PPCon 1.0: Biogeochemical Argo Profile Prediction with 1D Convolutional Networks

Gloria Pietropolli[1, 2], Luca Manzoni[1, 2], and Gianpiero Cossarini[1]

[1]National Institute of Oceanography and Applied Geophysics - OGS, Trieste, Italy.
[2]Dipartimento di Matematica e Geoscienze, University of Trieste, Trieste, Italy.

Correspondence: Gloria Pietropolli (gloria.pietropolli@phd.units.it)

Abstract. Effective observation of the ocean is vital for studying and assessing the state and evolution of the marine ecosystem, and for evaluating the impact of human activities. However, obtaining comprehensive oceanic measurements across temporal and spatial scales and for different biogeochemical variables remains challenging. Autonomous oceanographic instruments, such as Biogeochemical (BCG) Argo profiling floats, have helped expand our ability to obtain subsurface and deep-ocean measurements, but measuring biogeochemical variables such as nutrient concentration still remains more demanding and expensive than measuring physical variables. Therefore, developing methods to estimate marine biogeochemical variables from high-frequency measurements is very much needed. Current Neural Network (NN) models developed for this task are based on a Multilayer Perceptron (MLP) architecture, trained over ~~punctual~~ point-wise pairs of input-output features. However, MLPs lack awareness of the typical shape of biogeochemical variable profiles they aim to infer resulting in irregularities such as jumps and gaps when used for the prediction of vertical profiles. In this study, we ~~evaluate the effectiveness of a~~ present a novel one-dimensional Convolutional Neural Network (1D CNN) model to predict ~~nutrient profiles ,~~ profiles leveraging the typical shape of vertical profiles of a variable as a prior constraint during training. ~~We will present a novel model named PPCon (~~In particular, the Predict Profiles Convolutional ~~), which is trained over a dataset containing BCG Argo float measurements, for the prediction of~~ (PPCon) model predicts nitrate, chlorophyll and backscattering (bbp700) ~~,~~ starting from the date, geolocation, and temperature, salinity, and oxygen ~~. The effectiveness of the model is then accurately validated by presenting both~~ profiles. Its effectiveness is demonstrated using a robust BGC-Argo dataset collected in the Mediterranean Sea for training and validation. Results, which include quantitative metrics and visual representations~~of the predicted profiles. Our proposed approach proves capable of overcoming the limitations of MLPs, resulting in~~ , prove the capability of PPCon to produce smooth and accurate profile predictions improving previous MLP applications.

1 Introduction

Observation of the ocean is crucial for studying the state and evolution of the marine ecosystem and assessing the impact of human activities ~~(Campbell et al. (2016); Euzen et al. (2017))~~(Campbell et al., 2016; Euzen et al., 2017). Access to reliable and extensive oceanic measurements remains restricted due to the challenges of collecting comprehensive observations on

multiple temporal and spatial scales, as well as variability in the availability of observations across different biogeochemical variables ~~(Munk (2000))~~(Munk, 2000).

The introduction of autonomous oceanographic instruments such as Biogeochemical (BGC) Argo floats have notably expanded our ability to obtain subsurface and deep ocean measurements ~~(Miloslavich et al. (2019))~~(Miloslavich et al., 2019). BGC-Argo floats are autonomous profiling platforms that incorporate physical and biogeochemical sensors, enabling to collect time-series of vertical profiles across various sea conditions and throughout the complete annual cycle ~~(d'Ortenzio et al. (2014); Mignot et a~~ ~~)~~(d'Ortenzio et al., 2014; Mignot et al., 2014). Over the past decade, there has been a steady rise in the number of biogeochemical profiles acquired using these platforms ~~(Johnson et al. (2013); Johnson and Claustre (2016))~~(Johnson et al., 2013; Johnson and Claustre . These instruments are essential to advance our knowledge of the biogeochemical state of the ocean, as one of their principal ~~advantages~~ use cases is the assimilation into ocean biogeochemical models ~~(Mignot et al. (2019); D'ortenzio et al. (2020)~~ ~~)~~(Mignot et al., 2019; D'ortenzio et al., 2020). This assimilation process is particularly promising for variables such as oxygen, nitrate, and chlorophyll concentrations, as they serve as core state variables in most ocean biogeochemical models ~~(Teruzzi et al. (2021); Cossarini et al. (2019))~~(Teruzzi et al., 2021; Cossarini et al., 2019).

However, the measurement of biogeochemical variables such as nutrient concentration and carbonate system variables (e.g. nitrate, chlorophyll, and pH) remains more demanding and expensive compared to physical variables (e.g. temperature, salinity) and oxygen. In fact, among the BCG sensors, oxygen is the most commonly measured variable: there have been approximately $250,000$ oxygen profiles collected worldwide, which is twice the number of profiles for chlorophyll, and more than four times the number of profiles for nitrate and bbp700 (https://biogeochemical-argo.org. Thus, developing methods to estimate low-frequency marine biogeochemical variables from high-frequency measurements is essential to maximize the potential of observing systems such as the Argo program. Major efforts have been devoted toward the improvement of the long-term reliability and accuracy of autonomous measurements in recent years ~~(Sauzède et al. (2017))~~(Sauzède et al., 2017).

Artificial neural networks (ANNs) are computational models that are inspired by the structure and function of the human brain and have become a widely used approach for solving complex problems in a variety of fields, from computer vision and natural language processing to finance and engineering ~~(Krogh (2008))~~(Krogh, 2008). ANNs have emerged as a powerful tool for modeling complex, non-linear relationships also in the oceanographic field, where their use has seen a significant increase in recent years ~~(Ahmad (2019))~~(Ahmad, 2019). The use of these models has found applications in a wide range of areas, such as oceanic climate prediction and forecasting ~~(Mori et al. (2017))~~(Mori et al., 2017), species identification ~~(Goodwin et al. (2014)~~ ~~)~~(Goodwin et al., 2014), coastal morphological and morphodynamic modeling ~~(Goldstein et al. (2019))~~(Goldstein et al., 2019) , ocean current prediction ~~(Bolton and Zanna (2019))~~(Bolton and Zanna, 2019), interpolation and gap filling for remote sensing observation ~~(Sammartino et al. (2020))~~(Sammartino et al., 2020), and the integration of observation data into biogeochemical models ~~(Pietropolli et al. (2022))~~(Pietropolli et al., 2022). These examples demonstrate the broad utility of ANNs in advancing our understanding of the ocean and its processes.

Existing ANN-based techniques to infer low-sampled variables starting from high-sampled ones are based on Multilayer Perceptron (MLP) architecture, a type of feedforward NN that processes input data through interconnected layers of nodes, or neurons, with each neuron in a layer receiving inputs from all the neurons in the previous layer ~~(Taud and Mas (2018)~~

)(Taud and Mas, 2018). The initial model designed for this task was proposed by Sauzède et al. (2017), where a deterministic MLP network, named *CANYON*, was trained on a global ocean dataset to estimate biogeochemically relevant variables from concurrent in situ samples of temperature, salinity, pressure and oxygen and their latitude, longitude, depth, and date. Later, an improved version, called *canyon-b*, was introduced by Bittig et al. (2018). In this approach, a Bayesian framework was utilized, and experimental findings demonstrated that this method resulted in a more robust output. This methodology was subsequently limited to the Mediterranean Sea, resulting in the development of *canyon-med* by Fourrier et al. (2020), and empirical results validated the effectiveness of restricting the model to a smaller region. The latest advancement in this field is presented in Pietropolli et al. (2023a), wherein the authors enhance the performance related to Mediterranean Sea predictions by leveraging a more extensive training dataset and implementing a two-step quality check procedure to improve its quality.

Despite their widespread use, applications based on MLPs currently lack awareness of the typical shape of biogeochemical variable profiles they aim to infer. In fact, when When these methods are used to forecast predict profiles from Argo float measurements, they may generate jumps and irregularities in the reconstruction. This originates from the fact that MLPs are trained on individual data points and provide pointwise outputs, which makes the generation of regular profiles challenging as the NN does not take into account the vertical neighbors of predicted variables, possibly because they use point-wise data as input and output.

To solve this problem effectively, our idea consists of working directly with an architecture that infers the complete vertical profile. This approach takes advantage of architectures like the Convolutional Neural Network (CNN) that operate on vector inputs instead of individual points. CNNs are recognized as one of the most impressive forms of ANNs, especially for their effectiveness in tackling complex pattern recognition problems (O'Shea and Nash (2015); Gu et al. (2018))(O'Shea and Nash, 2015; Gu et al., 2018). While CNNs are well-known for their success in image classification tasks, they can also be used for other tasks such as speech recognition (Shan et al. (2018))(Shan et al., 2018), natural language processing (Collobert et al. (2011))(Collobert et al., 2011), and even drug discovery (Goh et al. (2017))(Goh et al., 2017).

In this study, we evaluate the effectiveness of a one-dimensional (1D) CNN model (Kiranyaz et al. (2021)) (Kiranyaz et al., 2021) for predicting nutrient vertical profiles from input data such as sampling time, geolocation, and profiles of temperature, salinity, and oxygen, using Argo float measurements as the training dataset. This approach, called PPCon (Predict Profile Convolutional) is applied to generate synthetic profiles of nitrate, chlorophyll, and backscattering (bbp700). Thanks to the intrinsic spatial-aware nature of its CNN architecture, PPCon is able to can leverage the typical shape of vertical profiles of a variable as a prior constraint during training. PPCon predictions are characterized by lower error with respect to the one obtained with MLP while also showing smoother predictions and the disappearing of phenomena such as gaps and irregularities in the generation of vertical profiles. approach is tested with a robust Argo dataset collected in the Mediterranean Sea. The Mediterranean Sea, a semi-enclosed marginal sea, presents a substantially high density of BGC-Argo profiles thanks to dedicated programs such as ARGO-Italy and the French NAOS initiative (D'ortenzio et al., 2020). This particularly fortunate situation has already made the Mediterranean a successful case for the development of biogeochemical modeling approaches based on BGC-Argo. For example, BGC-Argo is being integrated into the biogeochemical prediction model of the Mediterranean component of the Copernicus Marine Service (Cossarini et al., 2019; Teruzzi et al., 2021; Coppini et al., 2023).

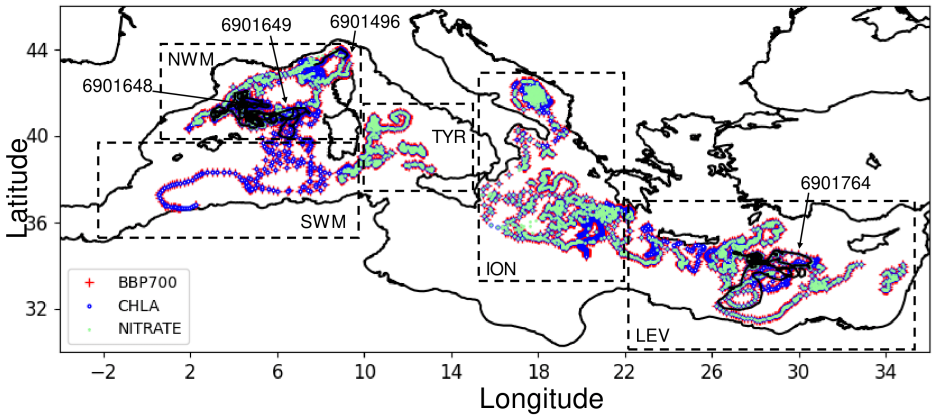

Figure 1. Position of BGC-Argo float profiles for bbp700 (red), chlorophyll (blue) and nitrate (green) that also have oxygen data. Position of the 4 BGC-Argo float profiles used for the external validation (black and numeric labels). Geographical limits of sub-regions (dashed boxes): North Western Mediterranean (NWM), South Western Mediterranean (SWM), Tyrrhenian (TYR), Ionian and southern Adriatic Sea (ION) and Levantine (LEV).

This paper is organized as follows: Section 2 presents the dataset utilized for training the deep learning (DL) architecture, including its key characteristics. Section 3 provides a detailed overview of the PPCon approach, encompassing the architecture, preprocessing techniques applied to input data, and the specialized loss function employed for network training. In Section 4, we outline the specific experimental settings employed to enable complete reproducibility of the PPCon architecture. Section 5 presents a summary of the key results obtained during the experimental campaign we conducted to validate our proposed techniques, and finally, Section 7 presents the conclusions drawn from our work and directions for future research.

2 Dataset: the Argo GDACs

The data used to train and test the architecture discussed in this paper comes from the ~~Array for Real-time Geostrophic Oceanography (Argo) program (Bittig et al. (2019))~~BCG-Argo program (Bittig et al., 2019), specifically the Argo float collecting also biogeochemical variables (BCG Argo float). ~~This program is an important part of the Global Ocean Observing System (GOOS) () and is dedicated to monitoring changes in the temperature and salinity of the upper ocean. The Argo program was primarily designed to observe pressure, temperature, and salinity (conductivity) within the upper 2000 meters of the ocean. However, due to advancements in float and sensor technologies, newly developed sensors now enable profiling floats to accurately observe biogeochemical properties. Over the past decade, there has been a consistent and substantial increase in the number of biogeochemical profiles obtained through BGC-Argo float platforms. For instance, by 2011, the global ocean had accumulated approximately 45,000 BGC-Argo profiles across all parameters, while, by 2017, this number had risen to almost 390,000 profiles. Thus, the BGC-Argo floats network has become a crucial component of the ocean observing system, enabling the monitoring, understanding, and prediction of changes in the ocean ecosystem. As of now, the Argo program has~~

amassed over 2 million data profiles, and analysis of this data has made significant contributions to basic research as well as national and international climate studies (Jayne et al. (2017)).

Our investigation specifically centers on the dataset made available by the Argo Global Data Assembly Centers (GDACs) (e.g. Coriolis, NOAA, among others) , which disseminate Our investigation used BGC-Argo S-profile data for the Mediterranean Sea downloaded from the Coriolis Argo GDAC (Argo, last visit on August 2022) and the analysis considered only Delayed Mode (DM) data procured from 11 data centers situated across 9 countries. DM data undergoes a more exacting quality control process and is typically released a few months later to their sampling (Li et al. (2020)). GDACs also supply and Adjusted Real-Time Mode (RT) data , however, given the lower quality of RT data, our analysis is based only on available for the period from $1-7-2013$ to $31-12-2020$ ensuring a larger number of DM data.

Our model was trained using data from the Mediterranean basin collected between 2015 and 2020. To ensure the data's reliability, we only selected profiles that were marked with quality flags (QFs)of 1The downloaded dataset was checked retrieving only complete profiles with Quality Flags 1 (good data), 2 , or (probably good data), 5 (value changed) and 8 for variables such as (interpolated) for temperature, salinity, nitrate , and chlorophyll. Furthermore, we applied a preprocessing step on the Furthermore, three specific quality check steps were applied for bbp700 variable based on the study by Dall'Olmo et al. (2022) , which introduced a new set of real-time quality-control tests for this variable, since no procedures have been agreed upon so far to quality control bbp700 data in real-time. Specifically, we applied three of the procedures introduced in this work to bbp700 profiles: the (Dall'Olmo et al., 2022): missing-data test , which detects and flags profiles that contain a (profiles with substantial amount of missing data; the); high-deep-value test , which flags (profiles with unusually high bbp700 values at depth; and the negative-BBP test , which flags data points or value at depth) and negative-bbp test (profiles with negative bbp700 values.). Finally, for each of the three variables (nitrate, chlorophyll and bbp700), profiles were only considered if the corresponding oxygen, salinity and temperature profiles were also available. As a result of the quality check, the number of profiles for each variable used in the train, test and validation is reported in Table 2, the float spatial distribution is shown in Figure 1 and the dataset is available at the following repository (Amadio et al., 2023).

3 PPCOn: Profile Prediction Convolutional Neural Network

Within the realm of DL, CNN has emerged as a prominent architecture (Albawi et al. (2017)). A CNN network functions as a feedforward NN capable of capturing data features through convolutional structures (Li et al. (2021)). While the two-dimensional (2D) CNN architecture is designed to extract spatial features from two-dimensional data such as images (O'Shea and Nash (2015)), the 1D CNN is specifically designed to extract temporal features from one-dimensional sequential data such as signals or time series data (Tang et al. (2020)). Due to their streamlined and efficient configuration that employs only 1D convolutions (scalar multiplications and additions), 1D CNNs offer advantages in terms of real-time processing and cost-effectiveness for hardware implementation. (Kiranyaz et al. (2021)).

This section introduces the PPCon architecture, which is primarily a 1D CNN with additional MLPs employed to transform punctual point-wise data into a vectorial shape - necessary for the training of the convolutional component. The input variables

~~for PPCon include~~ for PPCon includes sampling data, geolocation, temperature, salinity, and oxygen, while the ~~output variables comprise~~ PPCon output comprises vertical profiles for nitrate, chlorophyll, and BBP. Despite using the same architecture, a separate model is trained for each output variable, and different hyperparameters (number of epochs, weights of the loss function, and so on) are set for each of them. This separate tuning is necessary due to some intrinsic differences such as the numerosity of the training set and the variable ranges. The hyperparameters are tuned manually by comparing performance on the test set composed of unseen data, based on a fitness metric to be introduced later. A specific loss function is designed to promote good performances, generalization capabilities, and smooth predictions.

3.1 Input preprocessing

Table 1. The MLP component of the PPCon model illustrated in diagram form. All the 4 MLPs used in the PPCon architecture share the same architecture.

Layer	Output Size	Activation Function
Input	~~[32, 1]~~ 1	-
Linear	~~[32, 80]~~ 80	SELU
Linear	~~[32, 140]~~ 140	SELU
Linear	~~[32, 200]~~ 200	SELU
Output	~~[32, 200]~~ 200	-

The data considered for feeding the DL architecture comprises a collection of measurements, where each input-output pair consists of the information collected by a singular float profile. The inputs consist of two distinct categories of data, namely ~~punctual~~ point-wise and vectorial. ~~Punctual~~ Point-wise data encompasses temporal and geospatial parameters, such as the sampling date (specifically year and day) ~~, geolocation,~~ and geographic coordinates (latitude and longitude), while vectorial data encapsulates profiles of temperature, salinity, and oxygen, as recorded by the float instruments. Given that the 1D CNN architecture operates only on vectorial input data, a coherent transformation of ~~punctual~~ point-wise features into vectorial ones is required.

In this regard, we leverage an MLP architecture that accepts ~~punctual~~ point-wise input and transforms it into vectorial form. MLPs are employed ~~in order~~ to enable the NN to automatically learn how to weigh differently the importance of such ~~punctual~~ point-wise input features in correspondence of different levels of depth. A separate MLP is trained for each of the four pointwise inputs. The MLP architectures have the same number of layers and neurons contained in these layers (Table 1), since there are no a priori reasons to make them different.

During training, the weights of the MLP are optimized along with the weights of the 1D CNN architecture. Since the MLP operates as a non-linear function, this training approach enables the creation of a mapping between ~~punctual~~ point-wise input and its vectorial equivalent. This enables PPCon to effectively exploit ~~punctual~~ point-wise information and achieve optimal learning outcomes. The output vectors generated by the MLP are concatenated with the remaining vectorial input, yielding a seven-channel tensor that serves the input of the PPCon architecture.

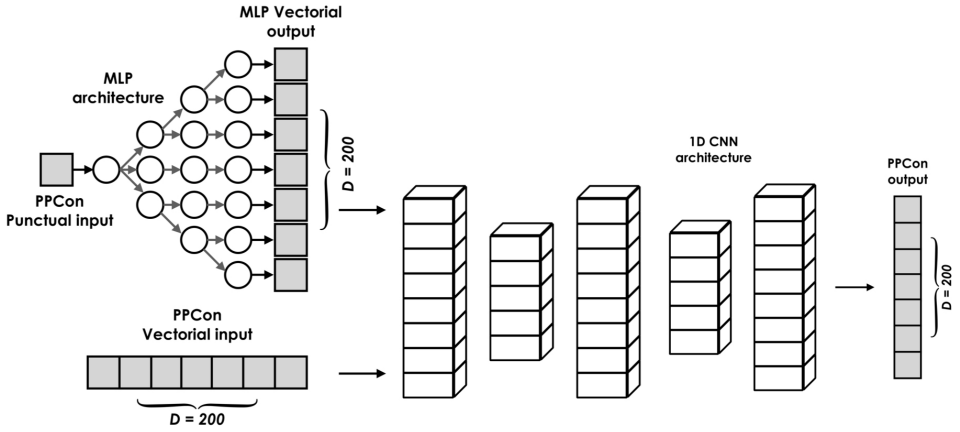

Figure 2. Illustration of the principal architectural components of the PPCon model: (i) MLP network to transform the point-wise inputs (day, year, latitude, and longitude) into vectorial form; ii) vectorial inputs (profiles of temperature, salinity, and oxygen and output of the MLP); iii) structure of encoder-decoder of a 1D CNN architecture; iv) output vector representing the vertical profile of one of the target variables (nitrate, chlorophyll, or backscattering).

Thus, to sum up, the input to the PPCon architecture consists of four point-wise inputs — latitude, longitude, day, and year — which are transformed into a vectorial input using an MLP architecture. In addition, the architecture uses for the training three 1×200 input vectors representing the profiles of temperature, salinity, and oxygen.

3.2 PPCon Architecture

The convolutional component of the PPCon architecture, summarized in Table 2, is a DL model comprising multiple 1D convolutional and deconvolutional layers.

The input tensor has a 1-dimensional shape, with a total number of channels equal to 7, one for each of the three variables to reconstruct, i.e. nitrate, chlorophyll and bbp700.

The architecture includes a total of nine layers, each of which applies a set of filters to the input tensor. These filters are designed to detect specific features or patterns, with the number and size of the filter kernels specified by the parameters of each layer. To enable effective feature extraction across different scales, various stride parameters are employed to specify the step size at which the filters are applied to the input tensor. To ensure that the output tensor has the same shape as the input tensor, padding parameters are incorporated, adding zero padding to the borders of the input tensor. The output tensor is then normalized through a batch normalization ~~(Santurkar et al. (2018))~~ (Santurkar et al., 2018) layer after each convolutional layer. The normalization process ensures that the output tensor has a mean of zero and a unit variance, thereby minimizing the effect of

Table 2. The convolutional component of the PPCon model is illustrated in diagram form. The key attributes of the NN are outlined, encompassing parameters, output size (represented as [~~batch size,~~ number of channels, input length]), as well as any additional layers. More specifically, "BN" denotes the batch normalization layer, "SELU" represents the non-linear selu() activation layer, and "Dropout" indicates the presence of a dropout layer along with the corresponding dropout rate.)

Layer	Kernel	Stride	Padding	Output Size	Additional Details
Input	-	-	-	~~[32, 7, 200]~~ [7, 200]	-
Conv. 1D	2	1	2	~~[32, 64, 203]~~ [64, 203]	BN, SELU, Dropout (rate: d_r)
Conv. 1D	2	2	1	~~[32, 128, 102]~~ [128, 102]	BN, SELU, Dropout (rate: d_r)
Conv. 1D	4	1	1	~~[32, 128, 101]~~ [128, 101]	BN, SELU, Dropout (rate: d_r)
Conv. 1D	4	1	2	~~[32, 128, 102]~~ [128, 102]	BN, SELU, Dropout (rate: d_r)
Deconv. 1D	2	2	2	~~[32, 128, 200]~~ [128, 200]	BN, SELU, Dropout (rate: d_r)
Conv. 1D	3	1	1	~~[32, 128, 200]~~ [128, 200]	BN, SELU, Dropout (rate: d_r)
Deconv. 1D	2	2	1	~~[32, 64, 398]~~ [64, 398]	BN, SELU, Dropout (rate: d_r)
Conv. 1D	2	2	1	~~[32, 32, 200]~~ [32, 200]	BN, SELU, Dropout (rate: d_r)
Conv. 1D	3	1	1	~~[32, 1, 200]~~ [1, 200]	BN, SELU
Output	-	-	-	~~[32, 1, 200]~~ [1, 200]	-

covariate shifts and enhancing the stability of the training process. Following normalization, the output tensor is passed through a scaled exponential linear unit (SELU) activation function ~~(Rasamoelina et al. (2020))~~(Rasamoelina et al., 2020), which is defined as:

$$f(x) = \begin{cases} \lambda x & \text{if } x \geq 0 \\ \lambda\alpha(e^x) & \text{if } x < 0 \end{cases} \tag{1}$$

where and $\lambda \approx 1.0507$ and $\alpha \approx 1.6732$. SELU has been selected as an activation function as it induces self-normalization properties. Dropout layers ~~(Baldi and Sadowski (2013))~~ (Baldi and Sadowski, 2013) are also incorporated to prevent overfitting during training, promoting robust generalization and enhancing the NN's ability to learn diverse features from the input data. These layers randomly drop out some of the network neurons, with the specific probability of dropout (d_r) specified for each layer in the architecture's hyperparameters.

The final convolutional layer produces a 1-channel output tensor, which represents the final prediction of the model.

3.3 Loss function

The choice and design of a loss function is a crucial step in the development of DL models, as it determines the objective to be optimized during training and can have a significant impact on the model's ability to generalize to new data. Besides the ability to skilfully reproduce output variable profiles, we want the PPCon architecture to mitigate overfitting and produce a smooth prediction curve.

To ~~fulfil~~ fulfill these objectives, we define a loss function comprising three components: first, the Root Mean Square Error (RMSE) between the target output and the PPCon architecture's prediction, to assess prediction quality. Second, to mitigate overfitting phenomena, a regularization term known as λ-regularization is employed, which penalizes complex curves in proportion to the square of the model's weights ~~(Zou and Hastie (2005))~~(Zou and Hastie, 2005). By promoting smaller weight values, this technique encourages the generation of more general predictions. The severity of this penalty is determined by a multiplicative factor λ, which is a hyperparameter of the model. The final component of the loss function is incorporated to promote the generation of a smoother output curve. This term, controlled by a hyperparameter ~~α~~α_s, serves as a regularization technique that penalizes sharp variations in the output. The final loss formula is as follows:

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{N} |\theta_i|^2 + \alpha_s \sum_{i=1}^{n-1} (\hat{y}_{i+1} - \hat{y}_i)^2 \tag{2}$$

where y represents the target value, $\hat{y}$ the output of PPCon model, n the length of both y and $\hat{y}$ and N the total number of weights of the DL architecture.

4 Experimental Study

This section presents the experimental settings for the PPCon architecture, which are defined for each predicted variable under consideration. The complete code for the reproducibility of the results presented in this paper is available at: https://zenodo.org/record/8369573.

4.1 Training

We divided the dataset into three subsets: training, testing, and validation. The training set was used for model training and parameter optimization. The testing set was utilized to evaluate the model's performance on unseen data and assess its generalization ability. Finally, the validation set was employed for hyperparameter tuning and model selection. The dataset was randomly partitioned, ensuring that each subset contained a representative distribution of the overall data characteristics. The sizes of the training, testing, and validation sets were chosen as ~~80, 10, and 10. This approach enabled us to assess and validate the performance of our model effectively~~80%, 10%, and 10% of the total number of measurements. Moreover, before operating this partition, a few float instruments have been selected and all of their measurements have been excluded from both the training, test, and validation set. These samples will be used as an external validation dataset. The metrics and the performances over this external validation dataset are a more effective indicator of the generalization capabilities of the PPCon model with respect to the metrics on the test set.

~~In order to~~ To train the NN model efficiently, the input dataset is partitioned into minibatches, where each minibatch contains 32 samples. The batch size ~~, which~~ is a hyperparameter that determines the number of samples processed before updating the model weights~~, is a hyperparameter that must be set prior to training~~. By processing multiple samples in a minibatch, the model can update its parameters more frequently, which can lead to faster convergence and improved generalization performance

	#samples	#epochs	batchsize	dropout rate	λ	~~α~~ α_s
nitrate	2337	50	32	0.2	0.001	0.001
chlorophyll	3189	150	32	0.2	0.0001	0.0001
BBP	3942	100	32	0.2	~~$1e^{-7}$~~ $1 \cdot 10^{-7}$	~~$1e^{-7}$~~ $1 \cdot 10^{-7}$

Table 3. Summary of hyperparameters and dataset sizes~~for all inferred variables~~.

~~(Bottou (2010)). Once all the mini-batches have been processed by the optimization algorithm, the model has completed an epoch of training.~~ (Bottou, 2010).

Adadelta ~~(Zeiler (2012))~~ (Zeiler, 2012) is the algorithm that is selected as the optimizer for training the network due to its ability to dynamically adapt over time using only first-order ~~information~~derivatives of the objective function. This method eliminates the need for manual tuning of the learning rate and has been found to exhibit robustness~~in the face of noisy gradient information, various data modalities, different model architecture choices, and hyperparameter selection~~.

It is worth recalling that the PPCon architecture includes a 1D CNN and four MLPs, which convert point-wise input into a vector form suitable for use by the CNN. The MLPs and the CNN component of PPCon were trained using the same optimizer, with concurrent weight updates across all networks. This approach enables the joint learning of optimal information transfer from point-wise input to vector form, as well as the accurate generation of predicted profiles based on the input tensor.

To accelerate the training process, the model was trained using a GPU (graphics processing unit), which allowed for parallelized computation of the forward and backward passes.

The model's performance was evaluated once every 25 ~~epochs~~ epoch by assessing its ability to predict outcomes on the test set, which consists of previously unseen data. To prevent overfitting and minimize computational burden, we introduced an early stopping routine. Specifically, the training was interrupted if the error metrics on the validation set increased for two consecutive evaluations (i.e. after 50 epochs of training). The final model selected was the one trained ~~prior to the two~~ before the two 25 consecutive test loss increases.

4.2 Experimental Settings

Since each output variable has intrinsic differences in training set size, range of values, and profile shapes and variabilities, a separate hyperparameter tuning step is performed for each of them. These hyperparameters were tuned using a systematic search over a range of values, guided by the performance of the model on a held-out validation set. To avoid overfitting in the test set, we employed cross-validation techniques to estimate the generalization performance of the model and selected the hyperparameters that yielded the best performance.

The hyperparameters used for training the three PPCon architectures are summarized in Table 3, together with the size of the dataset, the total number of epochs performed, and the batch size dimension which have already been discussed in previous sections.

	North West Med.	South West Med.	Tyrrhenian	Ionian	Levantine
Latitude	$40°N - 45°N$	$32°N - 40°N$	$37°N - 45°N$	$30°N - 45°N$	$30°N - 37°N$
Longitude	$-2°E - 9.5°E$	$-2°E - 9.5°E$	$9.5°E - 15°E$	$14°E - 22°E$	$22°E - 36°E$

Table 4. Geographical limits of the five areas in which the Mediterranean is divided for the posterior analysis.

In our experiments, we applied a dropout rate of 0.2, which was consistent across all trained models. This means that during training, each neuron in the NN has a 20% chance of being randomly excluded from the computation. Dropout regularization is a technique used to prevent overfitting by encouraging each neuron to encode information independently, thereby inhibiting co-dependencies among neurons.

Table 3 also reports the multiplicative factors that determine the relative contributions of different elements that compose the loss function defined in Section 3.3. The values of these hyperparameters vary depending on the variable being inferred, as these variables have different orders of magnitude and result in RMSE values that vary in magnitude as well. It is crucial to accurately balance the regularization term, governed by λ, and the smoothness term, governed by α, to prevent them from dominating the loss function's RMSE component. The optimal values reported in Table 3 guarantee a good and smooth prediction of the vertical profile.

The last implementation detail to be addressed concerns the creation of vectors used to feed the PPCon architecture. As previously discussed, vectorial inputs of different natures are fed into the CNN component of PPCon: firstly, the outputs of an MLP architecture; secondly, vectors representing input variables (temperature, salinity, oxygen) at different depths. To ensure that all input vectors have the same length, we adopted the following strategy: (i) the output and input variables have been interpolated on a regular grid of size 200, and (ii) the output of MLPs have the same length and discretization of the input variable vectors. For nitrate, we considered a depth range of 0 to 1000 meters with an interpolation interval of 5 meters, whereas, for chlorophyll and BBP, we considered a depth range of 0 to 200 meters with an interpolation interval of 1 meter. Then, we set the output layer dimension of the MLP to 200 to ensure that all input vectors have the same length. As a result, the final dimension of the input tensor is 7 (the number of inputs) x 200 (the length of the input vector) x the number of dimensions in the training set.

4.3 A posterior validation analysis of PPCon

To validate the PPCon architecture, we conducted a thorough analysis of its ~~average~~ performance in different geographic areas ~~and across various seasons. The choice underlying this investigation originates from the fact that diverse geographic areas and seasons are known to have distinct profile properties, such as the shape of the vertical profiles~~ (NWM, TYR, SWM, ION and LEV in Figure 1) and across the four seasons: winter (JFM), spring (AMJ), summer (JAS) and autumn (OND). The specific geographical limits related to different areas are reported also in Table 4. While the PPCon model is trained on the entire dataset, this subdivision is only used to analyze the performance retrospectively to check whether the non-uniform geographical and spatial distribution of the profiles and the natural variability of the profiles (e.g. depth and slope of the nitracline, or depth and

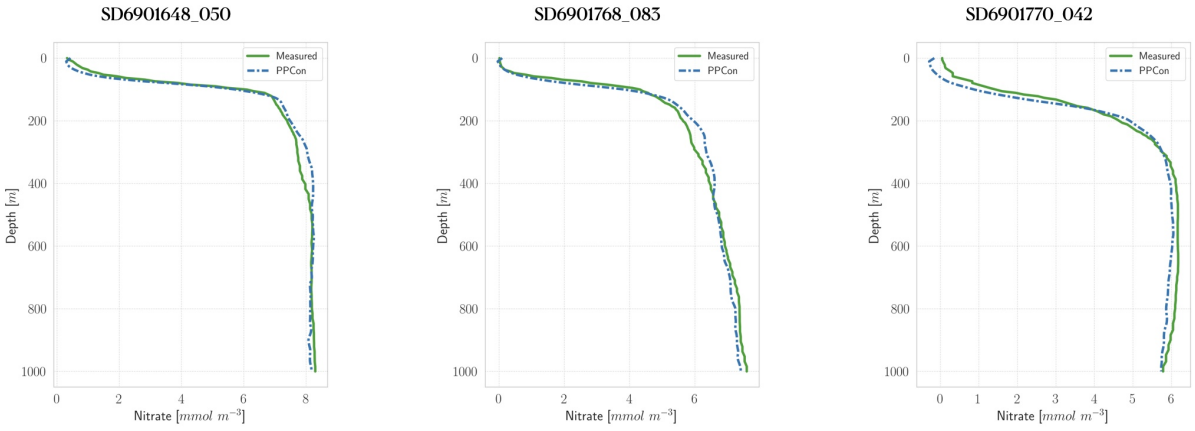

Figure 3. Profiles of nitrate for some selected floats (WMO numbers and cycles in the title)~~and~~ . Profile dates and geolocations are reported in Table 5. Comparison between measured profile (green lines) ~~, MLP reconstruction (azure dashed lines),~~ and PPCon reconstruction (blue ~~dash-dotted~~ dashed lines). Profiles are from the subset used for the test.

nitrate				chlorophyll			
WMO	date	lat	lon	WMO	date	lat	lon
~~6901653~~ 6901648	~~24~~27/~~07~~12/2014	~~41.85~~ 41.94	~~4.55~~ 4.02	~~6902901~~ 6902954	~~10~~01/~~10~~08/2019	~~42.91~~ 42.48	~~7.62~~ 7.12
~~6901769~~ 6901768	~~20~~17/~~12~~03/2016	~~39.11~~ 38.02	~~10.96~~ 18.87	~~6901776~~ 6901648	13/~~04~~08/~~2014~~ 2015	~~42.79~~ 40.60	~~7.01~~ 4.25
~~6901771~~ 6901770	~~18~~17/~~10~~11/2015	~~36.01~~ 36.44	~~20.12~~ 19.65	~~6901773~~ 6901496	~~19~~27/~~09~~12/~~2017~~ 2013	~~32.93~~ 43.49	~~31.24~~ 9.00

Table 5. WMO, date, and geolocation of the float profiles reported in Figure 1 – 3.

intensity of ~~DCM) and the values at the surface and in deep water. Our goal is to evaluate the model's ability to accurately capture these variations. We wish to once again note the model is trained on the entire dataset, and this division is purely for a posteriori analysis of the performances. This a posteriori analysis of the performances has the objective of identifying possible influence and bias of the uneven~~ the DCM) have an influence. In particular, the RMSE is calculated for the reconstructed profiles in each area and season to verify the presence of any bias in the accuracy of PPCon in capturing the spatial and temporal ~~distribution of the profiles on the performance of the PPcon model.~~ variability in the Mediterranean Sea.

~~Five different geographic areas are considered, namely: Northern West Mediterranean , Southern West Mediterranean, Tyrrhenian, Ionian, and Levantine. The geographical limits related to these variables are reported in Table4.~~

~~To gain a comprehensive understanding of how the model performs across different seasons, we analyzed four distinct time periods: winter (January to March), spring (April to June), summer (July to September), and autumn (October to December).~~

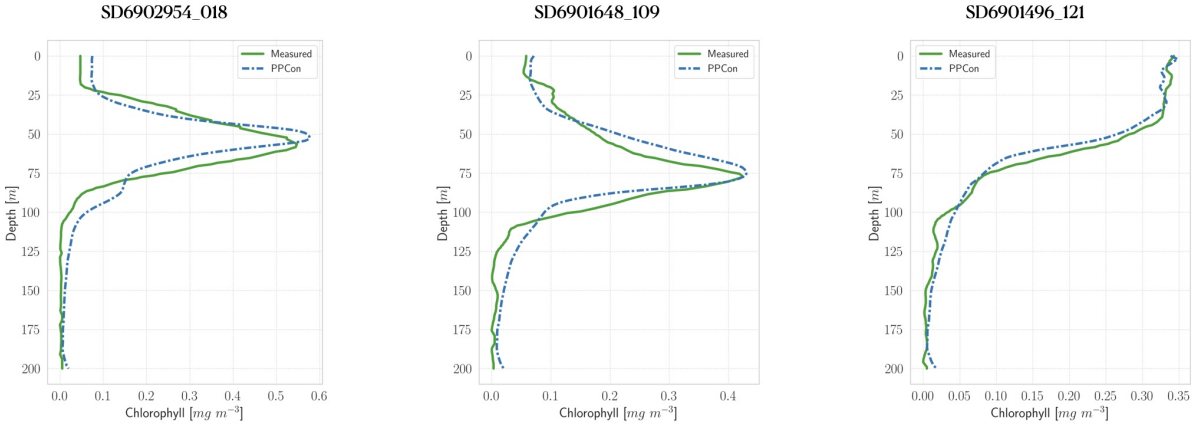

Figure 4. Profiles of chlorophyll for some selected floats (WMO numbers and cycles in the title)~~and~~ . Profile dates and geolocations are reported in Table 5. Comparison between measured profile (green lines) and PPCon reconstruction (blue dashed lines). Profiles are from the subset used for the test.

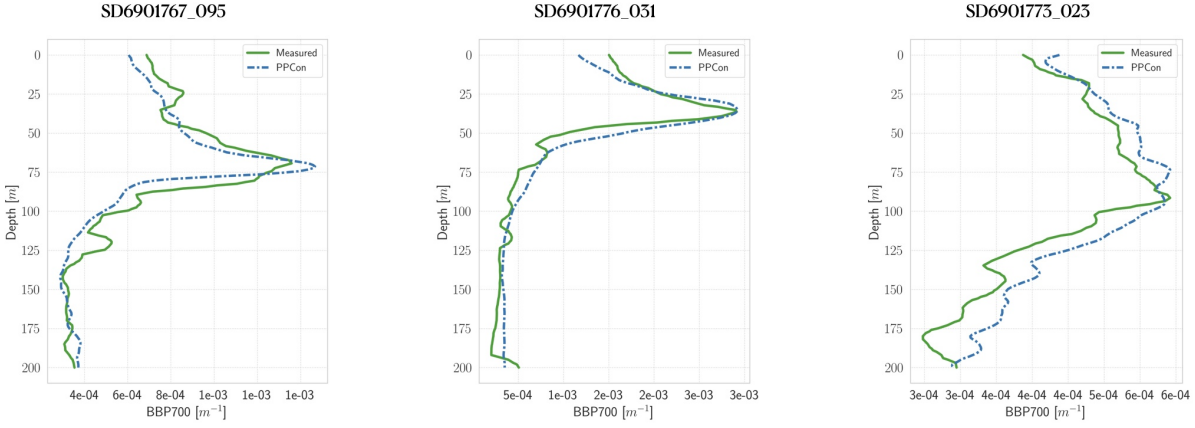

Figure 5. Profiles of bbp700 for some selected floats (WMO numbers and cycles in the title)~~and~~ . Profile dates and geolocations are reported in Table 5. Comparison between measured profile (green lines) and PPCon reconstruction (blue dashed lines). Profiles are from the subset used for the test.

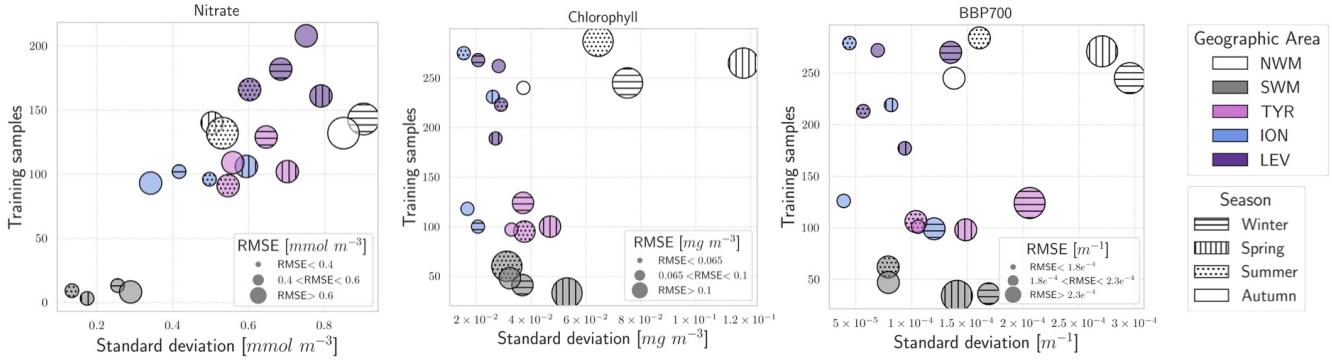

Figure 6. Plot of the RMSE distribution with respect to the data variability (on the $x-$axis) and training dataset size ~~of~~ (on the ~~different~~ $y-$axis). Different sub-areas ~~(colour~~ are represented by different colors of the symbol ~~)~~ and ~~the~~ different seasons ~~(~~are represented by different symbol fill ~~pattern)~~patterns. RMSE values are categorized by the size of the symbols~~.~~, and bigger symbols correspond to bigger RMSE values

.

	Winter		Spring		Summer		Autumn	
	train	test	train	test	train	test	train	test
Nitrate $[mmol/m^3]$	~~0.640~~ 0.51	~~0.665~~ 0.51	~~0.591~~ 0.51	~~0.610~~ 0.52	~~0.570~~ 0.51	~~0.615~~ 0.49	~~0.631~~ 0.48	~~0.64~~
Chlorophyll $[mg/m^3]$	~~0.082~~ 0.08	~~0.098~~ 0.07	~~0.129~~ 0.12	~~0.134~~ 0.13	~~0.069~~ 0.08	~~0.067~~ 0.08	~~0.045~~ 0.05	~~0.04~~
bbp700 $[\times 10^{-4} m^{-1}]$	~~2.561e^{-4}~~ 2.6	~~2.571e^{-4}~~ 2.4	~~3.101e^{-4}~~ 2.3	~~2.691e^{-4}~~ 2.6	~~1.871e^{-4}~~ 1.5	~~1.941e^{-4}~~ 1.4	~~2.211e^{-4}~~ 1.5	~~1.97~~

Table 6. RMSE calculated between the float measurements and the reconstructed values obtained from the PPCon architecture~~for all variables inferred~~. This metric is evaluated individually for the train and test sets. The RMSE is computed for different seasons of the year (described in Section 2).

	North Western Med		South Western Med		Tyrrhenian		Ionian	
	train	test	train	test	train	test	train	test
Nitrate $[mmol/m^3]$	~~0.514~~ 0.62	~~0.531~~ 0.65	~~0.637~~ 0.37	~~0.643~~ 0.38	~~0.706~~ 0.44	~~0.704~~ 0.44	~~0.426~~ 0.41	~~0.4~~
Chlorophyll $[mg/m^3]$	~~0.119~~ 0.14	~~0.139~~ 0.13	~~0.103~~ 0.10	~~0.098~~ 0.12	~~0.074~~ 0.08	~~0.075~~ 0.08	~~0.051~~ 0.04	~~0.0~~
bbp700 $[\times 10^{-4} m^{-1}]$	~~2.661e^{-4}~~ 2.6	~~2.641e^{-4}~~ 2.4	~~2.501e^{-4}~~ 2.1	~~2.581e^{-4}~~ 2.0	~~3.871e^{-4}~~ 2.3	~~2.831e^{-4}~~ 2.3	~~2.051e^{-4}~~ 1.4	~~2.0~~

Table 7. RMSE calculated between the float measurements and the reconstructed values obtained from the PPCon architecture~~for all variables inferred~~. This metric is evaluated individually for the train and test sets. The RMSE is computed for different geographic areas (described in Section 2).

5 Results

This section presents the results of the PPCon model in predicting nitrate, chlorophyll, and bbp700 profiles. The effectiveness of the model is evaluated by presenting both quantitative skill metrics (i.e., RMSE) and visual representations of the predicted profiles based on the test set.

Specifically, we assess the PPCon performance over different seasonal variations (Table 6), and different geographic areas (Table 7). The absence of overfitting is supported by reporting the RMSE for both the training and test sets, which exhibit non-dissimilar values.

In terms of performances across different geographic areas (Table 7), it can be seen that the lowest RMSE values for chlorophyll and bbp700 are in the eastern sub-basins, while for nitrate the lowest and highest values are in the two eastern sub-basins. Notably, the prediction accuracy for nitrate is significantly higher in the ~~Ionian Basin~~ION, with RMSE values below 0.5. Considering the temporal evolution of RMSE values (Table 6), the highest values of chlorophyll and bbp700 are in spring and winter, which appears reasonable given the higher variability of vertical pattern during these seasons ~~(Cossarini et al. (2019); Teruzzi et al. (202~~ ~~)~~(Cossarini et al., 2019; Teruzzi et al., 2021). Errors for nitrate are fairly homogeneous among the seasons, which are the highest during winter (e.g. vertical mixing season) and the lowest values during the stratification seasons (i.e. spring and summer). As for chlorophyll, the western basin of the Mediterranean shows higher RMSE values. This can be attributed to the naturally elevated chlorophyll levels observed in that specific area, which consequently lead to higher RMSE values.

For each variable investigated, we present three instances of vertical profile reconstruction using the PPCon architecture, compared to the profile measured by the float instrument whose corresponding identification number is indicated above each profile. To ensure geographic and seasonal diversity, we have selected profiles representing different regions, including at least one from the West Mediterranean and one from the East Mediterranean. Figure 3-5 displays examples of, respectively, reconstructed nitrate, chlorophyll, and bbp700 profiles. For the nitrate variable, the reconstruction performed by the MLP model is also reported ~~(Pietropolli et al. (2023a))~~(Pietropolli et al., 2023a). The information related to these profiles, such as the date and geolocation of sampling, are reported in Table 5.

The obtained results confirm the quality of the profiles generated by the PPCon architecture, which appears to better reconstruct the shape and smoothness of the profiles than the previous MLP architecture. Indeed, PPCon ~~is able to~~ can capture different profile shapes associated with different geographic and seasonal conditions, as ~~clearly~~ demonstrated by the predicted nitrate and chlorophyll profiles. ~~Higher~~ The visual inspection of all test profiles (not shown) revealed that higher quality in the prediction is achieved for the nitrate variable, followed by chlorophyll, and last by bbp700. This outcome is expected, as the nitrate variable exhibits lower variability in the values and profile shapes than ~~the other two variables~~chlorophyll and bbp700. For a more detailed analysis of the behavior of the PPCon architecture quality of the predicted profiles, Appendix A reports a comparison between the mean of PPCon predicted profiles and the mean of profiles measured by the float instruments (in the test set)~~for all investigated variables~~, providing a more specific insight on the PPCon performances in different geographic areas and seasons.

Nitrate $[mmol/m^3]$		Chlorophyll $[mg/m^3]$		bbp700 $[m^{-1}]$	
6901767	~~0,4288~~ 0.44	6901648	~~0,1739~~ 0.14	6901649	~~2,6231e^{-4}~~ $1.9 \cdot 10^{-4}$
6901764	~~0,4583~~ 0.31	6901496	~~0,0878~~ 0.13	6901496	~~1,5071e^{-3}~~ $2.2 \cdot 10^{-4}$

Table 8. RMSE calculated between the float measurements and the reconstructed values obtained from the PPCon architecture over the external validation dataset.

~~In order to~~ To understand the impact of the training set numerosity and of the variability of profiles on the quality of the PPCon predictions we investigated the relation between these quantities and the PPCon error. Specifically, Figure 6 ~~reports points whose size corresponds to the prediction RMSE (divided according to the four seasons and the~~ illustrates the RMSE values computed for the reconstructed profiles subdivided into five geographic areas ~~) and their relation with the~~ and four seasons. RMSE values, which are indicated by the size of the symbols, are plotted against the variability of the training set (~~on the x-axis),~~ quantified by the standard deviation ~~, and the numerosity~~ on the $x-$axis) and the size of the training set (on the ~~y-axis~~$y-$axis). This figure offers also valuable insight into the geographical and seasonal distribution of the training dataset dimension.

In terms of training size, the plots of the three variables shows that the SWM exhibits the smallest number of training profiles, while the largest numbers are in the NWM and LEV areas. Natural variability changes across sub-basins with higher values of standard deviations in the western sub-basins (i.e., SWM, NWM and TYR). Variability and sample size show a roughly homogeneous distribution among seasons.

The analysis of the nitrate plot reveals fairly homogeneous errors across ~~all sub-areas and seasons, suggesting a lack of a strong relationship between errors and variability or data availability~~natural variability and training sample size, excluding the SWM profiles. The NWM is the basin predicted with the lowest accuracy, while the SWM and ION have in general the lowest errors. In terms of ~~chlorophyll and~~ seasonal variation, the RMSE values appear slightly lower during summer compared to winter and spring.

Chlorophyll and bbp700 ~~, both variables~~ exhibit similar behavior (central and right plots in Figure 6). In particular, data availability appears to have no significant impact on the error, whereas RMSE tends to increase proportionally with the variability.

Regarding the chlorophyll, the performances of the western sub-basins (i.e., NWM, SWM and TYR) are lower than the eastern sub-basins (LEV and ION), likely due to higher profile variability. Winter and autumn are the seasons with lower RMSE, while the highest error is predicted in Spring.

Similarly, better performances for bbp700 are observed in LEV and ION compared to the western sub-basin.

Interestingly, summer and autumn performances are almost 50% better than winter and spring ones despite natural variability and sample size do not show appreciable differences among the seasons.

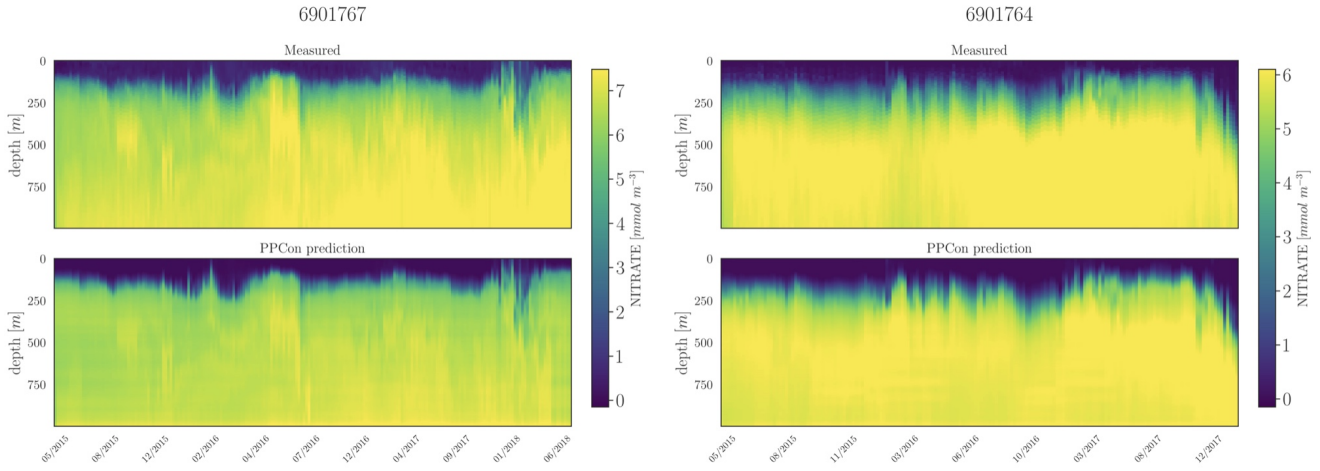

Figure 7. Hovmöller diagrams for the nitrate of two selected floats (WMO name in the title) belonging to the external validation set. BGC-Argo measurements (upper panels) and PPCon prediction (lower panels) are compared. WMO ~~6901648~~ 6901767 sampled the ~~$40°N - 42°N$~~ $39°N - 41°N$ and ~~$2°E - 6°E$~~ $10°E - 11°E$ area during ~~$2014 - 2016$~~$2015 - 2018$, whereas WMO 691764 sampled the $31°N - 34°N$ and $26°E - 40°E$ area during $2015 - 2017$.

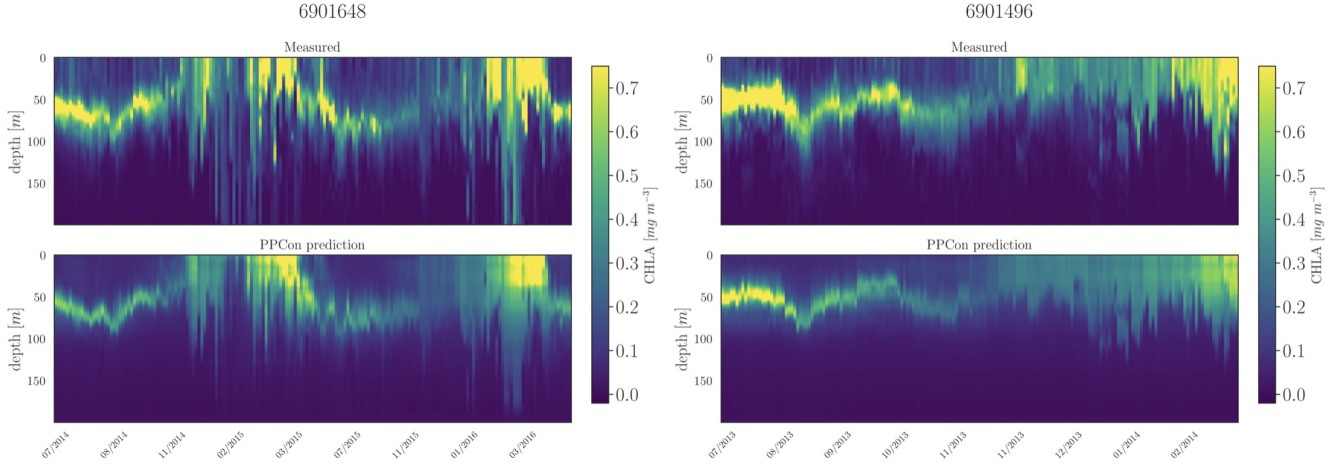

Figure 8. Hovmöller diagrams for the chlorophyll of two selected floats (WMO name in the title) belonging to the external validation set. BGC-Argo measurements (upper panels) and PPCon prediction (lower panels) are compared. WMO 6901648 sampled the $40°N - 42°N$ and $2°E - 6°E$ area during $2014 - 2016$, whereas WMO 6901496 sampled the $42°N - 43°N$ and $7°E - 12°E$ area during $2013 - 2014$.

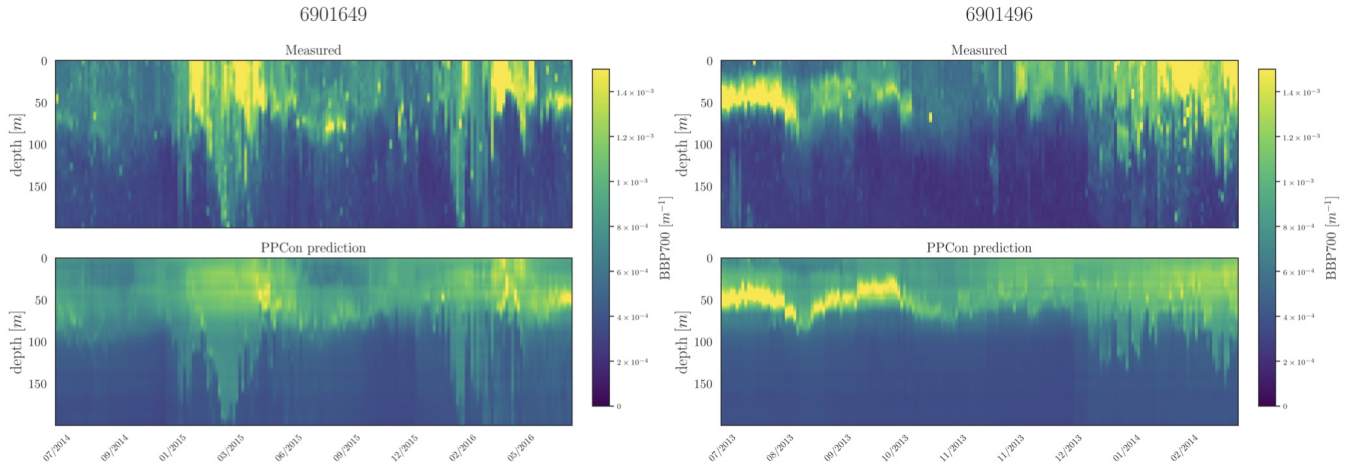

Figure 9. Hovmöller diagrams for the bbp700 of two selected floats (WMO name in the title) belonging to the external validation set. BGC-Argo measurements (upper panels) and PPCon prediction (lower panels) are compared. WMO 6901649 sampled the $39°N - 41°N$ and $3°E - 7°E$ area during 2014 – 2016, whereas WMO ~~6900807~~ 6901496 sampled the ~~$41°N - 44°N$~~ $42°N - 43°N$ and ~~$29°E - 35°E$~~ $7°E - 12°E$ area during ~~2014 – 2018~~2013 – 2014.

5.1 PPCOn performance over an external validation dataset

For each inferred variable, Figures 7-9 display Hovmöller diagrams of measured and reconstructed float instruments belonging to the external validation set, and Table 8 reports the corresponding RMSE values. This represents a particularly stringent validation test since none of the profiles measured by these floats have been encountered by the PPCon model during the training or validation phases. The figures compare the in situ float measurements (upper diagram) and the predictions generated by the PPCon architecture (lower diagram) for floats that have been specifically selected to cover different geographical regions of the Mediterranean Sea (e.g. one in the eastern and one in the western Mediterranean Sea). White lines in the diagram indicate float measurements that cannot be compared due to various reasons such as the sensor's temporary inability to measure the specific variable inferred or the absence of one of the inputs necessary for the PPCon architecture (e.g. at least one between temperature, salinity, and oxygen). This could be attributed to limitations in the sensor or unacceptable quality flags associated with the collected data. Nevertheless, the number of profiles that cannot be calculated by PPCon is rather low and does not degrade the very good capacity of the reconstructed profiles to reproduce the temporal evolution of the vertical dynamics shown by the measured floats.

These plots confirm the PPCon capability to perform accurate predictions also regarding float devices which are totally unseen by the model. The nitrate (Figure 7) reconstructions exhibit a very good performance of PPCon in predicting the vertical dynamics associated with the temporal evolution of the nutricline depth (i.e. the depth at which the sharp increase of the nitrate values is observed), the values in the deep layers (which are different in the sub-areas sampled by the two floats), and the occurrence of deep vertical mixing when surface concentration increases to values higher than $3mmol/m^3$. Particularly

impressive is the capability of PPCon to reconstruct the temporal dynamics of chlorophyll (Figure 8). The reconstruction effectively captures the evolution of the chlorophyll surface peaks during winter and the formation of the deep chlorophyll maximum during summer in both floats representing the two areas of the Mediterranean Sea. Among the three variables, the bbp700 (Figure 9) shows the least accurate predictions. However, the model still displays the ability to infer the key characteristics of the variable's temporal behavior. Nonetheless, the generated predictions for bbp700 appear slightly less detailed compared to the original sampling, indicating a partial limitation of the model in capturing small-scale variations.

Quantitatively, the prediction quality of the PPCon architecture (RMSE values in Table 8 are fairly well aligned with the metrics calculated over the test set, as indicated in Table 7. In particular, nitrate errors of the two floats are quite homogeneous and 30% lower than the RMSE values of the test set. The errors in chlorophyll and bbp700 predictions exhibit greater variability, with values almost double for the floats in the western Mediterranean with respect to the eastern ones. This is, however, in line with results reported in Table 7, and Figure 6 where higher errors are associated with higher variability.

6 Discussion

To our knowledge, the PPCon architecture is the first attempt to predict vertical BGC-Argo profiles ~~by means of~~ through a convolutional architecture. Its primary objective is the incorporation of typical profile shapes during the training phase, in contrast with previous architectures which all relied on MLP architectures and point-wise strategy. There are notable distinctions between the two approaches: MLPs were trained on cruise data, which are known to be more precise in collecting ~~variables~~ data than autonomous sensors such as the BGC-Argo ~~(Johnson et al. (2013); Johnson and Claustre (2016))~~(Johnson et al., 2013; Johnson and Claustre, 2016). However, while MLP architectures ~~can~~ have been demonstrated to provide good training and test errors ~~(Pietropolli et al. (2023a); Fourrier et al. (2020); Bittig et al. (2018); Sauzède et al. (2017)), they have been found to~~ for point-wise input and output (Pietropolli et al., 2023a; Fourrier et al., 2020; Bittig et al., 2018; Sauzède et al., 2017) , they can exhibit higher errors when predicting BGC-Argo profiles~~(Pietropolli et al. (2023a))~~, as demonstrated in Pietropolli et al. (2023a) and Appendix B. In contrast, the PPCon architecture, which relies directly on BGC-Argo float measurements for the training, showed very good test and external validation performances.

However, it should be noted that an intrinsic measurement error is introduced by the higher uncertainty of the variables measured throughout the autonomous sensors. We alleviated this limitation by using only DT and high-quality checked Argo and BGC-Argo floats data; however, the use of the present PPCon in operational oceanography ~~(Le Traon et al. (2021); Cossarini et al. (2019))~~ (Le Traon et al., 2021; Cossarini et al., 2019) should be considered cautiously given the lower reliability of Adjusted or near real-time (NRT) Argo data. According to the analysis conducted by Mignot et al. (2019), the BGC-Argo float data for ~~oxygen,~~ nitrate, and chlorophyll ~~concentrations~~ exhibit RMSE values evaluated at ~~$5.1 \pm 0.8 \mu mol/kg$, $0.25 \pm 0.07 \mu mol/kg$, and $0.03 \pm 0.01 mg/m^3$~~ $0.25\ mmol/m^3$, and $0.03\ mg/m^3$, respectively. On the other hand, ~~we have demonstrated that the RMSE for the PPCon architecture is 0.61~~PPCon architecture produced BCG-Argo profile reconstruction with RMSE values of $0.52\ mmol/m^3$ and $0.08\ mg/m^3$, for nitrate and chlorophyll, respectively. Therefore, a research question pertains to un-

derstanding how the measurement error of the float instrument impacts the performance of the PPCon architecture, and how to estimate an overall error that combines the contribution of the instrument error and the error associated with the PPCon.

Although both MLPs and PPCon employ similar input information (date, geolocation, temperature, oxygen, and salinity), their treatment of this data differs significantly. ~~MLPs~~ While the current MLP applications process the input and output as ~~discrete datapoints, while~~ point-wise data, PPCon utilizes vector representations of the vertical profiles. This approach ~~was necessitated to effectively exploit~~ effectively exploits the potential of a 1D CNN, which intrinsically preserves the characteristic profile shape of the input and output variables ~~Kiranyaz et al. (2021)~~(Kiranyaz et al., 2021). When comparing the predictive performance of these techniques in generating vertical profiles from float data, distinct differences emerge. MLPs ~~tend to produce profiles characterized by apparently artificial discontinuityand jumps~~can produce profiles affected by artificial discontinuity, while the profiles generated by PPCon exhibit a smoother and more realistic appearance (~~Figure 3). This improvement is confirmed also by the RMSE , which is lower when using the PPCon model ($RMSE_{PPCon} = 0.61$) compared to the state of the art of MLP architectures ($RMSE_{MLP} = 0.87$ according to Pietropolli et al. (2023a)~~Appendix B). Additionally, the RMSE values computed on the reconstructed nitrate profiles of the test sub-set confirm the better performance of the 1D CNN approach with respect to an MLP approach trained on point-wise data (Appendix B).

Moreover, the posterior study that we conducted shows that there is no significant variation in the error across different geographic areas and seasons of the year (Table 6-7), confirming that PPCon can be successfully applied to all the float profiles collected in the Mediterranean basin.

Specifically, the PPCon architecture serves as a valuable tool for significantly enriching the BGC-Argo dataset. This becomes useful as ocean observing systems - while essential for the monitoring of the health of the marine ecosystem ~~(Euzen et al. (2017))~~ (Euzen et al., 2017) - have considerable limitations given their sparse and scarce space-temporal coverage. Surface satellite observations are limited by cloud ~~covers and incomplete swaps~~ coverage and incomplete swaths of satellite sensors ~~(Donlon et al. (2012))~~(Donlon et al., 2012), while profiling the ocean interior is limited by the capacity of deploying and retrieving sensors and measurements with sufficient coverage. Gap filling and interpolation of satellite observations ~~(Volpe et al. (2018) , Sammartino et al. (2020); Alvera-Azcárate et al. (2005))~~ (Volpe et al., 2018; Sammartino et al., 2020; Alvera-Azcárate et al., 2005) are nowadays consolidated practices to provide gap-free and high-level products ~~(Barth et al. (2020); Sauzède et al. (2016))~~(Barth et al., 2020; Sauzède et al., 2016). Our PPCon architecture presents a valuable approach to harness the potential of the Argo and BGC-Argo network by enabling the synthetic generation of essential variables (chlorophyll, nitrate, and bbp700) even when these costly sensors are not present in the deployed floats. ~~The~~ For instance, the application of PPCon on ~~the GDACs BCG-Argo float dataset (spanning from 2015~~ Argo and oxygen profiles in the Mediterranean Sea for the period from 2013 to 2020 ~~)~~ enabled the generation of 5234 ~~synthetic nitrateprofiles~~(nitrate), 3879 ~~chlorophyll profiles, and~~ (chlorophyll), 3307 (bbp700) synthetic profiles, which means doubling the chlorophyll and bbp700 BGC-Argo profiles and more than tripling those of nitrate. Enhancing the float dataset through the inclusion of reconstructed nutrient profiles (and possibly other biogeochemical variables) has been ~~proved~~ proven successful in observing system simulation experiments ~~(Ford (2021); Yu et al. (2018))~~ (Ford, 2021; Yu et al., 2018) and in real assimilation numerical experiments (Amadio et al., (2023)). In particular, the assimilation of reconstructed profiles effectively corrects a widespread positive bias observed in the Operational System for Short-Term

Forecasting of the Biogeochemistry of the Mediterranean (MedBFM), and the addition of the reconstructed profiles increases the spatial impact of the BCG-argo network from 20% to 45% (Amadio et al., (2023))

7 Conclusions

This paper presents a novel approach for reconstructing low-sampled variables, namely nitrate, chlorophyll, and bbp700, using high-sampled variables such as date, geolocation, temperature, salinity and oxygen. The introduced model, named PPCon, utilizes a spatial-aware 1D CNN architecture that effectively learns the characteristic shape of the vertical profile, enabling precise and smooth reconstructions. PPCon represents a ~~notable advancement over previous techniques relying on MLPs~~potential advancement in predicting BCG-Argo profiles over previous MLP applications, which operate on ~~point-to-point~~ point-wise input and output~~, making it challenging to generate continuous curves when forecasting complete vertical profiles.~~ .

The training dataset consists of a collection of BGC-Argo float measurements in the Mediterranean basin. The proposed architecture has been specifically designed to handle both ~~punctual~~ point-wise and vectorial input, with careful tuning of the architecture and loss function for the task. An extensive hyperparameter tuning phase has been conducted to ensure the best architecture for each variable.

To evaluate the accuracy of the profiles generated by the PPCon architecture, both quantitative metrics and visual representations of the results have been provided. Additionally, the method has been validated on an external dataset to verify its generability. The results confirm the model's ability to predict high-quality synthetic profiles, with particularly accurate predictions for the nitrate variable, followed by chlorophyll, and lastly, bbp700. ~~The RMSE for nitrate reconstruction is reduced from $RMSE_{MLP} = 0.87$ to $RMSE_{PPCon} = 0.61$.~~

PPCon demonstrates its capacity to capture and learn distinct typical shapes in the profiles, which characterize the inferred variables across different seasons and geographic areas. Detailed error analysis confirms the model's robust performance, accounting for seasonal and regional variations, suggesting that PPCon's ability to learn these differences can make it successful for broader-scale training beyond the Mediterranean basin. Furthermore, the model exhibits accurate performance on an external validation dataset, confirming its potential for generalization.

Code and data availability. The datasets, source code, and model implementation used in this study are publicly available at https://github.com/gpietrop/PPCon for interested readers to access and replicate the results presented in this paper (Pietropolli et al., 2023b). In the present work, we present an optimized version of the architecture for the specific dataset of the Mediterranean Sea, but the release of the PPCon code allows arbitrary adjustments of the architecture.

Appendix A: Extended Results

Figure A1 presents a comparison between the mean values of the PPCon predicted profiles and the mean values of the sampled measurements obtained from the float instruments in the test set. ~~This comparison encompasses all the variables examined.~~ The

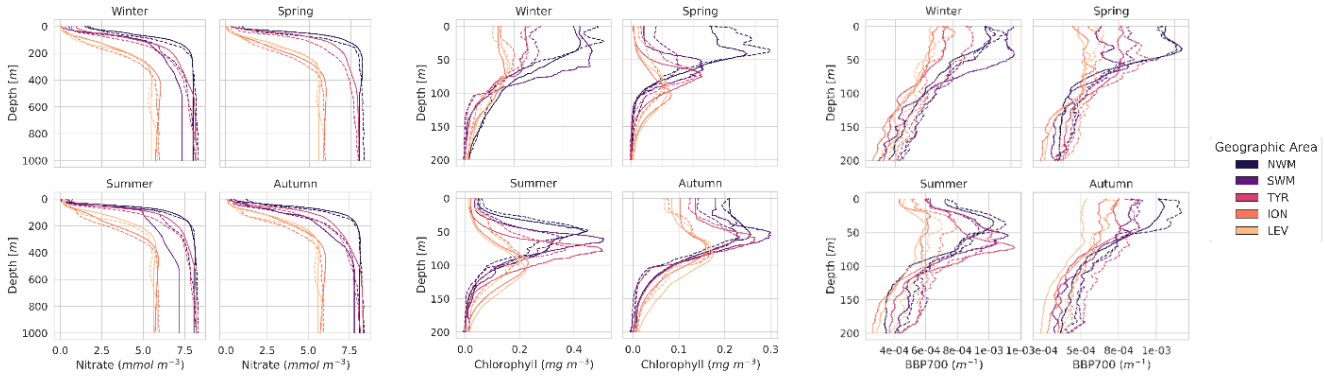

Figure A1. Comparison of the mean of PPCon predicted profiles with the mean of sampled values measured by the float instruments in the test set. Results are divided among different geographic areas, ~~for all~~ the ~~variables investigated~~dashed lines represent sampled values, while continuous lines represent PPCOn predictions.~~The mean is computed over all the profiles belonging to a fixed geographic area and a fixed season.~~

	PPCon	CANYON-Med (Fourrier et al., 2020)	MLP (Pietropolli et al., 2023a)
Nitrate RMSE $[mmol/m^3]$	0.52	0.78	0.98

Table B1. RMSE of the three ML architectures computed over the nitrate profiles of the sub-set BGC-Argo dataset of the test phase.

mean values are computed based on the profiles within a specific geographic area and season. These profiles serve as additional indicators to assess the reliability of predictions within different frameworks, providing valuable insights into the precision of predictions at various depth levels. These results confirm the previous observations discussed in Section 5, particularly the finding that the prediction quality is superior for the nitrate~~variables~~, followed by chlorophyll, and lastly, bbp700. Additionally, an interesting characteristic of the PPCon prediction is its higher quality in deep water compared to surface water. This can be attributed to the higher variability of profiles in the surface water, making it more challenging for the neural network to accurately capture the diverse shapes.

Appendix B: Comparison between reconstructed nitrate profiles by PPCon and MLP architectures

The present appendix aims to show the performance of three different ML architectures to reconstruct nitrate profiles that use Argo profiles of temperature, salinity and BGC-Argo profiles of oxygen. The three ML architectures are: the 1D CNN of the present work (PPCon), MLP trained on point-wise data from Emodnet (Pietropolli et al., 2023a) and MLP trained on point-wise data (Fourrier et al., 2020). The comparison is done on the sub-set of profiles used in the test phase. Figure B1 shows some measured and reconstructed float profiles. The visual comparison reveals the higher performance of PPCon to match the shape of the measured profiles (e.g., depth and intensity of the nitracline) and to reproduce the nitrate values of the deepest

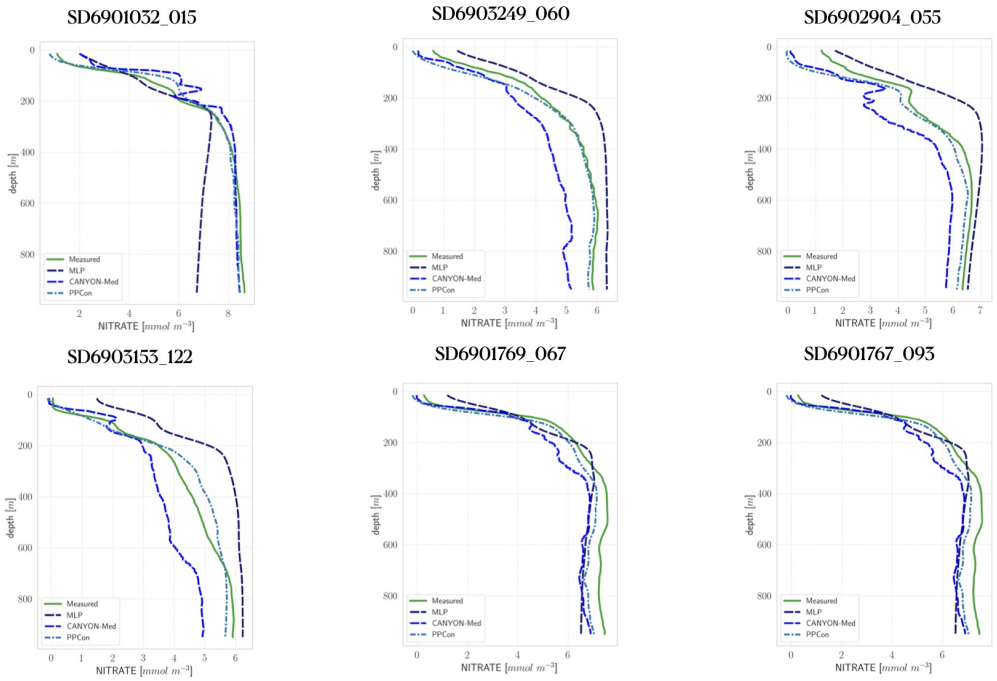

Figure B1. Nitrate profiles from BGC-Argo dataset (green, measured) and reconstructed by PPCon (cyan dashed line), MLP as in (Pietropolli et al., 2023a) (purple dashed line) and CANYON-Med (dark blue dashed line). Profiles are selected from the sub-set used in the test phase of the present work. Float positions are as follows: 6901032 in NWM, 6903249 and 6903153 in ION, 69032904 in LEV and 6901767 in TYR and 6901769 in SWM.

part of the profiles observed in the different Mediterranean sub-regions. The quantitative assessment of the performance of the three ML architectures is shown in Table B1 that reports the RMSE computed over all profiles of the sub-set used in the test phase. RMSE of the reconstructed profile by PPCon is more than 30% lower than that computed on the MLP reconstructions.

Author contributions. The authors contributed to this work as follows: GP, LM, and GC conceptualized the research, GP conducted data curation and analysis, GP developed the computational model, GP and GC contributed to the experimental design and methodology, LM and GC provided critical insights and supervision, GP performed validation experiments. All authors reviewed and approved the final version of the paper.

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

Acknowledgements. Data were collected and made freely available by the International Argo Program and the national programs that contribute to it (https://argo.ucsd.edu, https://www.ocean-ops.org). The Argo Program is part of the Global Ocean Observing System ~~(Argo (2000))~~(Argo, 2000).

This research has been partly supported by the MED-MFC "Mediterranean Monitoring and Forecasting Centre" of CMEMS, which is implemented by Mercator Ocean International within the framework of a delegation agreement with the European Union (Ref. n. 21002L5-COP-MFC MED-5500).

References

Ahmad, H.: Machine learning applications in oceanography, Aquatic Research, 2, 161–169, 2019.

Albawi, S., Mohammed, T. A., and Al-Zawi, S.: Understanding of a convolutional neural network, in: 2017 international conference on engineering and technology (ICET), pp. 1–6, Ieee, 2017.

Alvera-Azcárate, A., Barth, A., Rixen, M., and Beckers, J.-M.: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature, Ocean Modelling, 9, 325–346, 2005.

Amadio, C., TERUZZI, A., Feudale, L., BOLZON, G., DI BIAGIO, V., Lazzari, P., Álvarez, E., Coidessa, G., Salon, S., and COSSARINI, G.: Mediterranean Quality checked BGC-Argo 2013-2022 dataset, https://doi.org/10.5281/zenodo.10391759, 2023.

Argo: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC), https://doi.org/https://doi.org/10.17882/42182, last visited on August 2022., 2000.

Baldi, P. and Sadowski, P. J.: Understanding dropout, Advances in neural information processing systems, 26, 2013.

Barth, A., Alvera-Azcárate, A., Licer, M., and Beckers, J.-M.: DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations, Geoscientific Model Development, 13, 1609–1622, 2020.

Bittig, H., Wong, A., and Plant, J.: BGC-Argo synthetic profile file processing and format on Coriolis GDAC. Version 1.1, 2019.

Bittig, H. C., Steinhoff, T., Claustre, H., Fiedler, B., Williams, N. L., Sauzède, R., Körtzinger, A., and Gattuso, J.-P.: An alternative to static climatologies: Robust estimation of open ocean CO2 variables and nutrient concentrations from T, S, and O2 data using Bayesian neural networks, Frontiers in Marine Science, 5, 328, 2018.

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

Bottou, L.: Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers, pp. 177–186, Springer, 2010.

Campbell, L. M., Gray, N. J., Fairbanks, L., Silver, J. J., Gruby, R. L., Dubik, B. A., and Basurto, X.: Global oceans governance: New and emerging issues, Annual review of environment and resources, 41, 517–543, 2016.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P.: Natural language processing (almost) from scratch, Journal of machine learning research, 12, 2493–2537, 2011.

Coppini, G., Clementi, E., Cossarini, G., Salon, S., Korres, G., Ravdas, M., Lecci, R., Pistoia, J., Goglio, A. C., Drudi, M., et al.: The Mediterranean forecasting system. Part I: evolution and performance, EGUsphere, pp. 1–50, 2023.

Cossarini, G., Mariotti, L., Feudale, L., Mignot, A., Salon, S., Taillandier, V., Teruzzi, A., and d'Ortenzio, F.: Towards operational 3D-Var assimilation of chlorophyll Biogeochemical-Argo float data into a biogeochemical model of the Mediterranean Sea, Ocean Modelling, 133, 112–128, 2019.

Dall'Olmo, G., TVS, U. B., Bittig, H., Boss, E., Brewster, J., Claustre, H., Donnelly, M., Maurer, T., Nicholson, D., Paba, V., et al.: Real-time quality control of optical backscattering data from Biogeochemical-Argo floats, Open Research Europe, 2, 118, 2022.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., and Wimmer, W.: The operational sea surface temperature and sea ice analysis (OSTIA) system, Remote Sensing of Environment, 116, 140–158, 2012.

d'Ortenzio, F., Lavigne, H., Besson, F., Claustre, H., Coppola, L., Garcia, N., Laës-Huon, A., Le Reste, S., Malardé, D., Migon, C., et al.: Observing mixed layer depth, nitrate and chlorophyll concentrations in the northwestern Mediterranean: A combined satellite and NO3 profiling floats experiment, Geophysical Research Letters, 41, 6443–6451, 2014.

D'ortenzio, F., Taillandier, V., Claustre, H., Prieur, L. M., Leymarie, E., Mignot, A., Poteau, A., Penkerc'h, C., and Schmechtig, C. M.: Biogeochemical Argo: The test case of the NAOS Mediterranean array, Frontiers in Marine Science, 7, 120, 2020.

Euzen, A., Gaill, F., Lacroix, D., and Cury, P.: The ocean revealed, CNRS Éditions, Paris, 2017.

Ford, D.: Assimilating synthetic Biogeochemical-Argo and ocean colour observations into a global ocean model to inform observing system design, Biogeosciences, 18, 509–534, 2021.

Fourrier, M., Coppola, L., Claustre, H., D'Ortenzio, F., Sauzède, R., and Gattuso, J.-P.: A regional neural network approach to estimate water-column nutrient concentrations and carbonate system variables in the Mediterranean Sea: CANYON-MED, Frontiers in Marine Science, 7, 620, 2020.

Goh, G. B., Hodas, N. O., and Vishnu, A.: Deep learning for computational chemistry, Journal of computational chemistry, 38, 1291–1307, 2017.

Goldstein, E. B., Coco, G., and Plant, N. G.: A review of machine learning applications to coastal sediment transport and morphodynamics, Earth-science reviews, 194, 97–108, 2019.

Goodwin, J. D., North, E. W., and Thompson, C. M.: Evaluating and improving a semi-automated image analysis technique for identifying bivalve larvae, Limnology and Oceanography: Methods, 12, 548–562, 2014.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks, Pattern recognition, 77, 354–377, 2018.

Jayne, S. R., Roemmich, D., Zilberman, N., Riser, S. C., Johnson, K. S., Johnson, G. C., and Piotrowicz, S. R.: The Argo program: present and future, Oceanography, 30, 18–28, 2017.

Johnson, K. and Claustre, H.: Bringing biogeochemistry into the Argo age, Eos, Transactions American Geophysical Union, 2016.

Johnson, K. S., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Swift, D. D., and Riser, S. C.: Long-term nitrate measurements in the ocean using the In Situ Ultraviolet Spectrophotometer: sensor integration into the Apex profiling float, Journal of Atmospheric and Oceanic Technology, 30, 1854–1866, 2013.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J.: 1D convolutional neural networks and applications: A survey, Mechanical systems and signal processing, 151, 107 398, 2021.

Krogh, A.: What are artificial neural networks?, Nature biotechnology, 26, 195–197, 2008.

Le Traon, P., Abadie, V., Ali, A., Behrens, A., Staneva, J., Hieronymi, M., and Krasemann, H.: The Copernicus Marine Service from 2015 to 2021: six years of achievements, 2021.

Li, Z., Liu, Z., and Lu, S.: Global Argo data fast receiving and post-quality-control system, in: IOP Conference Series: Earth and Environmental Science, vol. 502, p. 012012, IOP Publishing, 2020.

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects, IEEE transactions on neural networks and learning systems, 2021.

Mignot, A., Claustre, H., Uitz, J., Poteau, A., d'Ortenzio, F., and Xing, X.: Understanding the seasonal dynamics of phytoplankton biomass and the deep chlorophyll maximum in oligotrophic environments: A Bio-Argo float investigation, Global Biogeochemical Cycles, 28, 856–876, 2014.

Mignot, A., d'Ortenzio, F., Taillandier, V., Cossarini, G., and Salon, S.: Quantifying observational errors in Biogeochemical-Argo oxygen, nitrate, and chlorophyll a concentrations, Geophysical Research Letters, 46, 4330–4337, 2019.

Miloslavich, P., Seeyave, S., Muller-Karger, F., Bax, N., Ali, E., Delgado, C., Evers-King, H., Loveday, B., Lutz, V., Newton, J., et al.: Challenges for global ocean observation: the need for increased human capacity, Journal of Operational Oceanography, 12, S137–S156, 2019.

Mori, U., Mendiburu, A., Keogh, E., and Lozano, J. A.: Reliable early classification of time series based on discriminating the classes over time, Data mining and knowledge discovery, 31, 233–263, 2017.

Munk, W.: Oceanography before, and after, the advent of satellites, in: Elsevier Oceanography Series, vol. 63, pp. 1–4, Elsevier, 2000.

O'Shea, K. and Nash, R.: An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458, 2015.

Pietropolli, G., Cossarini, G., and Manzoni, L.: GANs for Integration of Deterministic Model and Observations in Marine Ecosystem, in: Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings, pp. 452–463, Springer, 2022.

Pietropolli, G., Manzoni, L., and Cossarini, G.: Multivariate Relationship in Big Data Collection of Ocean Observing System, Applied Sciences, 13, 5634, 2023a.

Pietropolli, G., Manzoni, L., and Gianpiero, C.: PPCon 1.0: Biogeochemical Argo Profile Prediction with 1D Convolutional Networks, https://doi.org/10.5281/zenodo.8369573, 2023b.

Rasamoelina, A. D., Adjailia, F., and Sinčák, P.: A review of activation function for artificial neural network, in: 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 281–286, IEEE, 2020.

Sammartino, M., Buongiorno Nardelli, B., Marullo, S., and Santoleri, R.: An artificial neural network to infer the Mediterranean 3D chlorophyll-a and temperature fields from remote sensing observations, Remote Sensing, 12, 4123, 2020.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A.: How does batch normalization help optimization?, Advances in neural information processing systems, 31, 2018.

Sauzède, R., Claustre, H., Uitz, J., Jamet, C., Dall'Olmo, G., d'Ortenzio, F., Gentili, B., Poteau, A., and Schmechtig, C.: A neural network-based method for merging ocean color and Argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscattering coefficient, Journal of Geophysical Research: Oceans, 121, 2552–2571, 2016.

Sauzède, R., Bittig, H. C., Claustre, H., Pasqueron de Fommervault, O., Gattuso, J.-P., Legendre, L., and Johnson, K. S.: Estimates of water-column nutrient concentrations and carbonate system parameters in the global ocean: A novel approach based on neural networks, Frontiers in Marine Science, 4, 128, 2017.

Shan, C., Zhang, J., Wang, Y., and Xie, L.: Attention-based end-to-end speech recognition on voice search, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4764–4768, IEEE, 2018.

Tang, W., Long, G., Liu, L., Zhou, T., Jiang, J., and Blumenstein, M.: Rethinking 1d-cnn for time series classification: A stronger baseline, arXiv preprint arXiv:2002.10061, pp. 1–7, 2020.

Taud, H. and Mas, J.: Multilayer perceptron (MLP), Geomatic approaches for modeling land change scenarios, pp. 451–455, 2018.

Teruzzi, A., Bolzon, G., Feudale, L., and Cossarini, G.: Deep chlorophyll maximum and nutricline in the Mediterranean Sea: emerging properties from a multi-platform assimilated biogeochemical model experiment, Biogeosciences, 18, 6147–6166, 2021.

Volpe, G., Nardelli, B. B., Colella, S., Pisano, A., and Santoleri, R.: Operational Interpolated Ocean Colour Product in the Mediterranean Sea, New Frontiers in Operational Oceanography, pp. 227–244, 2018.

Yu, L., Fennel, K., Bertino, L., El Gharamti, M., and Thompson, K. R.: Insights on multivariate updates of physical and biogeochemical ocean variables using an Ensemble Kalman Filter and an idealized model of upwelling, Ocean Modelling, 126, 13–28, 2018.

Zeiler, M. D.: Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701, 2012.

Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net, Journal of the royal statistical society: series B (statistical methodology), 67, 301–320, 2005.