. Despite the relatively high mean value, the validation error does not increase during the whole period (1000 epochs), suggesting a robust fit.

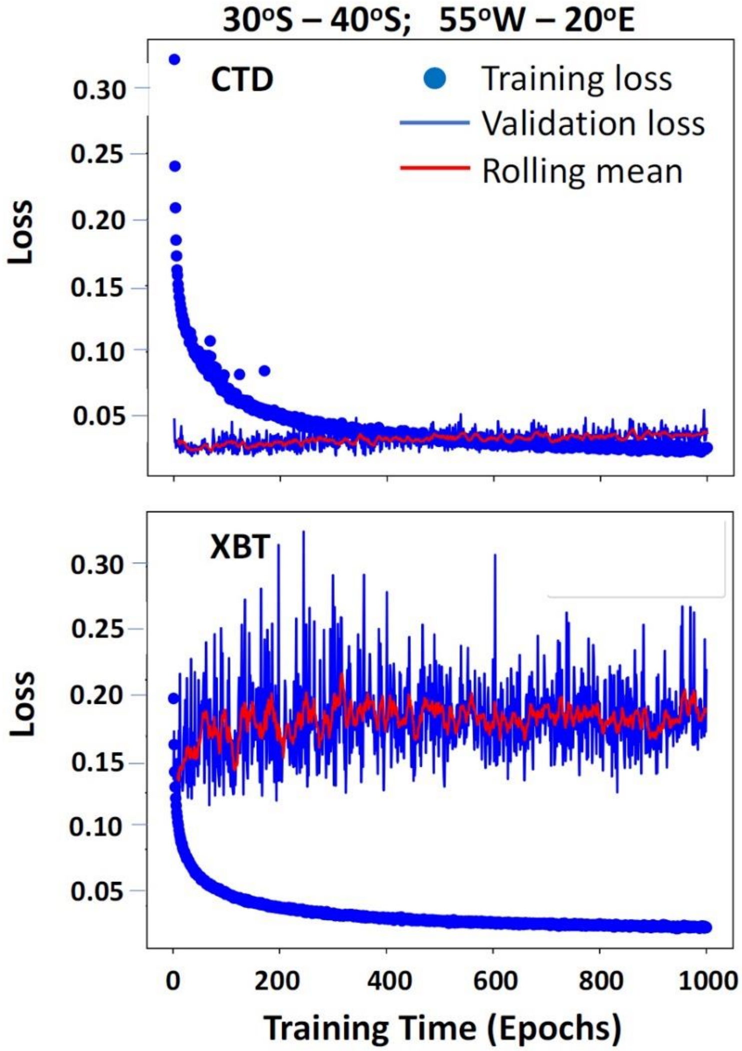


Figure 4: Training and validation loss versus training time, for the CTD (top) and XBT (bottom) datasets.

4. Results

4.1. Predictions

After having the model properly trained and validated, with both XBT and CTD datasets, the subsurface salinity field corresponding to the respective test datasets was predicted in the following cases:

(1) Prediction of the salinity corresponding to the CTD test with the model trained with the CTD train dataset.

(2) Prediction of the salinity corresponding to the XBT test with the model trained with the XBT train dataset.

(3) Prediction of the XBT salinity with the CTD trained model.

This sequence of experiments was motivated by the suspicion of some spurious values of the salinity estimates in the XBT datasets, as discussed in Section 2. The question was, how good are the predictions made with a model trained with a dataset containing suspicious information? To help answer this question, the mean absolute error was calculated using the redimensionalized variables, in order to have MAE expressed in salinity units.

The results are shown in Fig. 5. The left panels show the evolution of the absolute error over the 1000 iterations. On the right, the TS-diagrams constructed using the target salinity (the salinity in the test dataset), in blue, and the predicted salinity, in red color. In the legends of the left panels, the value labeled as non-dimensional MAE is the absolute error averaged over the last 50 iterations. Because the model uses demeaned and standardized variables, this error is also given in normalized units. The target and the predicted salinity were multiplied by the standard deviation of the train dataset, in order to get the error in physical units. In Fig. 5 this is indicated by appending PSU to the dimensional error. In all cases, the dimensional errors are similar and considerably low: 0.021 salinity units for case (1) and 0.031 and 0.024 for the cases (2) and (3), respectively. The good accuracy is confirmed by the almost perfect match between the target and predicted pairs T, S in the TS-diagrams. It means that, in all cases, the model does a good job in predicting the salinity. As shown by middle panels of Fig. 5, the largest error was found in case (2), where the model trained with the XBT data tries to reconstruct the same values of the salinity profile that are likely to be wrong. This should not be a surprise. After all, the model has learned how to correlate the salinity with the train input variables, no question asked if they are right or wrong. The effect of the spurious salinity on the model’s accuracy, the model was applied another XBT data (Goes et al., 2018), in which most of the spurious values had been removed. The dimensional MAE of the predictions with the model trained with XBT train using XBT test (Case 2), fell from 0.031 to 0.027 PSU.

4.2. Effects of Input Variables

In the ocean, the salinity depends primarily on the temperature, but the time of the year, the depth, and the geographic location of the sample are also relevant. In order to evaluate the effects of each of the input variables, a sequence of predictions was conducted with the model trained with different subsets of the CTD input vector. First, only temperature was considered. Then, other variables, in addition to temperature, were included as follows (Fig. 6):

(1) Temperature (T) only

(2) T and time (t)

(3) T and Level (L)

(4) T, t and L

(5) T, longitude (ϕ), and latitude (θ)

(6) T, t, L, ϕ, and θ

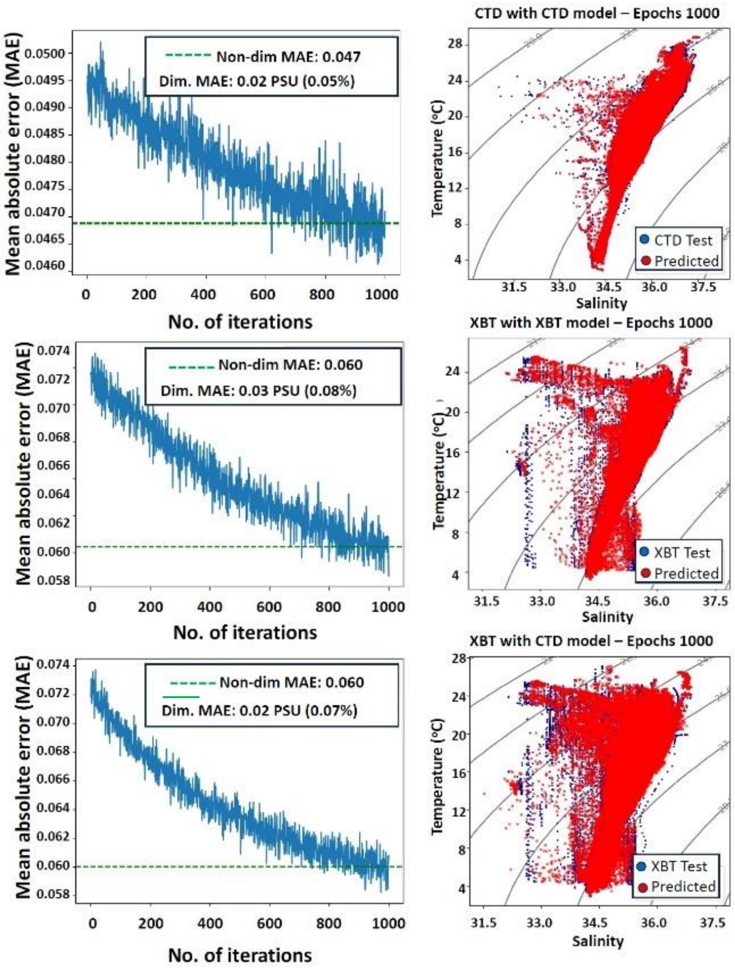


Figure 5: Mean absolute error versus time and TS-diagrams constructed with test and predicted values for the CTD and XBT.

In case (1), when the input is only temperature, the prediction is the poorest, with non-dimensional MAE = 0.24. In this case, the predicted TS-diagram (red dots) appears to simply reproduce the mean values of the T,S pairs in the T-S space. The inclusion of the month of the year, case (2), improves the prediction a little, with the MAE reduced to 0.18. The red dots now cover a larger area as compared to the previous case. The depth of the measurement seems to be more relevant than the time of the year the CTD sample was taken. In case (3), considering the temperature and level, the MAE falls to 0.16. The combined effect of time and depth (case 4) makes the error fall to 0.11. The location of the CTD station (longitude, latitude) appears to be the largest individual contribution to the model skill. In agreement with the results obtained by Thacker (2008), when only temperature, longitude and latitude are considered (case 5), the error is 0.10, the lowest as compared with the previous cases. Finally, as it should be expected, the best performance is obtained when all five variables are considered, with the non-dimensional MAE = 0.07 (case 6). It is important to mention here that these predictions were conducted with only 100 iterations. With this limited number of epochs, the model had not yet reached an optimum training. However, this does not invalidate the relative values obtained.

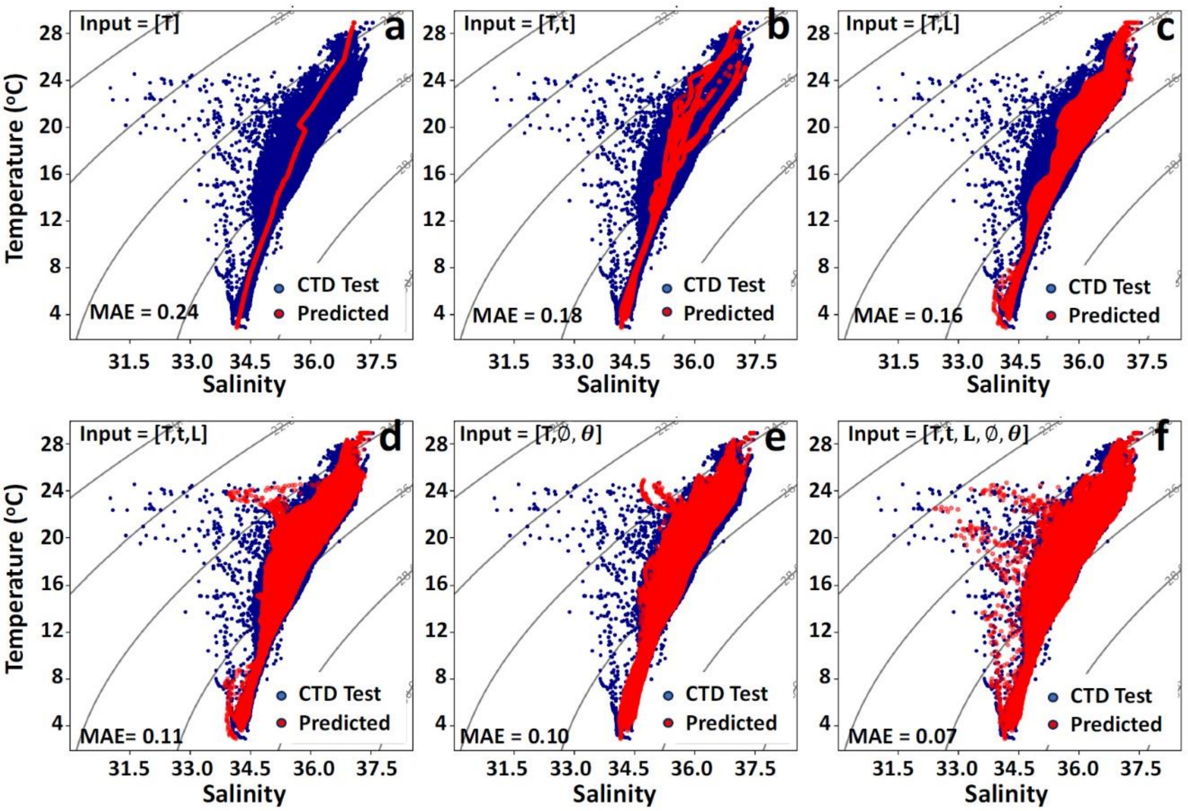


Figure 6: TS-diagrams constructed with data from the test dataset (blue) and the temperature from the test dataset and the predicted salinity (red), for different input variables: (**a**) temperature (T) only; (**b**) T and time (t); (**c**) T and Level (L); (**d**) T, t, and L; (**e**) T, longitude (ϕ), and latitude (θ); and (**f**) T, t, L, ϕ, and θ.

4.3. Comparison with Other Methods

Here, we compared our estimates using deep learning against other methods. Similar to Goes et al. (2018), we used the following methods in this comparison: Annual, Stommel, Thacker and RSEAS. The Annual method (Emery and Dewar, 1982) uses the mean annual climatology. The Stommel method (Stommel, 1947) uses the mean annual TS relationship. The Thacker method (Thacker, 2008) uses a linear regression of salinity at 85 levels against the predictors temperature, temperature squared, latitude, and longitude. The RSEAS method (Goes et al., 2018), a multivariate linear regression, was applied on 85 depth levels, using as regressors temperature, temperature squared, and the first and second harmonics, which were shown to reduce seasonal biases. This comparison was made on 2990 CTD data points, between 15°S and 40°S, which were not used for training. After estimating salinity with the five methods, residuals (model minus observations) were calculated, and the distribution of the residuals and the standard deviations (SD) for each depth were estimated.

The four methods used in this comparison against DL have similar performances in the upper ocean, with residuals reaching 0.3 in the top 50 m (Fig. 7a). From all methods, Annual is the one with the worst performance. The standard deviation in the Annual method is generally between 0.1 and 0.2 below 50 m. The Stommel method improves considerably over the Annual method, with SD ≈ 0.05 below 200 m, and increasing to about 0.2 at 50 m. The Thacker and

RSEAS methods have similar performance, with SD ≈0.02 below 100 m, with RSEAS slightly better and less noisy. The DL method outperforms the other methods, reaching SD ≈ 0.007 in the deeper regions, and improving the most near the surface levels. The distribution of residuals (Fig. 7b) shows that the DL method has a 90% of the salinity residuals fall under 0.1, in comparison to 0.2 for RSEAS and Thacker, and 0.3 for Stommel and Annual.

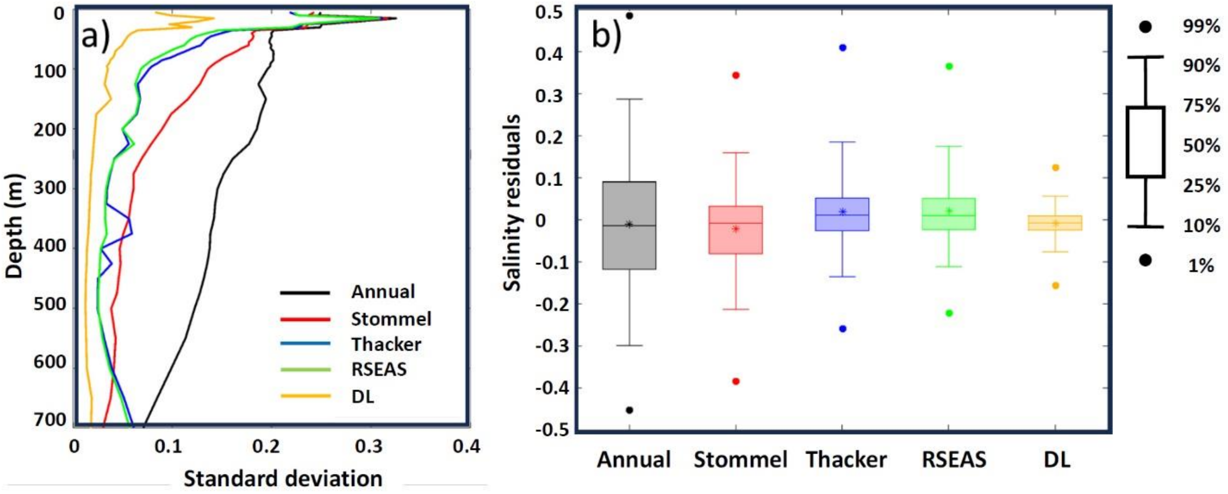


Figure 7: (a) Standard deviation for each depth and (b) distribution of all salinity residuals for the Annual (black), Stommel (red), Thacker (blue), RSEAS (green), and Deep Learning (orange) relative to 2990 CTD profiles between 15°S and 40°S.

5. Summary and Discussion

Data from XBT and CTD datasets, in the South Atlantic region between 30°S and 40°S, were used to train and validate a supervised deep learning model based on a feed forward neural network. The trained model was then used to reconstruct the salinity field corresponding to a subset of the XBT temperature data (test), previously unseen by the model. To evaluate the model's skill, the dimensional value of the predicted salinity was then compared with the target salinity (the salinity of the test subset). Both the mean absolute error (MAE) and the root mean squared error (MSE) were used in the model validation. In general, MAE was slightly higher that MSE and was adopted as the validation metric. Two different optimizers, RMSprop and Adam were tested, with the results not showing any significant differences. RMSprop was then used for the final experiments.

Three sets of experiments were conducted. In the first one, the model was trained with CTD data and then used to predict the salinity, using a test subset with CTD temperatures. In the second, the model was trained with XBT data and then used the XBT test for predictions. The third experiment used the CTD trained model to predict salinity with the test subset from the XBT temperature data. The TS-diagrams constructed with the predicted salinity matched quite well against the diagram constructed with the target salinity in the test datasets. The dimensional mean absolute error in the three experiments were, respectively, 0.021, 0.031 and 0.024 salinity units. These are small errors, confirming the good accuracy seen in the almost perfect match of the blue and red dots in the TS-diagrams. In the TS-diagrams, the main differences between the predicted and the target salinity are mostly in mixing regions. The relatively poorer performance of the XBT trained model used to predict salinity with the XBT test (MAE 0.031) can be explained by the presence of apparently spurious salinity in the XBT datasets. It is likely that, when trained with a dataset containing strange values, the model has more difficulty in learning the correlations between feature and target.

The contribution of the spatial and time dependence to the model's prediction skill was investigated. It was found that the longitude and latitude are the predictors that mostly improves the performance, followed by the depth and the month of the year of the sampling.

In a comparative analysis, our DL model outperformed four other methods, being more accurate than all the others in the entire water column. Our estimates are particularly better than the others in the upper levels of the ocean, where the dynamics are strongly affected by turbulent processes. This good performance of the DL, compared with the other methodologies, confirms that our model has a good generalization ability.

This was the first effort towards a machine learning model to estimate the salinity field corresponding to temperature sampled by expendable bathythermographs. In general, the results demonstrate that the machine learning approach is a powerful tool to reconstruct the salinity field given a sufficiently large set of known variables. The caveat is the availability of datasets with sufficient and reliable salinity fields for training the model. Here, data from some XBT lines and from CTD stations in a specific region were considered.

The next step will consider the inclusion of data from other sources, such as satellite sea surface temperature and height, surface drifters, Argo floats and gliders. The expansion of the methodology to be applied to entire XBT lines and the comparison with results obtained with other approaches will also be considered.

Acknowledgments

This work was partially supported by the National Oceanic and Atmospheric Administration (NOAA) Atlantic Oceanographic and Meteorological Laboratory, and carried out under the auspices of the Cooperative Institute for Marine and Atmospheric Studies, a cooperative institute of the University of Miami and NOAA (cooperative agreement number NA20OAR4320472), and by the SAMOC/SAMBAR Project, funded by the São Paulo State Foundation for Research Support (FAPESP, grant 2017/09659-6). The first author acknowledges a research fellowship from the Brazilian Council for Scientific and Technological Development (CNPq, 302503/2019-6).

Appendix A. Supplementary Information on the Deep Learning Approach

The deep learning model is composed of input and output layers and intermediate hidden layers. The hidden layers are dense, or fully connected layers. That is, each neuron is connected to each neuron of the previous and the following layers (Fig. A.1). The number of hidden layers can be one or more. The number of neurons in each hidden layer also varies. A dense layer can be interpreted as a function g that takes an input tensor $\mathbf{X}$ and returns another tensor $\widehat{\mathbf{Y}}$, as the output. This can be mathematically expressed as:

$$\widehat{\mathbf{Y}} = \mathrm{g}(\mathbf{W} \cdot \mathbf{X} + \mathbf{b}) \tag{A.1}$$

In Eq. A.1, g is the *activation function*, and $\mathbf{W}$ and $\mathbf{b}$ are the weights and biases, respectively. The activation function is a modeler's choice. The weights and bias are the *parameters* to be learned by the model. To have a model capable of learning from a set of known information and make accurate predictions based on new, independent data, the modeler also needs to learn how to decide for the best set of hyperparameters, which are number of layers or the size of the layers, and other conditions for constructing an efficient model. This was achieved by a long sequence of experiments in which different values of hyperparameters, learning rates and optimizers were tested.

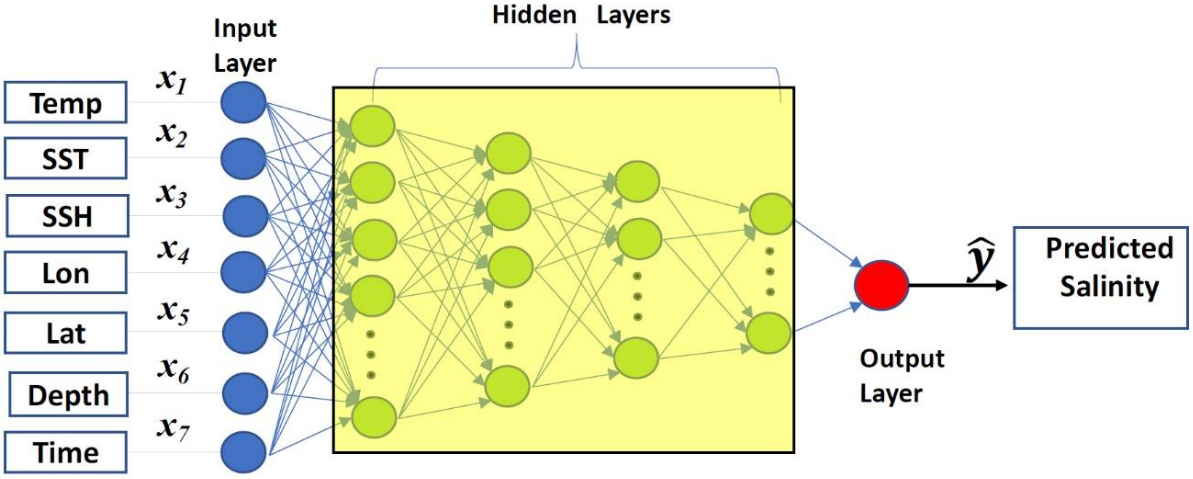


Figure A.1: Schematic of the FFNN model basic structure: the input layer, a set of intermediate hidden layers and the output layer (based on Fig. 1 of Tian et al., 2022).

Appendix A.1. Optimization vs Generalization

As described in the literature (Chollet, 2021), a fundamental issue in machine learning is how to optimize the model to get the best results in making predictions. The optimization, the process of fitting the model to the training data, is under the modeler's control. However, generalization, or how well the model performs during prediction, is not. A good fit to the training data does not necessarily mean a good generalization. Sometimes, overfitting could happen and the model starts to fit too well against its training data and fails to perform accurately against unseen data.

In machine learning, the original data is usually split in two parts. The train set, used for training the data and the test set, for the predictions. One way to look for overfitting is to break the train in two pieces, one to be used for the actual training, and another to evaluate the training process. Then, compute and plot the loss (*J*), defined by Eq. A.2, versus the training time (Fig. A.2). In Eq. A.2, θ_n are model parameters (number of layers, number of nodes in each layer, etc.); *X* is the input matrix; *y* is the output; *n* is the number of parameters; *m* the number of variables in the input; and h_θ is the model for training. The training objective is to reach a minimum value for *J*.

$$J(\theta_0, \theta_1, \dots \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} [h_\theta \; (X^{(i)} - y^{(i)})]^2 \tag{A.2}$$

In the canonical overfitting behavior (Fig A.2), losses for both train and predictions (validation) are correlated at the beginning. After a certain number of iterations (training time), the validation loss starts to increase, indicating a degradation in the predictions in spite of the continuing decrease in the training loss (overfitting). The idea is to train the model until it starts to overfit, and then stop.

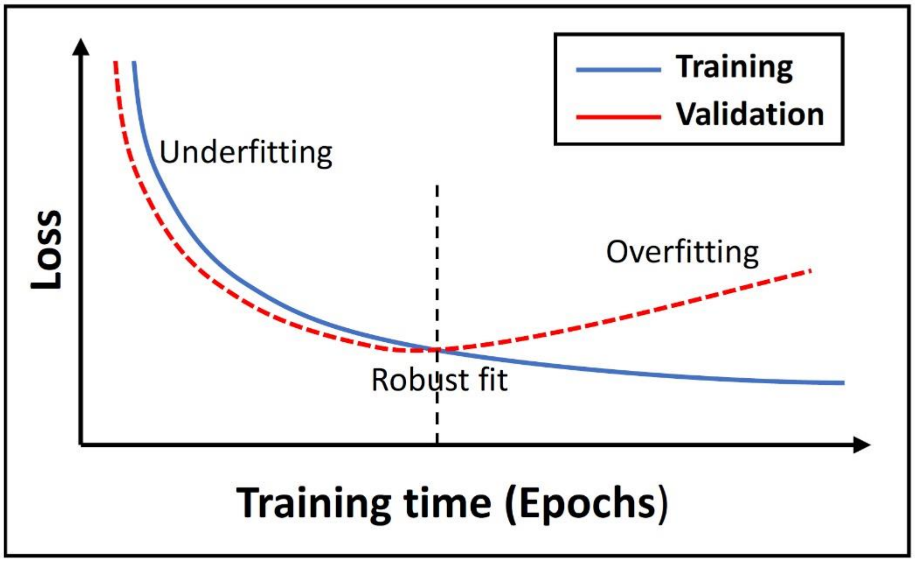


Figure A.2: Canonical overfitting behavior (based on Fig. 5.1 of Chollet, 2021).

References

Abbas, M., Min, Z., Liu, Z., Zhang, D., 2024. Unravelling oceanic wave patterns: A comparative study of machine learning approaches for predicting significant wave height. Applied Ocean Research 145, 103919.

Abraham, J., Baringer, M., Bindoff, N., Boyer, T., Cheng, L., Church, J., Conroy, J., Domingues, C., Fasullo, J., Gilson, J., Goni, G., Good, S., Gorman, J., Gouretski, V., Ishii, M., Johnson, G., Kizu, S., Lyman, J., Macdonald, A., Minkowycz, W., Moffitt, S., Palmer, M., Piola, A., Resenghetti, F., Shuckmann, K., Trenberth, K., Velicogna, I., Willis, J., 2013. A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change. Review of Geophysics 51, 450–483, doi:10.1002/rog.20022.

Berlinghieri, R., Trippe, B.L., Burt, D.R., Giordano, R., Srnivasan, K., Ozgokmen, T., Xia, J., Broderick, T., 2023. Gaussian Processes at the Helm(holtz): A more fluid model for ocean currents . Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii., arXiv:2302.10364 [stat.ME], doi: 10.48550/arXiv.2302.10364.

Bolton, T., Zanna, L., 2019. Applications of deep learning to ocean data inference and subgrid parameterization. Journal of Advances in Modeling Earth Systems 11, 376–399, https://doi.org/10.1029/2018MS001472.

Boyer, T., Baranova, O., Coleman, C., Garcia, H., Grodsky, A., Locarnini, A. Mishonov, A., paver, C., Reagan, J., Seidov, D., Smolyar, I., Weathers, K., Zweng, M., 2018. World ocean database 2018. a.v. mishonov, technical ed. NOAA Atlas NESDIS 87, https://www.ncei.noaa.gov/sites/default/files/2020–04/wod intro 0.pdf.

Chakraborty, S., Gangopadhyay, A., 2018. Deep learning in oceanography: A review. Ocean Engineering 152, 365–378, doi: 10.1016/j.oceaneng.2018.01.062.

Chang, Y.S., Rosati, A.J., Zhang, S., Harrison, M.J., 2014. Objective analysis of monthly temperature and salinity for the world ocean in the 21st century: Comparison with World Ocean Atlas and application to assimilation validation. Journal of Geophysical Research 114, C02014, doi:10.1029/2008JC004970.

Chen, G., Xie, J., 2020. Deep learning for ocean data analysis: A Review. Journal of Oceanography 76(2), 59–78, doi:10.1007/s10872–019–00502–z.

Cheng, L., Zhu, J., Cowley, R., Boyer, T., Wijffels, S., 2016. XBT science: Assessment of instrumental biases and errors. Bulletin of the American Meteorological Society 97, 924–933, https://10.1175/BAMS–D–15–00031.1.

Chollet, F., 2021. Deep Learning with Python . Manning Publications, New York, NY 2nd Ed., ISBN: 9781617296864.

Denvil-Sommer, A., Gehlen, M., Vrac, M., Mejia, C., 2019. LSCE-FFNN-v1: A two-step neural network model for the reconstruction of surface ocean $p\mathrm{CO}_2$ over the global ocean . Geoscience Model Development 12, 2091–2105, https://doi.org/10.5194/gmd–12–2091–2019.

Dorfschäfer, G., Tanajura, C., Costa, F., Santana, R., 2020. A New Approach for estimating salinity in the southwest Atlantic and its application in a data assimilation evaluation experiment. Journal of Geophysical Research-Oceans 125(19), e2020JC016428, https://doi.org/10.1029/2020JC016428.

Emery, W., Dewar, J., 1982. Mean temperature-salinity, salinity-depth and temperature-depth curves for the North Atlantic and the North Pacific. Progress in Oceanography 11, 219–305. doi:10.1016/0079–6611(82)90015–5.

Gao, J., Li, X., 2018. Deep learning for ocean remote sensing: An overview and future direction. Remote Sensing of Environment 216, 126–143, https://doi:10.1016/j.rse.2018.06.012.

Goes, M., Christophersen, J., Dong, S., Goni, G., Baringer, M., 2018. An updated estimate of salinity for the Atlantic Ocean sector using temperature–salinity relationships. Journal of Atmospheric and Oceanic Technology 35, 1771–1784, https://doi.org/10.1175/JTECH–D–18–0029.1.

Goni, G.J., Sprintall, J., Bringas, F., Cheng, L., Cirano, M., Dong, S., Domingues, R., Goes, M., Lopez, H., Morrow, R., Rivero, U., Rossby, T., Todd, R.E., Trinanes, J., Zilberman, N., Baringer, M., Boyer, T., Cowley, R., Domingues, C.M., Hutchinson, K., Kramp, M., Mata, M., Reseghetti, F., Sun, C., TVS, U.B., Volkov, D., 2019. More than 50 years of successful continuous temperature section measurements by the global expendable bathythermograph network, its integrability, societal benefits, and future. Frontiers in Marine Science 6, 452, https://doi.org/10.3389/fmars.2019.00452.

Guo, D., Huang, B., 2020. Deep learning in oceanography. A review. Journal of Oceanography 76(4), 225–249, doi: 10.1007/s10872–020–00516–5.

Haddad, S., Killick, R.E., Palmer, M.D., Webb, M.J., Prudden, R., Capponi, F., Adams, S.V., 2022. Improved infilling of missing metadata from expendable bathythermographs (XBTs) using multiple machine learning methods . Journal of Atmospheric and Oceanic Technology 39(9), 1367–1385, DOI: 10.1175/JTECH–D–21–0117.1.

Hansen, D.V., Thacker, W.C., 1999. Estimation of salinity profiles in the upper ocean . Journal of Geophysical Research 104(C4), 7921–7933.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A., 2018. Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research 18, 1–52. URL: http://jmlr.org/papers/v18/16-558.html.

Li, Y., Zhan, P., 2019. Deep learning for ocean modeling: A review. Journal of Oceanography 75(4), 343–360, doi: 10.1007/s10872–019–00501–0.

Nardelli, B.B., 2020. A deep learning network to retrieve ocean hydrographic profiles from combined satellite and in situ measurements . Remote Sensing 12, 3151, doi:10.3390/rs12193151.

Stommel, H., 1947. Note on the use of the T-S correlation for dynamic height anomaly computations. Journal of Marine Research 6, 85–92.

Thacker, W., 2008. Estimating salinity between 25° and 45°S in the Atlantic Ocean using local regression. Journal of Atmospheric and Oceanic Technololgy 25, 114–130, doi:10.1175/2007JTECH0530.1.

Thacker, W., Sindlinger, L., 2007. Estimating salinity to complement observed temperature: 2. Northwestern Atlantic . Journal of Marine Systems 65(1-4), 249–267. https://doi.org/10.1016/j.jmarsys.2005.06.007.

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, https://doi.org/10.5194/essd–14–5037–2022.

Vieira, F., Cavalcante, G., Campos, E., Taveira-Pinto, F., 2020. A methodology for data gap filling in wave records using artificial neural networks. Applied Ocean Research 98, 102109, https://doi.org/10.1016/j.apor.2020.102109.

Xu, G., Ji, C., Xu, Y., Yu, E., Cao, Z., Wu, Q., Lin, P., Wang, J., 2023. Machine learning in coastal bridge hydrodynamics: A state-of-the-art review. Applied Ocean Research 134, 103511, doi: 10.1016/j.apor.2023.103511.

Zhang, H., Huang, B., Chen, G., Ge, L., Radenkovic, M., 2022. An efficient oceanic eddy identification method with XBT data using Transformer . IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 9860–9872, doi: 10.1109/JSTARS.2022.3221113.