


Article

# Modelling the Vertical Distribution of Phytoplankton Biomass in the Mediterranean Sea from Satellite Data: A Neural Network Approach

Michela Sammartino <sup>1,3,\*</sup>, Salvatore Marullo <sup>5,6</sup> , Rosalia Santoleri <sup>2</sup> and Michele Scardi <sup>3,4</sup><sup>1</sup> Istituto di Scienze Marine, Consiglio Nazionale delle Ricerche (ISMAR-CNR), 00133 Rome, Italy<sup>2</sup> Istituto di Scienze Marine, Consiglio Nazionale delle Ricerche (ISMAR-CNR), 30122 Venice, Italy; rosalia.santoleri@cnr.it<sup>3</sup> Dipartimento di Biologia, Università di Roma “Tor Vergata”, 00133 Rome, Italy; mscardi@mclink.it<sup>4</sup> Consorzio Nazionale Interuniversitario Per Le Scienze Del Mare (CoNISMa), 00196 Rome, Italy<sup>5</sup> Agenzia nazionale per le nuove tecnologie, l’energia e lo sviluppo economico sostenibile (ENEA), Centro Ricerche Frascati, 00044 Frascati, Italy; salvatore.marullo@enea.it<sup>6</sup> Istituto di Scienze dell’Atmosfera e del Clima, Consiglio Nazionale delle Ricerche (ISAC-CNR), 00133 Rome, Italy

\* Correspondence: michela.sammartino@artov.isac.cnr.it; Tel.: +39-064993-4262

Received: 30 September 2018; Accepted: 14 October 2018; Published: 21 October 2018



**Abstract:** Knowledge of the vertical structure of the bio-chemical properties of the ocean is crucial for the estimation of primary production, phytoplankton distribution, and biological modelling. The vertical profiles of chlorophyll-*a* (Chla) are available via *in situ* measurements that are usually quite rare and not uniformly distributed in space and time. Therefore, obtaining estimates of the vertical profile of the Chla field from surface observations is a new challenge. In this study, we employed an Artificial Neural Network (ANN) to reconstruct the 3-Dimensional (3D) Chla field in the Mediterranean Sea from surface satellite estimates. This technique is able to reproduce the highly nonlinear nature of the relationship between different input variables. A large *in situ* dataset of temperature and Chla calibrated fluorescence profiles, covering almost all Mediterranean Sea seasonal conditions, was used for the training and test of the network. To separate sources of errors due to surface Chla and temperature satellite estimates, from errors due to the ANN itself, the method was first applied using *in situ* surface data and then using satellite data. In both cases, the validation against *in situ* observations shows comparable statistical results with respect to the training, highlighting the feasibility of applying an ANN to infer the vertical Chla field from surface *in situ* and satellite estimates. We also analyzed the usefulness of our approach to resolve the Chla prediction at small temporal scales (e.g., day) by comparing it with the most widely used Mediterranean climatology (MEDATLAS). The results demonstrated that, generally, our method is able to reproduce the most reliable profile of Chla from synoptical satellite observations, thus resolving finer spatial and temporal scales with respect to climatology, which can be crucial for several marine applications. We demonstrated that our 3D reconstructed Chla field could represent a valid alternative to overcome the absence or discontinuity of *in situ* sampling.

**Keywords:** ocean colour; SST; chlorophyll-*a* vertical profile; Mediterranean Sea; Artificial Neural Network; error-backpropagation

## 1. Introduction

Phytoplankton constitutes the autotrophic component of the plankton biomass.

As a bio-indicator, the alteration of algal community composition and distribution provides information on some of the principal climate-driven effects on environmental forcing, and consequently, on marine ecosystem equilibrium [1–4].

One of the most common proxies used to estimate the phytoplankton biomass and trophic condition of seawater is chlorophyll-*a* concentration (Chla) [5–8].

The monitoring of Chla pattern at global and regional scales is pursued through several approaches (e.g., oceanographic cruises, modelling studies, and satellite sensors).

On the surface, the monitoring of Chla has strongly improved in terms of time and spatial coverage with the introduction of satellite sensors. In fact, nowadays, the low cost and high resolution of satellite imagery have improved the comprehension of the marine environment, providing information on several environmental parameters, e.g., surface biomass, Sea Surface Temperature (SST), Photosynthetically Active Radiation (PAR), etc. [9–11].

However, satellite observations give us information only about the near surface pattern. In fact, considering the light penetration depth in the ocean, satellite Chla represents only one fifth of the total Chla content within the euphotic zone, ignoring the algal biomass variability at deeper layers [12]. The acquisition of information about the vertical biomass structure is more challenging and often time consuming. Accurate estimates of Chla are achieved through the collection of discrete water samples and relative analyses. Today, the most popular techniques used to quantify the Chla concentration in a sample are, e.g., spectrophotometric, chromatographic measures (High-Performance Liquid Chromatography, HPLC), or by a fluorimeter.

The characterization of Chla vertical field is principally based on *in situ* samples, profiling floats, autonomous platforms, or fluorescence profiles, which are, in turn, calibrated with discrete samples that are not easy to collect.

Nevertheless, if compared with ocean-color satellite sensors, the space-time coverage of ship-based sampling is definitely lower, and its economic effort hinders the collection of a large number of data, preventing a complete description of the Chla variability on a large scale [13]. For these reasons, it becomes necessary to find a complementary use of the available instruments and data, in order to supply the discontinuous nature of the actual *in situ* databases.

In this study, we employed an Artificial Neural Network (ANN) to reconstruct the 3D Chla field in the Mediterranean Sea from surface satellite estimates.

Today, the interest in the possible linkage of surface information with the vertical profile of oceanic variables has increased. Several works propose different approaches to extrapolate surface information to subsurface layers both for physical and biological processes. Some of them, mostly focused on the vertical reconstruction of physical variables, differ from each other for the parameters, statistical approach, and area of investigation. For example, Buongiorno Nardelli et al. [14], through the use of statistical methods as the Coupled Pattern Reconstruction (CPR) and multivariate Empirical Orthogonal Function reconstruction (mEOF-R), are able to extrapolate the subsurface geostrophic velocities from surface and integrated data, for the case study of the Sicily Channel in the Mediterranean Sea. Other methods, used for the extrapolation of physical processes to deeper layers, are based on machine learning techniques as, e.g., shown in Gueye et al. [15], in which an unsupervised neural network (Self-Organizing-Map, SOM) is applied on physical surface measurements of the tropical Atlantic Ocean to infer the most probable salinity vertical profile. A similar neural approach was also adopted by Sauzède et al. [16] to estimate one of the most important optical properties of the water column, the backscattering coefficient. Here, a Multi-Layer Perceptron (MLP) is adopted to infer the backscattering coefficient's vertical profile from the combined use of ocean-color measurements and autonomous floats (Argo data), employed for the collection of temperature, salinity, and current data in ice-free oceans.

The chance to extrapolate the vertical structure of an oceanic variable from its surface pattern varies according to several factors, e.g., the nature of the analyzed parameters or the seasonal and trophic regimes of the area of interest.

For example, in most of temperate areas, such as the Mediterranean Sea, the strong seasonality determines the well-known seasonal dynamics of Chla shape, making the Chla vertical structure rather predictable. This further strengthens the attempt to infer the vertical dynamics from surface data also for biological processes.

The first studies on the 3-Dimensional (3D) reconstruction of biological parameters from surface estimates are dated back to 1989, when Morel and Berthon [12], adopting a statistical relationship between the pigment concentration in the upper layer and within the water column, determined seven trophic situations and computed the vertical profile for each of them. Some years later, following the Morel and Berthon [12] model, Uitz et al. [17] suggested an empirical parameterization, based on a big global dataset, to infer, from near-surface Chla values, the column-integrated phytoplankton biomass, its vertical distribution, and its community composition.

However, most of the works previously cited are not based on machine learning techniques, whose use has increased in recent years. Today, in fact, most of these techniques are employed to extend the biological surface information to deeper layers. One example is Charantonis et al. [18], who presented a combined use of a Self-Organizing Map with the Hidden Markov Models to infer 3-Dimensional Chla fields starting from 2-Dimensional (2D) imaging of several variables (surface Chla -Chla<sub>surf</sub>, Sea Surface Elevation- SSH, radiation, and wind). Furthermore, a cascade of neural networks, trained on the quasi-Newton method, was proposed in Dall Cortivo et al. [19] to estimate the sub-surface Chla concentration in Case 1 waters from the upwelling radiation. A similar attempt to infer the vertical Chla profile, by using a Multi-Layer-Perceptron (MLP), was shown in Sauzède [20], in which, differing from our approach, the output is predicted from surface ocean-color estimates and depth-resolved physical properties, derived from temperature-salinity profiles measured by Argo profiling floats.

Beyond the Chla variability forecast, these techniques were also employed in the primary production estimates, becoming a valid alternative to well-known empirical models. Richardson et al. [21], for example, used an unsupervised network (SOM) to classify the vertical Chla shape from known satellite parameters and then to compute the column-integrated primary production.

In Scardi [22], the Chla was considered as one of the neural network co-predictors (in addition to surface irradiance, euphotic zone depth, light extinction coefficient, and depth) used to estimate vertically integrated primary production in marine estuaries environment of the USA coast. A depth-resolved model of marine primary production was, instead, proposed in Scardi [23], and, more recently, in Mattei et al. [24] for the Gulf of Naples and Global Oceans, respectively.

In this context, the estimate of the vertical Chla profile from surface observations remains one of the most unresolved oceanic issues, representing a new challenge for the modelling of biological processes occurring in the Mediterranean Sea. Therefore, taking into account the advantages provided by the use of remotely sensed data (high spatial and temporal resolution) and the accuracy of *in situ* profiling instrument, here, we propose, for our basin, a tool that is able to use both information in a complementary way. For this purpose, in this study, we advance a model where an Artificial Neural Network (ANN), trained with an error back-propagation algorithm ([25–27]), is combined with satellite variables in order to infer the vertical Chla profile, starting from sea-surface measurements only. The main aim of our manuscript is to propose and demonstrate that a specific machine learning technique such as the Artificial Neural Network can be employed to realize the extrapolation of surface information to deeper layers. The main potential of this approach relies on the ability to overcome the discontinuity of the sampling frequency of vertical biomass profiles, exploiting the high frequency and coverage of satellite measurements. In fact, one of the major advantage of Machine Learning approaches relies in their capability to adapt their inner structure in order to extract the non-linear functional relationship between the variables without *a priori* models, while only exploiting the numerical relations, thanks to a suitable algorithm [25]. This means that they do not require knowledge of the nature of the underlying process between considered variables to treat it mathematically.

They are usually employed in problems that are unsolvable for the deterministic physics [28] because of incomplete or unknown governing equations or unrealistically expensive computational costs.

Here, the neural network used to predict the Chla profile from surface data is a Multi-Layer-Perceptron, trained by an error back-propagation algorithm (BPN) ([29]; a brief description of the network structure and functioning will be presented in Section 2.3).

The BPNs are data-driven models that are able to identify the existing underlying relationship between the input and the output throughout training examples [28].

In our case, the availability of a large Mediterranean dataset of *in situ* variables that are useful for training has made this basin the most suitable study area to show the applicability of our network to a regional scale.

This work aims to describe a novel approach, which represents the first attempt at creating a working alternative to classical models and overcoming the discontinuous nature of *in situ* sampling. In Section 2 (Materials and methods), we will describe the regional dataset used for the training and testing of the network, its configuration, and the processing of its inputs. Section 3 will present the results of the calibration and the test of the network performance, first applied on *in situ* surface data and then applied on satellite estimates. In this section, we will also give an example of BPN applicability in a specific case study. In Section 4 we will show the comparison of the network's performance against the widest-used Mediterranean climatology (MEDATLAS) in order to highlight the prediction capability of the network. Finally, Section 5 will include the main conclusions of our study and the future perspectives and implementation of the work, such as a future recalibration of the model according to an updated training dataset.

## 2. Data and Methods

Every neural network has two phases: a training phase and a test phase. Here, to employ the BPN to infer the Chla profile from its surface concentration, it was necessary to create a reference database, for the training and testing of the network learning. In our model, the BPN Chla profiles have been reconstructed by using either surface data acquired *in situ* or remotely sensed from satellite sensors and, separately, validated against *in situ* Chla profiles.

### 2.1. In Situ Database and Processing

*In situ* database comprises all available data of concurrent and quality-controlled chlorophyll-*a* and temperature profiles, collected during 20 oceanographic cruises in the Mediterranean Sea, covering different seasons, during the period 1998 to 2015 (Figure 1). Most of them were carried out by the GOS group (Global Ocean Satellite monitoring and marine ecosystem studies) of the Italian National Research Council (CNR). The number of acquired profiles was 2710. Chlorophyll-*a* *in vivo* fluorescence profiles were acquired with a fluorometer installed on a Conductivity-Temperature-Depth (CTD) rosette.

All the fluorescence profiles were calibrated into chlorophyll-*a* using discrete water samples that were concurrently acquired in each oceanographic cruise.

The data passed all the SeaDataNet standards quality control (<https://www.seadatanet.org/Standards/Data-Quality-Control>), in addition to a package of internal procedures implemented at CNR that include:

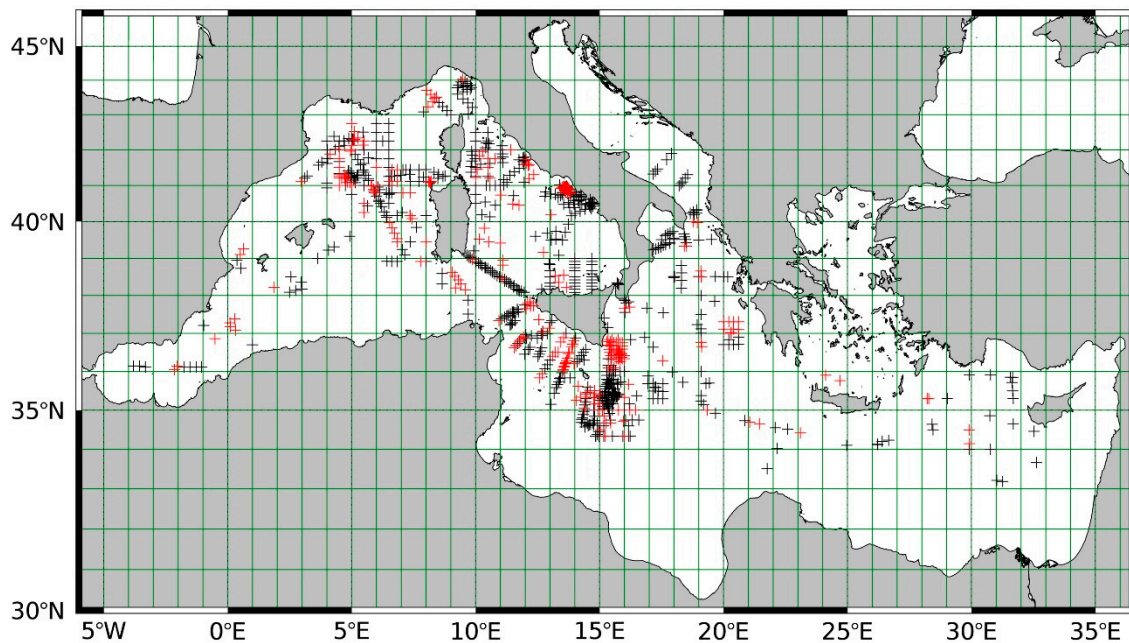
- Visual check (in order to ensuring the consistency of the database);
- Control on depth of the first acquisition (between 3 and 4 m and no deeper than 4 m, to avoid the use of the first acquired measures, which are usually most affected by noise and vessel oscillation);
- Control of missing values within the profile (missing values were linearly interpolated every meter from the deeper acquisition up to 3 m);
- Control on deeper acquisition (only profiles in region deeper than 150 m were retained);
- Profiles containing too few data samples were eliminated in order to avoid over-interpolation.



Same requirements were used for temperature profiles acquired through the CTD sensor. The closer CTD acquisition to the sea surface was assumed as *in situ* Sea Surface temperature.

After this quality control, from the initial 2710 profiles, 1213 remained available for the analysis. They constituted the final reference database used for the BPN development.

The dataset was divided into training and test sets containing 70% (849 stations) and 30% (364 stations) of the data, respectively, taking care that both training and test data points were uniformly distributed in space (Figure 1). In addition, a sub-set (~30%) was extracted from the training set to carry out the internal validation of the network's performance, during the training phase.



**Figure 1.** Spatial distribution of the stations used for the training set (black crosses in the image, about 70% of the total stations) and test set (red crosses in the image, corresponding to 30% of the total stations). The map grid ( $1^\circ \times 1^\circ$ ) is used to select the training and test set stations uniformly in the spatial domain.

## 2.2. Satellite Database Processing

Mediterranean 4 km spatial resolution Chla and SST maps (L4) (1998–2012), produced by the Copernicus Marine Environment Monitoring Service (CMEMS), were downloaded and used for the extraction of the input values.

Chla is a merged product (SeaWiFS, MERIS, and MODISAqua) that comes from the processing of the European Space Agency–Climate Change Initiative (ESA-CCI) input Remote Sensing Reflectance (Rrs) spectrum. The algorithm applied on Rrs to obtain Chla is MedOC4 [30]. It is a regional algorithm, which was developed by the GOS of CNR specifically for the optical properties of the Mediterranean Sea (more details about the processing in the Quality Information Document, <http://marine.copernicus.eu/>). In this context, the use of a regional algorithm improves the accuracy of the Chla estimates from space, reducing the error that usually affects the remote sensing variables. Moreover, the use of a merged product, as ESA-CCI-derived Chla, instead of a single sensor, improves the temporal and spatial coverage of the satellite dataset, thus increasing the number of available matchup points.

The SST data, here employed, are interpolated products, processed by the GOS of CNR and derived from the Pathfinder V5.2 (PFV52) Advanced Very High-resolution Radiometer (AVHRR) sensor, covering a period from 1981 until 2016 (more details about the processing in the Quality Information Document (<http://marine.copernicus.eu/>)).

These temperature and chlorophyll-*a* data were used to produce matchups of *in situ* and satellite estimates to compare BPN reconstruction using *in situ* or satellite sea surface parameters with *in situ* Chla profiles.

The matchup file includes both training and test points and was constructed using a  $3 \times 3$  pixels box centered on each *in situ* station. Only matchups in which all pixels in the  $3 \times 3$  Chla box were cloud free were included in the matchup file.

This process determined a number of 1103 matchup points containing *in situ* Chla observations from 1998 to 2012.

For the final comparison of the *in situ* profiles and those inferred by remote sensing data, satellite SST and Chla inputs were scaled into  $[0, 1]$  interval, as done for the training and test set.

### 2.3. Initialization of the Input Data

Since the mutable pattern of the Chla in the Mediterranean Sea may affect the learning rate and efficiency of the network, we decided to explore the training dataset features before starting the learning phase.

Therefore, the training set has undergone an explorative analysis, which aimed to characterize the features of the dataset. Moreover, the profiles were classified depending on different classes of surface Chla ranges, in order to recognize “typical” Chla shapes and their distribution within the dataset. This analysis revealed that some profiles were poorly represented within the training set. Specifically, it revealed that profiles typical of the bloom conditions or coastal zones turned out to be the less represented. Consequently, some of the rarest profiles were replicated in order to balance their occurrence in the learning phase [31].

During the learning phase of the network, a subset from training (~30%) was extracted to be used for the BPN internal validation. In fact, during the BPN learning, the training set is passed to the network to learn the intrinsic relationship among the chosen co-predictors, while the validation set is used to drive the learning and its duration. Once concluded the training phase, the BPN’s prediction skills are assessed on the test set.

In our BPN, the vertical Chla profile at fixed pressure levels (here, representative of the depth) is inferred from the following co-predictors: latitude, longitude, day of the year, surface temperature and Chla values. Here, the choice of using only surface data (Cha and sea temperature) and always readily available (latitude, longitude, and day of the year) is supported by the potential capability of the network to find the implicit relationship existing among them, even in non-linear cases.

Moreover, the chosen input parameters are strictly related to the ecology of the investigated problem. In fact, we know that spatial and temporal variables (day of the year/latitude and longitude) are important for providing implicit information about the season or the geo-location of the profile to be predicted, as well as the temperature, which alone or combined with the other inputs, deeply influences the vertical profile of Chla. Before the training, the input values were scaled into a  $[0, 1]$  interval, according to maximum/minimum ranges, respectively, which were bigger and smaller than the real ones (Table S1). This choice was made to distance the transformed values from the outers of the sigmoid activation function within each BPN node and thus improve the learning. The sampling date was projected on a circle, as follows:

$$\begin{aligned} day_1 &= \frac{1}{2} \left[ \cos \left( \frac{2\pi \text{ day of the year}}{365} \right) + 1 \right] \\ day_2 &= \frac{1}{2} \left[ \text{sen} \left( \frac{2\pi \text{ day of the year}}{365} \right) + 1 \right] \end{aligned}$$

Finally, in order to better represent the distribution of Chla and modulate the influence of highest concentrations on the training, the output target was log-transformed before being scaled into  $[0, 1]$  interval. In fact, considering the nature of the Chla, which usually assumes a normal distribution if treated in log-scale, the use of log-transformation allows to balance the importance of high Chla

concentrations with respect to the lowest ones, avoiding that the highest values of Chla drive the training toward themselves.

#### 2.4. BPN Training

The error back-propagation neural network (BPN), here employed, was developed by the research group of the Department of Biology, in the Laboratory of Experimental Ecology and Aquaculture, University of “Tor Vergata”, Rome (for more details on the BPN functioning see Appendix 1 in Scardi et al. [22]).

The back-propagation theory was presented for the first time by Werbos in 1974 [32] and then proposed to the research community by Rumelhart et al. [29].

This network is based on supervised learning, which allows the network to extract solely the relationship that exists between the input and the output [25]. The principal goal of a BPN is to create a generalized system, in which the error between  $\text{Output}_{\text{estimated}}$  and  $\text{Output}_{\text{target}}$  is minimized.

Our BPN is a Multilayer Perceptron, which consists of a series of layers (input-hidden-output) connected to each other by weights, responsible for the information flows from the input to the output. Each layer contains the non-linear elements (neurons or nodes) characterized by an activation function, which allows the estimate of the signal throughout the system, node by node.

During the training, a set of inputs/outputs of reference is provided to the network iteratively.

Here, the algorithm used to train the BPN is based on the backward propagation of the errors. Today, the back-propagation algorithm is widely used for convergence problem-solving. When the network provides an output, the synaptic weights are adjusted in order to reduce the error that exists between the estimated and target outputs. This weight adjustment is based on the theory of a gradient descent of the error on a surface [22,25].

After several tests aimed at selecting the optimal structure of the network, three different BPNs were considered [33]. They had the same number of hidden nodes but differed according to the number of epochs (defined as one weight update or training iteration), possible outliers to be discarded from the final error computation, and the learning constant (the step length of the weight correction).

The performances of the three networks were evaluated according to the lowest Mean Square Error (MSE, sum of the square deviations of the neural network outputs from the target values) and  $r^2$  gained in the internal validation, in addition to the statistics obtained on the test set. The best performing network was the Network 1, characterized by the lowest MSE and by the following configuration: 7-10-1 network, or 7 input nodes, 10 hidden nodes, and 1 output (Figure 2).

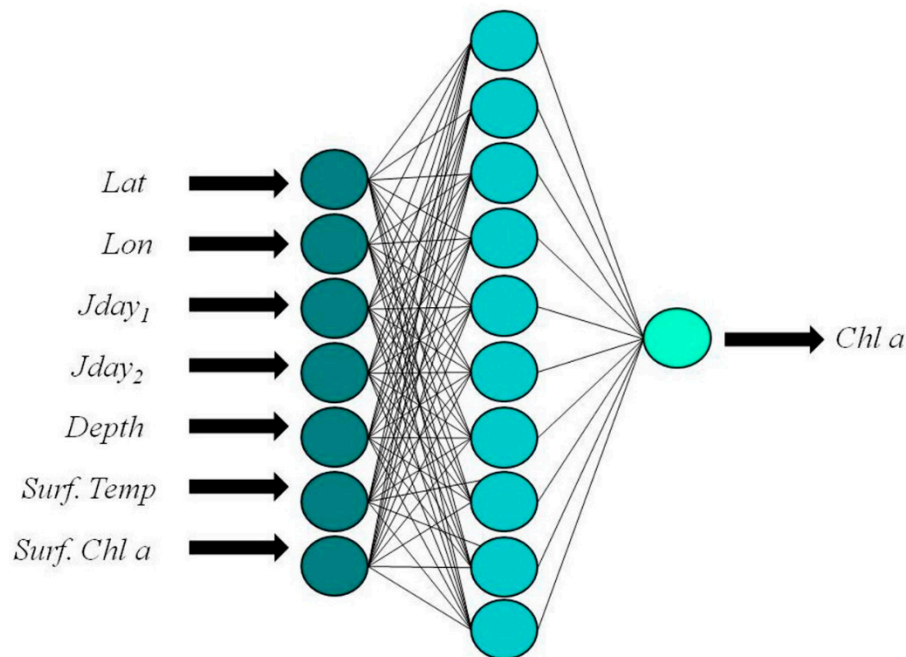
This BPN was characterized by 8000 maximum number of epochs, a learning constant of 0.7 and a momentum of 0.1. Moreover, to improve the generalization capability of the network, a small amount of Gaussian noise  $[-0.02, +0.02]$  was added to the input values. This technique known as “jittering” was aimed at introducing artificial training patterns (very close to the original ones) to disturb the system and at avoiding an overtraining of the network [27,34]. Today, this technique is widely employed and enables the network to learn to converge toward neighborhood values and not to fit precisely the original target, reducing its possible overfitting. Additionally, the weights were at first initialized randomly, and then a very small noise (0.01) was added to them to improve the network regularization.

During the learning phase, only a subset (50%) of the training set was randomly selected and passed to the BPN at each epoch, favoring the generalization.

The strategy adopted to interrupt the network training was an early stopping criterion. This approach is based on the minimization of the MSE against the validation set. When the error on validation set starts to increase and deviates from the MSE of the training set that still decreasing, the learning phase is stopped and the actual weights are saved. The use of this technique is widely used, in conjunction with other techniques such as jittering, to induce regularization in network training and avoiding possible overfitting.

One of the most important points to stress about the artificial neural network is that there is not a univocal configuration or best solution [22]. This is due to the random initialization of the synaptic weights. However, a neural network shows the great advantage of being able to manage non-linear problems, which are otherwise impossible to solve using conventional methods.

The number of units in the hidden layer was chosen on the basis of the performances obtained in several empirical simulations of different networks. In fact, among some tests, characterized by different momentum, learning rates and number of units in the hidden layers, the following configuration resulted in the best performance relative to the expected output.



**Figure 2.** BPN structure and configuration (7-10-1), with 7 input variables, 10 hidden nodes, and 1 output.

### 3. Results and Discussion

#### 3.1. Training Evaluation

The statistical analysis of the BPN training was carried out considering the *in situ* data as reference for the evaluation of the error on the predicted Chl *a* values, neglecting the error that can affect also the target data.

Figure 3a shows the plot density of the BPN training results. Even though the data are scattered, most of them are deployed close to the bisector, revealing a coherent agreement between the observed and modelled Chl *a* values. The formulas for the computation of statistics are given in Table 1. We considered those statistical parameters used in similar works to be comparable with them.

**Table 1.** Basic statistical quantities used for the assessment of the BPN performance with respect to the observed Chla values. N is the number of observations, and x is the observed value.

Determination Coefficient	$r^2 = \left( \frac{\sum_i (\text{Model}_i - \overline{\text{Model}})(x_i - \bar{x})}{\sqrt{\sum_i (\text{Model}_i - \overline{\text{Model}})^2} \sqrt{\sum_i (x_i - \bar{x})^2}} \right)^2$
Root Mean Squared Error	$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\text{Model}_i - x_i)^2}$
Bias	$\text{MBE} = \frac{1}{N} \sum_{i=1}^N (\text{Model}_i - x_i)$
Mean Absolute Percentage Difference	$\text{MAPD} = \frac{1}{N} \sum_{i=1}^N \left  \frac{\text{Model}_i - x_i}{x_i} \right  \times 100$
Median Absolute Percentage Difference	$\text{MdAPD} = \text{median} \left  \frac{\text{Model}_i - x_i}{x_i} \right  \times 100$
Mean Percentage Bias Error	$\text{MBE}\% = \frac{1}{N} \sum_{i=1}^N \left( \frac{\text{Model}_i - x_i}{x_i} \right) \times 100$

The high determination coefficient ( $r^2 = 0.71$ ) and the low bias ( $-0.02 \text{ mg m}^{-3}$ ) suggest the good predicting performance of the BPN, with a tendency to underestimate the observed Chla concentrations (Table 2).

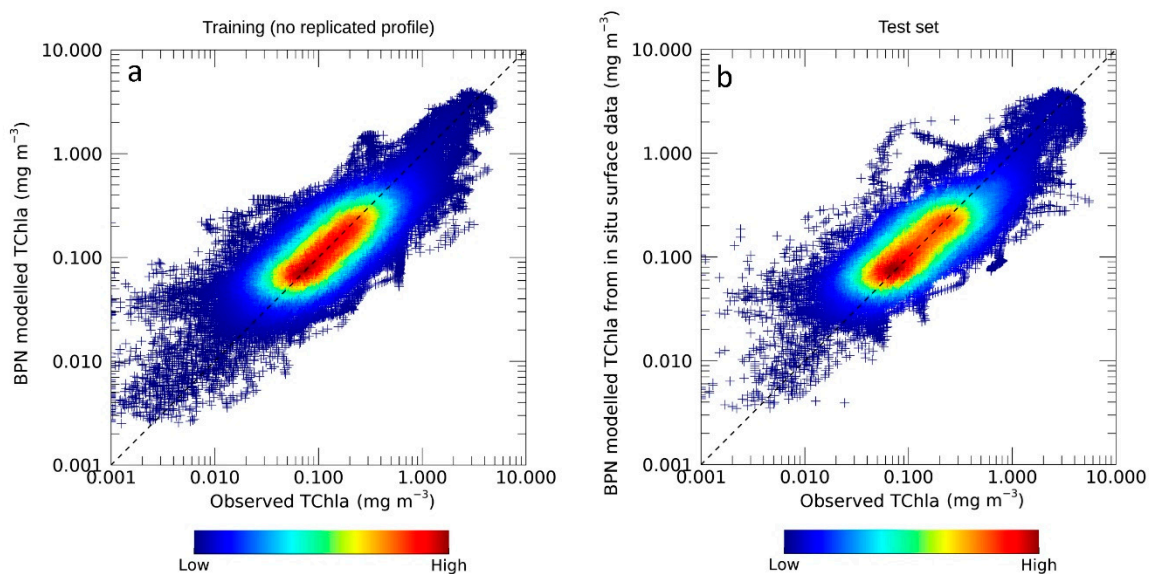
From the statistical analysis, an RMSE equal to  $0.17 \text{ mg m}^{-3}$  resulted, revealing a small error in the fit of predicted values vs. observed values. The results indicate that the BPN prediction capability is robust. The MBE% and the MdAPD are, respectively, 22% and 28%. Both statistical parameters fall within acceptable values, especially taking into account the variability that Chla has in the Mediterranean basin and that, here, its modelling is made from surface variables only.

The statistical results and the test executed during the training phase confirmed the goodness of the BPN performance.

**Table 2.** Statistical results of the comparison between the observed Chla values vs. the predicted Chla concentrations, as resulted, respectively, from the training phase, the validation on test set, and against BPN values inferred from satellite remote data. The statistics refer to linear values and not to log-transformed values.  $r^2$  refers to the determination coefficient, RMSE to the Root-Mean-Square Error, MAPD denotes the Mean Absolute Percentage difference, MdAPD the Median Absolute Percentage Difference, and MBE% represents the Mean Percentage Bias Error.

	Training Set (n = 125,652)	Test Set (n = 53,872)	Satellite Matchup Set (n = 163,244)
$r^2$	0.71	0.69	0.63
RMSE	0.17	0.26	0.23
bias	-0.02	-0.04	-0.04
MAPD	48.68	56.46	48.33
MdAPD	28.46	33.58	30.75
MBE%	21.59	27.71	16.98





**Figure 3.** Training phase (left panel) results and network's performance evaluation on test set (right panel). Density plot of the observed Chla concentration (on x axis) vs. the BPN estimated Chla values (y axis). The Chla concentrations are reported in log-scale. The statistics for log-transformed values are  $r^2 = 0.72$ , RMSE = 0.23, and bias = 0.01 for the training and  $r^2 = 0.72$ , RMSE = 0.23, and bias = 0.02 for the test set, respectively. In black, the 1:1 bisector.

### 3.2. Evaluation of the BPN Performance on the Test Set

After the training phase, the BPN was tested on an unknown dataset. The test set represents 30% of the total dataset (364 stations accounting for 53,872 set of values), including *in situ* Chla profiles and surface temperature values.

For the assessment of the BPN performance on the test set, the trained network was running in feed-forward direction (see supplementary material for the feed-forward run code of the network and the weight file of the present BPN). The inputs were treated and scaled as done for the training, and the output was scaled back to the original unit.

The results of this evaluation are shown in Figure 3b, where the good agreement between predicted and target values is highlighted in the plot density. Most of the data are deployed along the 1:1 line, especially for mean Chla concentrations (from 0.03 to 0.3  $\text{mg m}^{-3}$ ). For extreme domains, as highest and lowest concentrations, the scatter increases, revealing a general overestimation for small Chla values (0.001–0.03  $\text{mg m}^{-3}$ ; Figure 3b) and an underestimation for high concentrations.

As occurs in such models, the error in the predicted variable can be partially influenced by the quality of target data themselves. Usually, low Chla concentrations are most affected by high noise/signal ratio. Nevertheless, taking into account that, here, the predicted Chla values are inferred only from surface measures, the results on the test set are very reasonable.

As provided for the training phase, the statistical indices computed for the BPN assessment of the test set are given in Table 2. These results ( $r^2 = 0.69$ ; bias =  $-0.04 \text{ mg m}^{-3}$ ; RMSE =  $0.26 \text{ mg m}^{-3}$ ) are comparable to those of the training, highlighting the feasibility of applying a BPN to infer the vertical Chla field from surface estimates. The MBE% (28%) and MdAPD (34%) were slightly higher with respect to those of the training, but with the same order of magnitude. The comparable statistics between training and test set validation underline the generalization capability of such model acquired during the training phase.

In order to better visualize the BPN modelling capability, the Chla profiles (observed and predicted) were grouped according to different surface concentrations (Figure 4). Based on the trophic categories defined in Uitz et al. [17] and successively in Sauzède et al. [35], nine classes have been

chosen. All profiles falling within each class were averaged to obtain the mean profile for both cases: the observed and the BPN modelled one.

In Figure 4, the black and red lines represent, respectively, the mean profile of all *in situ* observations and those simulated by the network from *in situ* surface values, belonging to the same *in situ* surface Chla class (shown on the top of each panel), while the shaded area identifies their relative standard deviation.

Figure 4 shows that, generally, the vertical modelled pattern is very close to the real field. For classes of lowest Chla concentrations ( $0.001\text{--}0.04\text{ mg m}^{-3}$ ;  $0.04\text{--}0.08\text{ mg m}^{-3}$ ; Figure 4a,b), the estimated profile reproduces the shape of the observed one with high accuracy. In most cases, the two profiles overlap, especially in the depth range closer to the bottom. Here, in fact, the vertical Chla variability decreases, making its modelling easier. This is mostly evident for the class  $0.04\text{--}0.08\text{ mg m}^{-3}$  (Figure 4b) and, generally, occurs in all the classes, after the inflection point of the vertical Chla shape.

As already observed in the plot of Figure 3b, in low surface-concentration domain ( $0.001\text{--}0.04\text{ mg m}^{-3}$  class, Figure 4a), the network tends to overestimate the real profile. This behavior is probably related to the quality of target data. In fact, when very low Chla concentrations are considered, the noise-to-signal ratio is usually quite high, determining poorer predictability of the network in this ranges.

On the contrary, for the remaining Chla classes ( $0.08\text{--}2.2\text{ mg m}^{-3}$ , Figure 4c–h), the network predictions are underestimated with respect to the observed values. This occurs mainly for the classes from  $0.04\text{--}0.08\text{ mg m}^{-3}$  to  $0.4\text{--}0.8\text{ mg m}^{-3}$  (Figure 4b–h), while for classes of higher Chla concentrations ( $2.2\text{--}4.0\text{ mg m}^{-3}$ ; Figure 4i), the prediction is less accurate, especially in the first layers of water column, in which the BPN overestimates the observed Chla values. This is probably due to the high variability that Chla assumes in case of high concentrations, which means that high values in surface of Chla can reflect very dissimilar scenarios (e.g., bloom or coastal profiles), and thus create a poor prediction, especially in the upper layers, which are most affected by environmental processes of different nature. In the comparison of the two profiles, it is worth noting that their relative shaded areas are mostly overlapped, revealing that the variability of predicted values is comparable to that of the observed ones. This suggests that, even though with some uncertainties, such networks may be used to infer the shape of the Chla vertical field from surface data only, as a response to a particular combination of inputs. This further strengthens the ability of these approaches to integrate in their response the variability of the Chla field.

From Figure 4, the capability of the network to fit the surface Chla value emerges, especially for classes of lower concentrations, in which, most of the time, the two values coincide. Toward ranges of higher surface Chla, the network does not fit the *in situ* surface concentration as well (Figure 4h,i). This may be due to the number of profiles typical of eutrophic conditions within the training set. Even though the rare profiles were replicated, the random selection of the patterns in the learning phase could have affected the BPN reconstruction ability for those specific vertical Chla shapes. Moreover, the number of epochs, the type of activation function, and the learning constant may have an influence on the overfitting and underfitting of real data.

This validation allowed us to observe how the BPN is able to generalize the great variability of the trophic status that characterizes the Mediterranean Sea at specific spatio-temporal scales, which is, in turn, attested by the large grey shaded area of Figure 5. The results of this analysis provide the evidence of using BPN as a valid alternative to resolve a complex scenario such as the reconstruction of the 3D algal biomass pattern, which is usually affected by modelling issues.

### 3.3. Evaluation of the BPN Chla Profiles Inferred from Satellite Measurements

The Chla concentrations inferred by the BPN from surface satellite data were compared with field observations, in correspondence with the matchup stations, which accounted for 163244 total patterns.

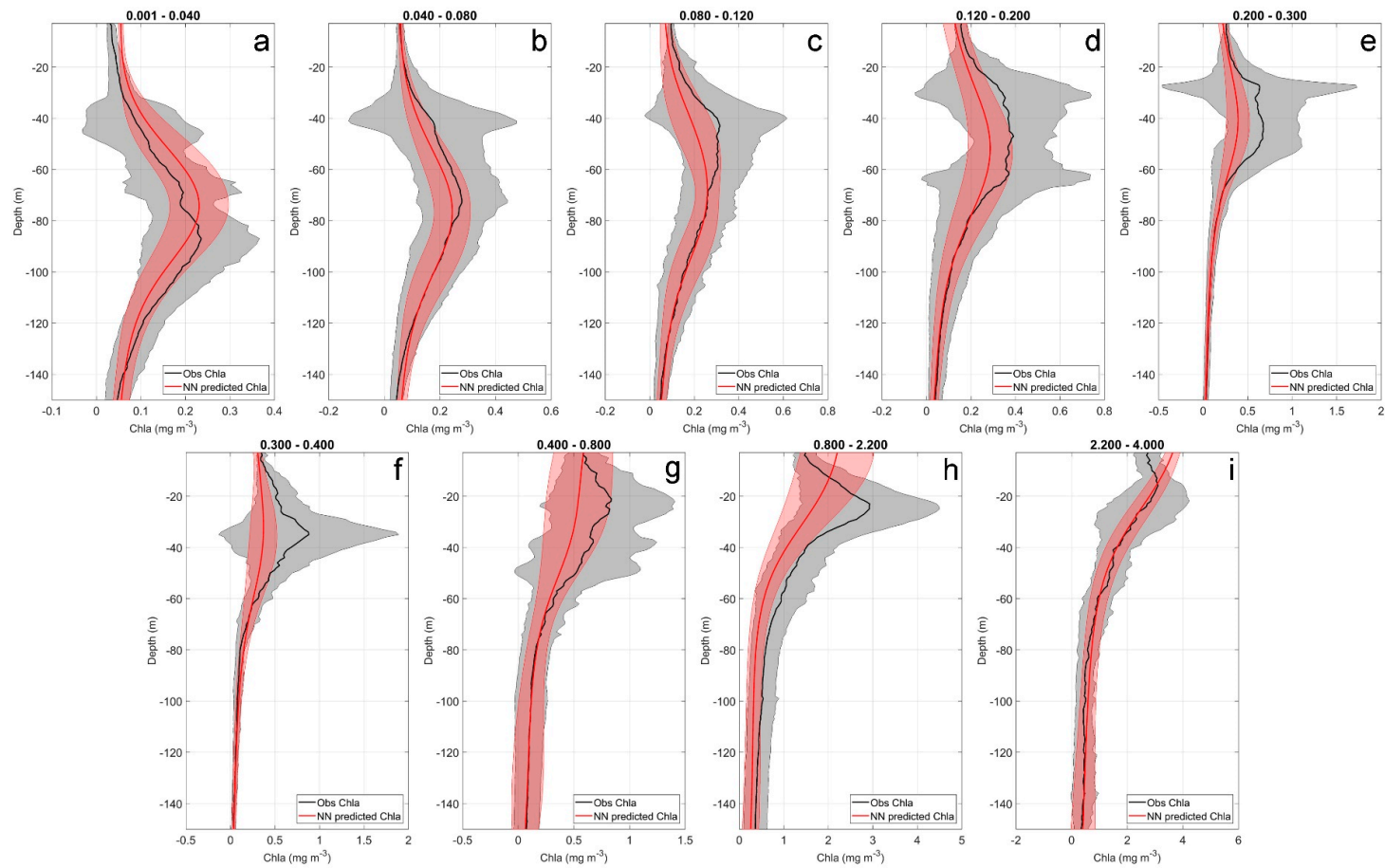
To estimate the BPN prediction capability, the same statistical parameters that were used for the overall evaluation of the BPN performance were considered (see satellite matchup set in Table 2).

The determination coefficient and the MdADP ( $r^2 = 0.63$ ; MdADP = 31%) were very close to those obtained on test set ( $r^2 = 0.69$ ; MdADP = 28%). Analogous results were obtained in previous and similar works as, e.g., that of Sauzède [20], in which the vertical profile of Chla is inferred by the combined use of Argo and satellite data, obtaining an  $r^2$  and MdADP of 0.67 and 44%, respectively. The comparison of our statistical results with those of Sauzède highlights that, even with some uncertainties and some limits, our network, applied on surface measurements only, shows a prediction accuracy very close to other models, in which both surface and vertical components are required.

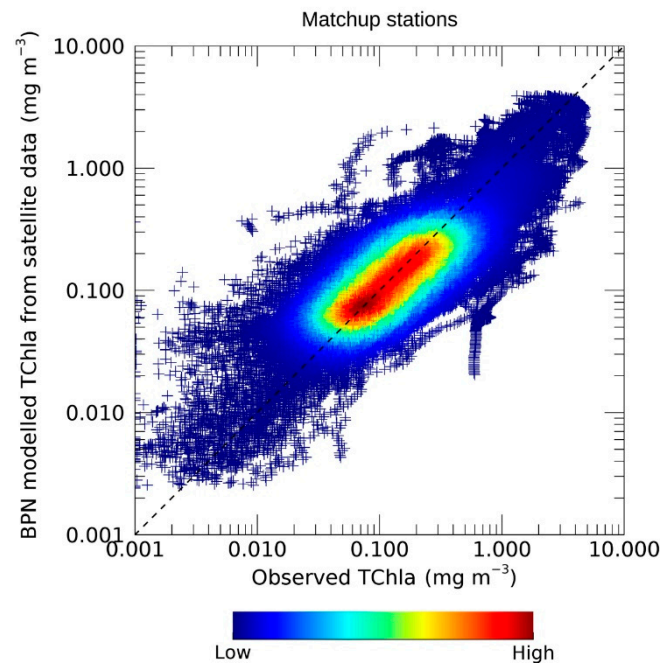
Here, the  $r^2$  was lower than the previous tests, revealing a possible influence of the satellite retrieval uncertainties on the prediction, independently from the reference dataset. In fact, same results are given, even if we consider separately the matchup stations belonging to the training set ( $r^2 = 0.61$ ) or the test set ( $r^2 = 0.66$ ).

The others statistical parameters, MBE% = 17% and the bias =  $-0.04 \text{ mg m}^{-3}$ , were of the same order of magnitude of those obtained on training and test set (Table 2).

As done in previous section, a plot density of remotely-sensed estimated values against observed ones is given in Figure 5. The plot shows the outputs of the BPN applied on surface remote-sensing data against the Chla targets. A high density of data points is concentrated around the bisector, especially in the range from 0.04 to  $1.0 \text{ mg m}^{-3}$ . On contrary, the scatter increases in the lower and higher domain of Chla concentrations, in which the data density is reduced, revealing, respectively, an overestimation and underestimation of the BPN predicted values.



**Figure 4.** Comparison of the mean profiles of observed values (black line) and BPN predictions (red line) inferred from *in situ* surface data and grouped according to different surface Chla concentration classes (in bold on top of each panel; a–i). For each panel (a–i) the black line represents the mean profile computed from the average of all observed profiles falling within a specific surface Chla class and the red line is the mean profile of all correspondent BPN-predicted profiles. The shaded area identifies the standard deviation of the two mean profiles, respectively, the original (grey area) and network-estimated ones (red area). The classes for which there were no profiles to be averaged are not represented. These profiles are related to the test set.



**Figure 5.** Comparison of the observed Chla concentrations vs. BPN modelled Chla inferred from satellite data in the matchup stations. Density plot of the observed Chla concentration (on x axis) vs. the BPN estimated Chla values (y axis). The Chla concentrations are reported in log-scale. For log-transformed values, the statistics are  $r^2 = 0.69$ ,  $RMSE = 0.24$ , and  $bias = -0.01$ . In black, the 1:1 bisector.

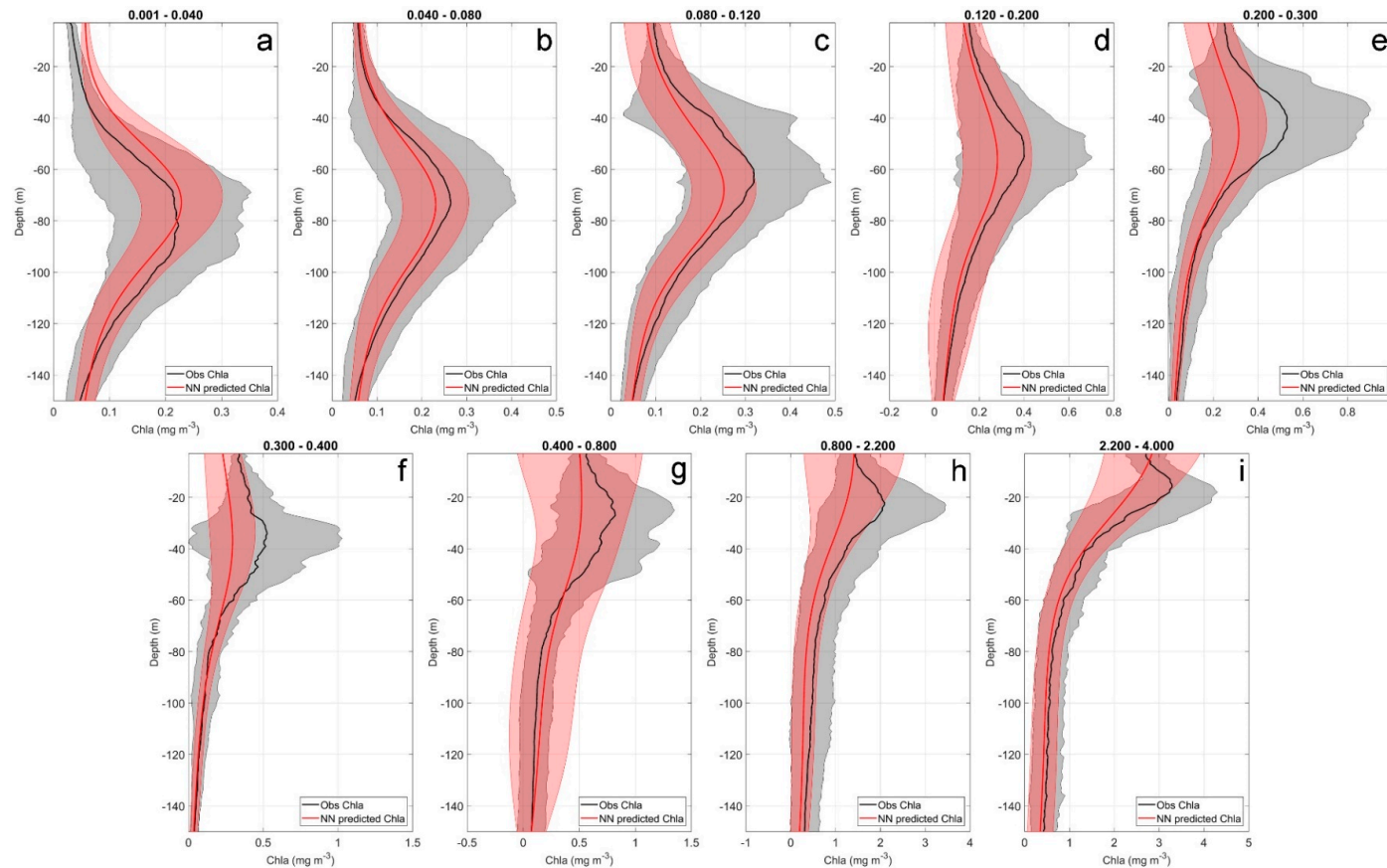
In Figure 6, the comparison of the mean BPN predicted profile versus the real one is shown. As done for Section 3.1, nine surface Chla classes were chosen, and all the profiles falling within each of them were averaged.

From Figure 6, the good agreement between the BPN modelled profile (red line) and the observed one (black line) is evident. In general, the two lines are overlapped or follow the same shape. One of the most interesting features is that the BPN is particularly efficient at recognizing the subsurface Chla maximum (Deep Chlorophyll Maximum-DCM) in most of the classes. This is of fundamental importance, if we take into account the mutable pattern that Chla can have in our basin, here, again, attested by the grey shaded area relative to the standard deviation of *in situ* profiles (Figure 6).

In Figure 6, it is shown that, generally, the BPN reproduces the vertically observed Chla profile with confidence.

The classes of lowest Chla ( $0.001\text{--}0.04\text{ mg m}^{-3}$ ,  $0.04\text{--}0.08\text{ mg m}^{-3}$ ; Figure 6a,b) show a BPN predicted mean profile very close to the real one, demonstrating the great prediction accuracy of the network in these Chla concentration domains. In the first class (Figure 6a), a slight overestimation of the BPN output occurs in the first 60–80 m of depth. Moving toward higher surface Chla intervals (from  $0.08$  to  $4.0\text{ mg m}^{-3}$ ; Figure 6c–i), the estimated profile is slightly underestimated in most of the classes, especially close to the DCM (20–65 m of depth). Here, in fact, the strong Chla variability, also attested by the larger grey shaded area, can affect the network accuracy in the prediction. However, below the DCM, the two lines (red and black) coincide in most of the classes, especially for the intermediate Chla ones as, e.g.,  $0.12\text{--}0.2\text{ mg m}^{-3}$ ;  $0.2\text{--}0.3\text{ mg m}^{-3}$ ; and  $0.3\text{--}0.4\text{ mg m}^{-3}$  (Figure 6d–f). In Figure 6, it is highlighted the high capability of the BPN to predict surface values very close to the observed ones, for all classes, except the  $0.2\text{--}0.3\text{ mg m}^{-3}$  and  $0.3\text{--}0.4\text{ mg m}^{-3}$  intervals (Figure 6e,f). These results emphasize the applicability of a neural approach to extend the remotely-sensed information to deeper layers.





**Figure 6.** Comparison of the mean profiles of observed values (black line) and BPN predictions (red line) inferred from satellite data and grouped according to different surface Chl *a* concentration classes (in bold on top of each panel; a–i). For each panel (a–i) the black line represents the mean profile computed from the average of all observed profiles falling within a specific surface Chl *a* class and the red line is the mean profile of all correspondent BPN-predicted profiles. The shaded area refers to the standard deviation of the two mean profiles, respectively: the original (grey area) and network-estimated ones (red area). The classes for which there were no profiles to be averaged are not represented. These profiles are related to the assessment of the BPN performance to infer vertical profiles from satellite data.

### 3.4. An Example of the 3D Chla Field Reconstruction from Satellite Data: the North-Western Mediterranean Sea Transect

Even though the Mediterranean Sea is an oligotrophic basin characterized by a mean Chla concentration less than  $0.2 \text{ mg m}^{-3}$ , significant deviations from this mean state occur in several regions throughout the year. Among these, the North-western Mediterranean Sea (NwMed) is one of the most investigated areas due to its high spatial and temporal variability characterized by extreme bloom events during late winter/early spring [36]. During these phenomena, the depth of the DCM varies seasonally according to the bloom timing, producing a variety of trophic regimes that can occur in the same area, making it more challenging to model the Chla vertical profile using only surface information.

In this section, rather than perform a further NwMed local validation against *in situ* data, we assess the capability of the BPN to reproduce the seasonal evolution of the DCM in the main bloom area.

Figure 7 shows the intra-annual variability of the vertical Chla field for the year 2009 as inferred by the BPN applied on satellite Chla and SST monthly mean, along the south to north NwMed transect (shown in Figure 8).

In the modelled field of Figure 7, the Chla reaches maxima values with different magnitudes and shapes in February, March, and April. In February, the Chla values seem to be overestimated with respect to the mean concentrations ( $0.3\text{--}0.4 \text{ mg m}^{-3}$ ) that characterize this bioregion in the late winter [37]. The overestimation can be justified by the absence, in the training set, of profiles for this month. This implies that the network learns from the closer month, e.g., March, which is usually characterized by bloom conditions.

Nevertheless, the high Chla values of February are not completely surprising. In fact, in this period, the water column is strongly mixed with the consequent disruption of the DCM and a Chla profile characterized by a sigmoid-like shape, typical of the mixed waters [38]. This process determines a lift of nutrients from bottom towards surface layers, with a consequent increase of the phytoplankton biomass. Moreover, as hypothesized by D'Ortenzio and Ribera d'Alcalà [36], in January and February, the Mixed Layer Depth (MLD) is deeper than the other months, causing a slight dispersal of the biomass within the water column, which can be represented by an anomaly of Chla concentrations (higher with respect to the seasonal means), within the water column.

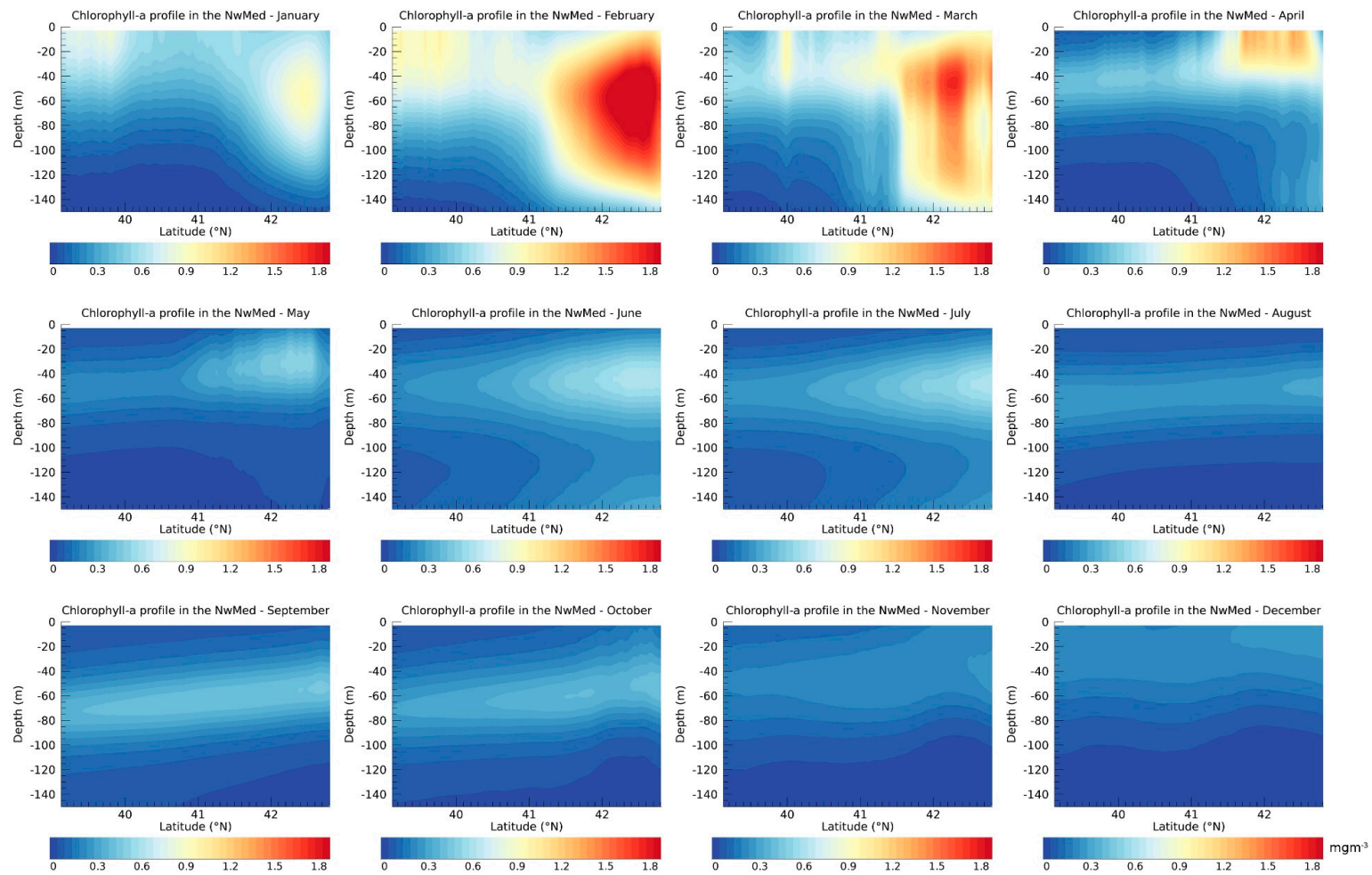
In March, the decrease in intensity of the vertical mixing and the increase of temperature causes an initial stabilization of the water column, which, in favorable conditions of light and nutrients, determines the onset of an evident phytoplankton bloom [11,36,39]. From March to April, the bloom rises up and remains constrained in surface, revealing a distinctive DCM within the first 40 m of depth. The high Chla values, typical of the bloom events and observed in our reconstruction, are also noticeable from satellite imageries (Figure 8).

At the surface, the lowest concentrations ( $<0.1 \text{ mg m}^{-3}$ ) occur in August, when the DCM (at 60 m of depth) becomes weak (Figure 7). In fact, in summer, the increase of the temperature induces a strong stratification of the water column, which determines very low concentrations ( $<0.1 \text{ mg m}^{-3}$ ).

In September/October, a slight increase of Chla values occurs (Figure 7). It is usually recognized as a second minor bloom, which is triggered by an intrusion of nutrients in surface layers due to the partial erosion of thermocline [36].

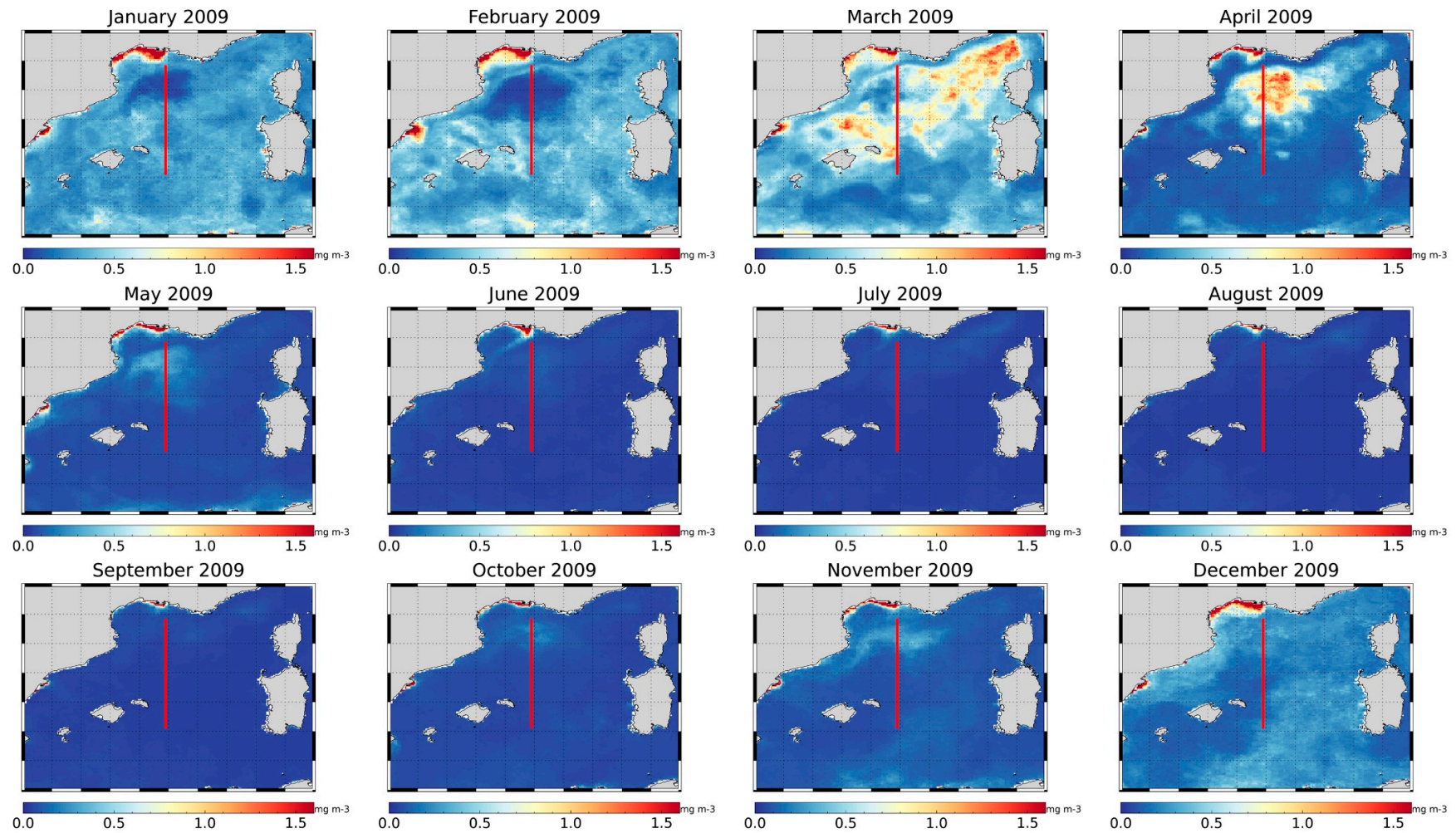
The results of this analysis highlighted the fact that the modelling of Chla can be very challenging, according to the deep variability that characterizes this bio-region of Mediterranean Sea.

Nevertheless, it should not be ignored that the well-known oligotrophy of the Mediterranean Sea determines, in surface, low Chla concentrations that are usually associated with several different shapes. This can increase the degree of freedom of Chla and makes the prediction of vertical profile from surface information only more difficult. However, the use of a neural network approach can straighten this issue, providing some solutions, for example, by adding as co-predictors variables that preserve information on the history evolution of the system toward specific trophic states (e.g., the surface Chla concentrations centered on some days earlier and later than the day examined; or MLD position as representative of the column mixing).



**Figure 7.** An example of the BPN application on a specified transect in the NwMed (Latitude: 39.12°–42.8° N, Longitude: 4.89° E). Contour plots of the vertical Chla field estimated by the BPN using monthly satellite data as inputs, for the year 2009.





**Figure 8.** Satellite monthly maps of Chla concentration for the year 2009. The red line refers to the transect of vertical reconstruction in Figure 7.

#### 4. Comparison of the BPN Reconstructed Chla Field and MEDATLAS Climatology

In order to assess the efficiency of our method, a qualitative comparison has been done between the BPN Chla reconstructed field and the MEDATLAS climatology.

As a global reference dataset, the MEDATLAS climatology represents one of the most commonly used instruments to approximate and define the variability of the Chla field along the water column. This climatology is based on the collection of multi-disciplinary, *in situ* hydrographic and bio-chemical data of the Mediterranean and the Black Seas, through widespread cooperation of countries (<http://modb.oce.ulg.ac.be/backup/medar/contribution.html>).

In Figure 9, we compare the observed Chla vertical section of a chosen transect with the Chla field inferred by the BPN once from *in situ* surface data and then from satellite estimates. Then, we compare these fields with the correspondent fall climatological pattern.

Even with some uncertainties, both reconstructions, from *in situ* (Figure 9b) and satellite surface data (Figure 9c), are coherent with the observed field (Figure 9a), recognizing the position of the DCM along the water column and its vertical variability. On contrary, the climatology pattern is less representative of the original profiles. It shows high Chla values on surface up to subsurface layers that are not recognized in the observed field. The prediction coming from *in situ* surface observations (Figure 9b) seems to be more confident than that inferred from satellite inputs (Figure 9c). We suggest that one reason for this discrepancy, among *in situ* and satellite predictions, may be due to the uncertainties that can affect Chla retrieval from satellite sensors. Nevertheless, both reconstructions do not differ significantly.

A more quantitative evaluation of the error in the BPN prediction with respect to the observed data is given by the dark line in plot of Figure 9e,f. As shown in the plots, the absolute error line is computed as the mean absolute difference between modelled and observed data for each quota.

For both modelled fields (predicted from *in situ* surface values and from satellite data; Figure 9e,f), the error is the same, ranging from  $\sim 0.003$  to  $0.08 \text{ mg m}^{-3}$ ; on contrary, the absolute error relative to climatology (Figure 9g) is higher and more variable with respect to the errors of BPN predictions. The highest discrepancy is evident at 80 m depth, where the climatology overestimates the original data, with an absolute error about of  $0.18 \text{ mg m}^{-3}$ .

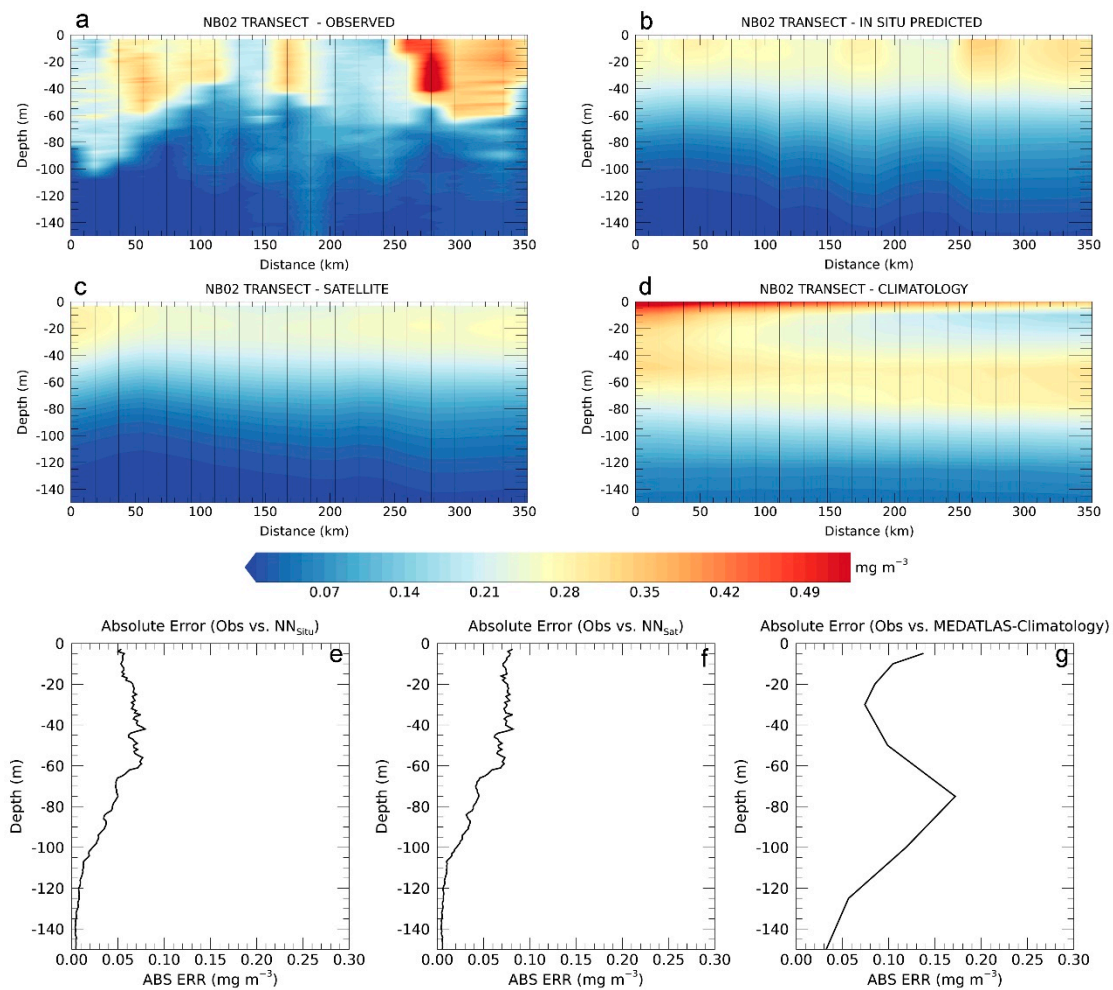
We hypothesize that, being characterized by a large number of coastal observations, the MEDATLAS climatology can be affected by the propagation of the high surface Chla concentrations through the interpolation, thus becoming a source for large dissimilarities with respect to the original profiles [11].

A second analysis was aimed at evaluating the general fit of the BPN and climatological estimates on the matchup database (Figure 10). For each station of this database (see Section 2.2), the correspondent Chla value was extracted from the BPN outputs and annual climatology, in order to compare them to the same number of data points. Figure 10 shows the scatter plots relative to this assessment: in blue, the results of the comparison of observed Chla against the BPN values predicted from *in situ* surface data; in red the same for values predicted from satellite observations; and finally, in green, the comparison against the annual MEDATLAS climatology.

As shown in Figure 10, the correlation between the observed and modelled data is high, for both predictions (from *in situ* surface and satellite inputs, Figure 10a,b). This is also confirmed by the statistical results given in Table 3, in which the  $r^2$  are 0.77 and 0.67 for Chla concentrations inferred from *in situ* and satellite measures. On the contrary, the comparison with the climatology shows low co-variability, with  $r^2 = 0.20$ . In Figure 10a, the colored dotted lines give a quantitative estimate of the MBE and absolute error associated with this comparison. On surface, the climatology (Figure 10a, green dotted line) shows higher discrepancy from the origin than the *in situ* (blue) and satellite (red) predictions. This is particularly evident for those depths at which a DCM occurs (between 0–60 m), while for deeper layers, the MBE decreases up to zero both for our method and climatology. As well as the MBE, the highest values of absolute error are registered by the climatology, constantly along the water column (Figure 10b). These results are also supported by the statistical analysis given in Table 3,



with an MBE% equal to 49% for the climatology comparison, which is higher than the 16% and 19% registered for satellite and *in situ* validations, respectively (see Table 3 for the statistics).



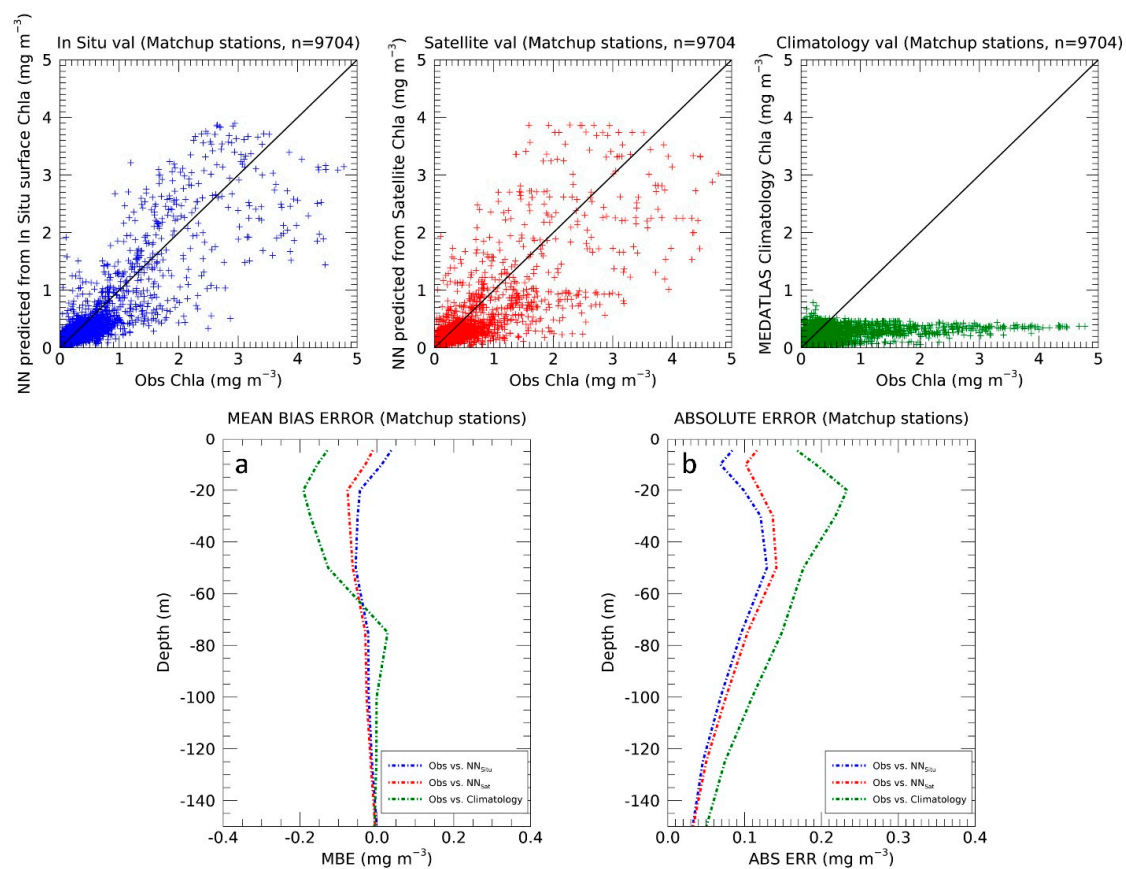
**Figure 9.** Contour plot of the chosen transect extracted from NB02 oceanographic cruise. The data are sampled in the Gulf of Lion from 18–12–2001 to 20–12–2001. Panels show the vertical section of (a) observed Chla field, (b) reconstructed field from *in situ* surface Chla, (c) reconstructed field from satellite data, and (d) extracted from the fall MEDTALAS climatology. The black lines in (a), (b), (c), (d) identify the station points. On the bottom, the three plots show, for each depth, the mean absolute error of computed differences between the observed and modelled data as (e) predicted from *in situ* surface Chla, (f) from satellite observations, and (g) extracted from fall MEDTALAS climatology.

Even if the consistency and the usefulness of the MEDATLAS climatology is nowadays clearly established, our comparison demonstrated that the use of novel approaches, as the method here proposed, represents important advantages to carry out an analysis of vertical Chla variability at finer temporal resolution, which are unsolvable with a climatological approach.

In fact, on a small scale, the Chla can strongly vary in different forms that cannot be represented by mean field as provided by the climatology. Even if our approach needs to be improved, it represents a first attempt at creating a valid alternative to reproducing Chla variability on seasonal, daily, and smaller spatio-temporal scales.

**Table 3.** Statistical results of the comparison between the observed Chla values vs. the modelled Chla concentrations predicted by the BPN (from surface *in situ* and satellite observations), and between the observed Chla values vs. MEDATLAS annual climatology. The analysis is carried out on the matchup database stations (see Section 2.2 for the technical details).

	In Situ Validation Set MATCHUP TOT (n = 9704)	Satellite Validation Set MATCHUP TOT (n = 9704)	Climatology MATCHUP TOT (n = 9704)
$r^2$	0.77	0.67	0.20
RMSE	0.21	0.25	0.41
bias	−0.02	−0.04	−0.08
MAPD	44.82	47.38	86.00
MdAPD	28.27	31.41	48.12
MBE%	19.38	16.31	49.38



**Figure 10.** Upper panel shows the scatterplots relative to the comparison on matchup database (for details see Section 2.2) of observed vs. BPN modelled data and climatology. In blue, the comparison of observed values (x axis) against predictions from superficial *in situ* data (y axis); in red, comparison against the Chla values predicted from satellite data (y axis); and finally, in green, the comparison against annual climatology (y axis). N is the number of available data. Bottom panel shows (a) the MBE and (b) the mean absolute error computed for each depth, for the three cases.

## 5. Conclusions and Future Perspectives

Several approaches have been advanced to characterize the phytoplankton assemblage structure.

As a well-known proxy for algal biomass [5,6], the Chla has a fundamental role in marine ecosystem monitoring research.

Nowadays, the employment of satellite sensors provides estimates of Chla at high spatial and temporal resolution, and it has become an efficient tool with which to follow the surface evolution of this pigment, as well as the phytoplankton.

Through the water column, however, the analysis of algal biomass dynamics is still hindered by practical difficulties. In fact, most of the techniques used to analyze Chla vertical variability are based on punctual or discrete samples. This heterogeneity between surface and vertical components represents an obstacle in the analysis of algal biomass in the 3D view.

In this work, a neural network approach was adopted to better investigate the existing relationship between the bio-physical surface parameters and vertical Chla shape. Taking into account the extrapolation ability of Machine Learning techniques [40], we combined the use of Artificial Neural Network with satellite variables (Chla and SST) to create an operative tool to assess the Chla vertical profile from its surface pattern. Specifically, a neural network trained with an error back-propagation algorithm (BPN) was employed to infer the Chla shape in the Mediterranean Sea from only surface variables.

Our method was validated against *in situ* data showing very promising results. We demonstrated the potential applicability of a BPN applied to satellite data to extend from 2D to 3D the algal biomass analysis.

Previous works also confirmed the feasibility of our approach. In fact, our results provide statistics comparable to those obtained in other works, e.g., Sauzède [20], in which, unlike our method, both information about surface and the water column are required to infer the vertical Chla field.

In this work, an example of the application of our model on an NwMed transect was shown. Satellite Chla and SST estimates were used to infer the vertical Chla field on seasonal scales. This analysis highlighted the ability of the BPN to recognize the typical seasonal cycle of the algal biomass in this area, which is usually very challenging to model.

The comparison of the BPN reconstruction with MEDATLAS climatology, at basin and local scales, represented an indirect validation of our method. This comparison demonstrated that, if the temporal scale of the analysis is small (e.g., day), our method may represent an improved approach to resolving Chla predictability at fine temporal scales than the use of the climatology.

Moreover, even if we know that an absolute optimization of a neural network cannot exist, its application to remote sensing data represents a novel instrument with which to extend from surface to deeper layers the spatial and temporal domain of the phytoplankton distribution analyses. This is of fundamental importance, taking into account the different trophic states of the Mediterranean Sea and the fragmented nature of the *in situ* datasets, which are nowadays used to describe the vertical Chla pattern.

In this work, it was observed how an artificial neural network could solve some of those issues coming from the complexity of the vertical algal biomass distribution characterizing the Mediterranean Sea. The neural network is potentially able to generalize the vertical shape of the Chla, only knowing its surface pattern and other variables, which are easily and constantly provided by satellite sensors.

Even if with some uncertainties, the employment of an Artificial Neural Network, in conjunction to the remote sensing products, represents an innovative opportunity to overcome the lack of information and divergence existing between surface and deeper Chla fields.

In this work, the use of surface information (Chla and SST) to infer the vertical algal biomass profile represents a first step to build up a tool able to incorporate into the output the implicit effects of variable bio-physical processes. In fact, here we considered those co-predictors that have definitely an influence on the output prediction; however, the structure of the network can be made more complex by including further and significant inputs. In this context, a future improvement of our method will be the employment of additional co-predictors, e.g., PAR and irradiance, or oceanographic parameters (e.g., mixing depth, wind components) that affect the Chla vertical shape, in addition to some statistical parameters (e.g., the standard deviation of Chla values in  $3 \times 3$  pixel matchup box), which together can improve the profile prediction. In fact, as known in machine learning techniques, one of the main

advantage of such artificial network is the chance to add new variables as predictors without knowing *a priori* the specific relationship between the new variables and the estimated output.

Nevertheless, since the good dimension and quality of the *in situ* database represents one of the principal requirements for a good training of the network, a future improvement will be the inclusion of new available data to expand the training set in view of a more accurate prediction also in those areas less represented in our dataset, where the prediction deserves more attention. This will be also accompanied by longer satellite time series, e.g., ESA-CCI v3 (1998-on going) at 1 km of resolution, released today by CMEMS service.

As already done for Chla, the neural network method may be used to infer the profile of other environmental variables; an example is temperature, which is usually strongly correlated with the trophic state of algal biomass.

This work demonstrated the usefulness of this model as an alternative to conventional models for understanding the 3D space of the relationships between phytoplankton biomass and physical phenomena that occur in the oceans.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2072-4292/10/10/1666/s1>. ANN feed-forward.pro: code to run the BPN in feedforward direction; weights.txt: BPN weight file; Appendix A1: instructions to use the code. Table S1: Arbitrary maxima and minima used to scale the BPN inputs in [0, 1] interval.

**Author Contributions:** All authors conceived and design the study. M.S. performed the processing of the dataset and images, the training and test of the network, and drafted most of the manuscript under the supervision and inputs from M.Sc., R.S., and S.M. All authors contributed to the result analysis and final discussion. All authors approved the submitted manuscript.

**Funding:** This research was supported by the Copernicus Marine Environmental Services—Ocean Colour Thematic Assembling Center Project (Grant number 77-CMEMS-TAC-OC-N). The research was also supported by the Italian Flagship Project RITMARE (la Ricerca Italiana per il MARE).

**Acknowledgments:** The authors would like to thank the research group of the GOS of the Italian CNR and the research group of the Department of Biology, in the Laboratory of Experimental Ecology and Aquaculture, University of “Tor Vergata”, Rome. We are very grateful to Simone Colella for his great technical support and his advices.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Guinder, V.; Molinero, J. Climate Change Effects on Marine Phytoplankton. In *Marine Ecology in a Changing World*; CRC Press: Boca Raton, FL, USA, 2013; pp. 68–90.
2. Sathyendranath, S.; Aiken, J.; Alvain, S.; Barlow, R.; Bouman, H.; Bracher, A.; Brewin, R.; Bricaud, A.; Brown, C.W.; Ciotti, A.M.; et al. *Phytoplankton Functional Types from Space. (Reports of the International Ocean-Colour Coordinating Group (IOCCG); 15)*; International Ocean-Colour Coordinating Group: Dartmouth, NS, Canada, 2014; pp. 1–156.
3. Reynolds, C.S. *The Ecology of Phytoplankton*; Cambridge University Press: Cambridge, UK, 2006.
4. Kyewalyanga, M. Phytoplankton Primary Production. In *UNEP-Nairobi Convention and WIOMSA The Regional State of the Coast Report: Western Indian Ocean*; UNEP and WIOMSA: Nairobi, Kenya, 2015; p. 546. Available online: <http://www.indiaenvironmentportal.org.in/files/file/WIO%20Regional%20State%20of%20Coast%20Report.pdf> (accessed on 22 March 2016).
5. Huot, Y.; Babin, M.; Bruyant, F.; Grob, C.; Twardowski, M.; Claustre, H. Does chlorophyll a provide the best index of phytoplankton biomass for primary producers? *Biosci. Discuss.* **2007**, *4*, 707–745. [[CrossRef](#)]
6. Volpe, G.; Nardelli, B.B.; Cipollini, P.; Santoleri, R.; Robinson, I.S. Seasonal to interannual phytoplankton response to physical processes in the Mediterranean Sea from satellite observations. *Remote Sens. Environ.* **2012**, *117*, 223–235. [[CrossRef](#)]
7. Colella, S.; Falcini, F.; Rinaldi, E.; Sammartino, M.; Santoleri, R. Mediterranean ocean colour chlorophyll trends. *PLoS ONE* **2016**, *11*, 1–16. [[CrossRef](#)] [[PubMed](#)]



8. Sathyendranath, S.; Brewin, R.J.W.; Jackson, T.; Mélin, F.; Platt, T. Ocean-colour products for climate-change studies: What are their ideal characteristics? *Remote Sens. Environ.* **2017**, *203*, 125–138. [[CrossRef](#)]
9. Boyer, J.N.; Kelble, C.R.; Ortner, P.B.; Rudnick, D.T. Phytoplankton bloom status: Chlorophyll a biomass as an indicator of water quality condition in the southern estuaries of Florida, USA. *Ecol. Indic.* **2009**, *9*, 56–67. [[CrossRef](#)]
10. Klein, T.; Nilsson, M.; Persson, A.; Håkansson, B. From Open Data to Open Analyses—New Opportunities for Environmental Applications? *Environments* **2017**, *4*, 32. [[CrossRef](#)]
11. Lavigne, H.; D’Ortenzio, F.; Ribera D’Alcalà, M.; Claustre, H.; Sauzède, R.; Gacic, M. On the vertical distribution of the chlorophyll a concentration in the Mediterranean Sea: A basin-scale and seasonal approach. *Biogeosciences* **2015**, *12*, 5021–5039. [[CrossRef](#)]
12. Morel, A.; Berthon, J.F. Surface Pigments, Algal Biomass Profile, and Potential Production of the Eutrophis Layer: Relationships Reinvestigated in View of Remote Applications. *Limnol. Oceanogr.* **1989**, *34*, 1545–1562. [[CrossRef](#)]
13. Sauzède, R.; Lavigne, H.; Claustre, H.; Uitz, J.; Schmechtig, C.; D’Ortenzio, F.; Guinet, C.; Pesant, S. Vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A first database for the global ocean. *Earth Syst. Sci. Data* **2015**, *7*, 261–273. [[CrossRef](#)]
14. Buongiorno Nardelli, B.; Cavalieri, O.; Rio, M.-H.; Santoleri, R. Subsurface geostrophic velocities inference from altimeter data: Application to the Sicily Channel (Mediterranean Sea). *J. Geophys. Res.* **2006**, *111*, 1–22. [[CrossRef](#)]
15. Gueye, M.B.; Niang, A.; Arnault, S.; Thiria, S.; Crépon, M. Neural approach to inverting complex system: Application to ocean salinity profile estimation from surface parameters. *Comput. Geosci.* **2014**, *72*, 201–209. [[CrossRef](#)]
16. Sauzède, R.; Claustre, H.; Uitz, J.; Jamet, C.; Dall’Olmo, G.; D’Ortenzio, F.; Gentili, B.; Poteau, A.; 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. *J. Geophys. Res. Oceans* **2016**, *121*, 2552–2571. [[CrossRef](#)]
17. Uitz, J.; Claustre, H.; Morel, A.; Hooker, S.B. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *J. Geophys. Res. Oceans* **2006**. [[CrossRef](#)]
18. Charantonis, A.A.; Brajard, J.; Moulin, C.; Bardan, F.; Thiria, S. Inverse method for the retrieval of ocean vertical profiles using self organizing maps and hidden markov models: Application on ocean colour satellite image inversion. In Proceedings of the International Conference on Neural Computation Theory and Applications, Paris, France, 24–26 October 2011; pp. 316–321.
19. Cortivo, F.D.; Chalhoub, E.S.; Velho, H.F.C.; Kampel, M. Chlorophyll profile estimation in ocean waters by a set of artificial neural networks. *Comput. Assist. Methods Eng. Sci.* **2015**, *22*, 63–88.
20. Sauzède, R. Etude et paramétrisation de la distribution verticale de la biomasse phytoplanctonique dans l’océan global. Océan, Atmosphère. Ph.D. Thesis, Université Pierre et Marie Curie, Paris, France, 2015.
21. Richardson, A.J.; Pfaff, M.C.; Field, J.G.; Silulwane, N.F.; Shillington, F.A. Identifying characteristic chlorophyll a profiles in the coastal domain using an artificial neural network. *J. Plankton Res.* **2002**, *24*, 1289–1303. [[CrossRef](#)]
22. Scardi, M. Artificial neural networks as empirical models of phytoplankton production. *Mar. Ecol. Prog. Ser.* **1996**, *139*, 289–299. [[CrossRef](#)]
23. Scardi, M. Neural Network Applications in Coastal Ecological Modeling. *Elsevier Oceanogr. Ser.* **2003**, *67*, 505–532.
24. Mattei, F.; Franceschini, S.; Scardi, M. A depth-resolved artificial neural network model of marine phytoplankton primary production. *Ecol. Model.* **2018**, *382*, 51–62. [[CrossRef](#)]
25. Lek, S.; Giraudel, J.L.; Guégan, J.-F. Neuronal networks: Algorithms and architectures for ecologists and evolutionary ecologists. In *Artificial Neuronal Networks*; Springer: Berlin, Germany, 2000; pp. 3–27.
26. Scardi, M.; Harding, L.W. Developing an empirical model of phytoplankton primary production: A neural network case study. *Ecol. Model.* **1999**, *120*, 13–223. [[CrossRef](#)]
27. Scardi, M. Advances in neural network modeling of phytoplankton primary production. *Ecol. Model.* **2001**, *146*, 33–45. [[CrossRef](#)]
28. Lek, S.; Scardi, M.; Verdonchot, P.; Descy, J.-P.; Park, Y.-S. Modelling Community Structure in Freshwater Ecosystems. In *Modelling Community Structure in Freshwater Ecosystems*; Springer: Berlin, Germany, 2005.



29. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [[CrossRef](#)]
30. Volpe, G.; Santoleri, R.; Vellucci, V.; Ribera d'Alcalà, M.; Marullo, S.; D'Ortenzio, F. The colour of the Mediterranean Sea: Global versus regional bio-optical algorithms evaluation and implication for satellite chlorophyll estimates. *Remote Sens. Environ.* **2007**, *107*, 625–638. [[CrossRef](#)]
31. Thessen, A. Adoption of Machine Learning Techniques in Ecology and Earth Science. *One Ecosyst.* **2016**, *1*, e8621. [[CrossRef](#)]
32. Werbos, P.; Paul, J. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*; Harvard University: Cambridge, MA, USA, 1974.
33. Sammartino, M. Modelling the Vertical Distribution of Phytoplankton Biomass in the Mediterranean Sea. Ph.D. Thesis, University of "Tor Vergata" Rome, Rome, Italy, 2016.
34. Györgyi, G. Inference of a rule by a neural network with thermal noise. *Phys. Rev. Lett.* **1990**, *64*, 2957–2960. [[CrossRef](#)] [[PubMed](#)]
35. Sauzède, R.; Claustre, H.; Jamet, C.; Uitz, J.; Ras, J.; Mignot, A.; D'Ortenzio, F. Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications. *J. Geophys. Res. Ocean.* **2015**, *120*, 451–470. [[CrossRef](#)]
36. D'Ortenzio, F.; D'Alcalà, M.R. On the trophic regimes of the Mediterranean Sea: A satellite analysis. *Biogeosciences* **2009**, *6*, 139–148. [[CrossRef](#)]
37. Mayot, N.; D'Ortenzio, F.; Uitz, J.; Gentili, B.; Ras, J.; Vellucci, V.; Golbol, M.; Antoine, D.; Claustre, H. Influence of the Phytoplankton Community Structure on the Spring and Annual Primary Production in the Northwestern Mediterranean Sea. *J. Geophys. Res. Ocean.* **2017**, *122*, 9918–9936. [[CrossRef](#)]
38. Mignot, A.; Claustre, H.; Uitz, J.; Poteau, A.; Ortenzio, F.D.; Xing, X. Global Biogeochemical Cycles. *Glob. Biogeochem. Cycles* **2014**, *32*, 856–876. [[CrossRef](#)]
39. Siokou-Frangou, I.; Christaki, U.; Mazzocchi, M.G.; Montresor, M.; Ribera d'Alcala, M.; Vaque, D.; Zingone, A. Plankton in the open mediterranean Sea: A review. *Biogeosciences* **2010**, *7*, 1543–1586. [[CrossRef](#)]
40. Olyae, E.; Banejad, H.; Chau, K.W.; Melesse, A.M. A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: A case study in United States. *Environ. Monit. Assess.* **2015**, *187*, 189. [[CrossRef](#)] [[PubMed](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).