# Accurate deep-learning estimation of chlorophyll-a concentration from the spectral particulate beam-attenuation coefficient

SEBASTIAN GRABAN,[1,2,4] GIORGIO DALL'OLMO,[1,2,5] STEPHEN GOULT,[1,2] AND RAPHAËLLE SAUZÈDE[3]

[1]*Plymouth Marine Laboratory, Prospect Place, Plymouth, PL1 3DH, United Kingdom*
[2]*National Centre for Earth Observation, Plymouth, United Kingdom*
[3]*Sorbonne Université, CNRS INSU, Institut de la mer de Villefranche, Villefranche-sur-Mer, France*
[4]*sebastian.graban@live.co.uk*
[5]*gdal@pml.ac.uk*

**Abstract:** Different techniques exist for determining chlorophyll-a concentration as a proxy of phytoplankton abundance. In this study, a novel method based on the spectral particulate beam-attenuation coefficient ($c_p$) was developed to estimate chlorophyll-a concentrations in oceanic waters. A multi-layer perceptron deep neural network was trained to exploit the spectral features present in $c_p$ around the chlorophyll-a absorption peak in the red spectral region. Results show that the model was successful at accurately retrieving chlorophyll-a concentrations using $c_p$ in three red spectral bands, irrespective of time or location and over a wide range of chlorophyll-a concentrations.

## 1. Introduction

The chlorophyll-a pigment is common to all microscopic algae and its concentration (*Chl-a*) in natural waters is commonly used as a proxy for phytoplankton abundance. Different techniques have been developed over the years to quantify *Chl-a*.

*Chl-a* can be measured in the laboratory on discrete water samples that are collected in the field. The phytoplankton cells from these samples are concentrated on filters, their pigments extracted using a solvent, and *Chl-a* quantified using techniques such as *in-vitro* fluorometry, spectrophotomery or high-performance liquid chromatography (HPLC) [1]. Of these laboratory methods, HPLC is currently considered as the gold standard for determining *Chl-a* (as well as other accessory pigments) on discrete samples and can reach a maximum accuracy of ~10% and a reproducibility of ~20% [2]. Despite its excellent performance, HPLC has some drawbacks (e.g., high costs both for collecting and analysing samples as well as increased uncertainty from handling samples in laboratories as compared to in-situ automatic methods), which limit the achievable accuracy and the number of samples that can be generated.

To overcome these limitations, field techniques have been developed to estimate *Chl-a in vivo* (i.e., on living phytoplankton cells) using sensors. The most commonly employed sensor to estimate *Chl-a in situ* is the fluorometer that exploits the red fluorescence emitted by chlorophyll-a when cells are illuminated with blue light (e.g., [3]). Fluorescence is approximately proportional to *Chl-a* and *in-situ* fluorometers are compact, low-power, and relatively sensitive sensors, which makes them ideal to be installed on autonomous platforms such as gliders and Biogeochemical-Argo floats [4]. Nevertheless, the proportionality between fluorescence and *Chl-a* varies with phytoplankton species (up to 7 times [5]) and physiological status (up to 6 times [6]), rendering *in-vivo* fluorescence-based estimates of *Chl-a* rather inaccurate.
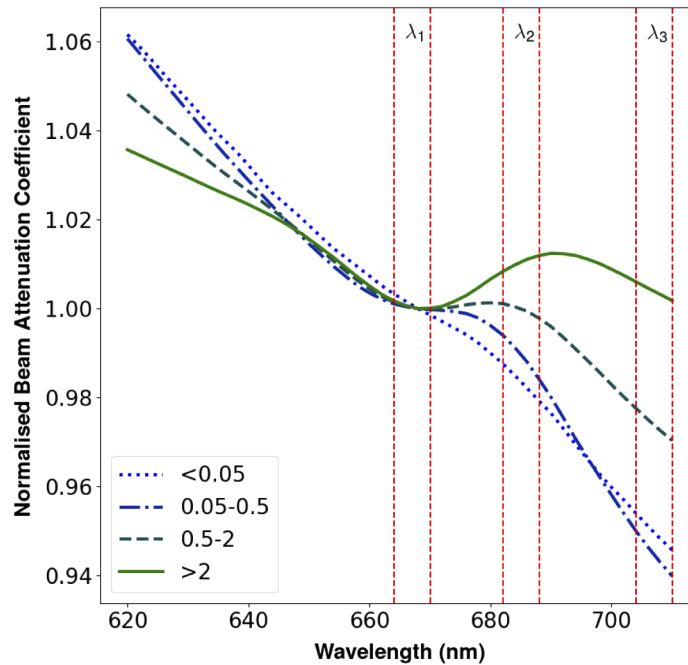
**Fig. 1.** Average normalised (at 660 nm) spectra of particulate beam-attenuation coefficient for different ranges of chlorophyll-a concentration (indicated in the legend by different line types in mg/m$^3$). $\lambda_1$, $\lambda_2$ and $\lambda_3$ represent the ranges of wavelengths used to train the model.

More recently a new *in-vivo* technique, based on measuring the particulate absorption peak of chlorophyll-a in the red spectral region (around 676 nm), has been developed using an absorption and attenuation meter (WETLabs ACS or AC9, [7,8]) and demonstrated to estimate *Chl-a* with high accuracy regardless of the phytoplankton species and physiological status [9–14]. Thus, estimates of *Chl-a* based on particulate absorption can overcome the problems of *in-vivo* fluorometers, but still suffer from their own limitations. For example, *in-situ* measurements of particulate absorption can only be conducted inside a closed path (e.g., [7]), which in turn requires a pump to draw sample water in it, thus adding complexity to the measurement system and preventing its installation on autonomous platforms.

Another optical property of the particles suspended in natural waters is the particulate beam attenuation coefficient ($c_p$), which is the sum of the particulate absorption and scattering coefficients, the latter contributing the majority of $c_p$ in open-ocean waters [15]. $c_p$ has the advantage of being considerably easier to measure than absorption, as for example it does not require a closed path (e.g., [15]) and has also be installed on autonomous platforms.

Interestingly, peaks in the particulate absorption coefficient correspond to a spectral signature in $c_p$ (Fig. 1) (e.g., [16]). This spectral feature in $c_p$ is caused by the so called "anomalous dispersion" [17]. Briefly, the optical properties of particles (including phytoplankton cells) are controlled by their complex refractive index, $n = m + im'$. The real part $m$ of the refractive index affects the scattering coefficient and thus $c_p$, and typically displays a monotonic decrease as a function of wavelength also known as *normal dispersion* [18]. Conversely, the complex part of the refractive index, $m'$, determines its absorption coefficient and can display characteristic local maxima and minima [18]. In the vicinity of a peak of $m'$, $m$ displays an *anomalous dispersion*, whereby $m$ decreases (increases) at wavelengths shorter (longer) than the peak in $m'$ [17]. As a consequence of this anomalous dispersion of $m$, the beam-attenuation coefficient also displays

spectral features in correspondence of the red absorption peak of *Chl-a* [19,20], but the shape of the $c_p$ spectral features depends on the size parameter and therefore is, in general, different from that of $m$ [17,19].

Thus measurements of $c_p$ in the spectral region around 676 nm contain information that is directly related to *Chl-a*. However, this information is obscured by other variable factors including the concentration, composition and size distribution of all the other particles suspended in the water that generate a $c_p$ signal. As a consequence, only very recently has the estimation of *Chl-a* from $c_p$ been attempted [21].

Here, we hypothesised that, with a dataset large enough to be representative of a wide range of *Chl-a* values and optical conditions, a machine-learning algorithm could be trained to estimate *Chl-a* from $c_p$ spectra. Luckily, tens of thousands of particulate absorption and attenuation measurements have been collected in the surface ocean in underway mode [22] in many ocean regions [9,10,12–14,23,24]. Importantly, all these data are freely available to the scientific community in the NASA SeaBASS public database (seabass.gsfc.nasa.gov, [25]).

The main objective of this study was to answer the following question. Can we accurately estimate *Chl-a* from $c_p$ measurements using few bands in the red spectral region and over a wide range of *Chl-a* values? Below we will show that indeed deep-learning neural networks can be trained to accurately estimate *Chl-a* from $c_p$ measurements at three bands in the red spectral region. These results open the way to new applications of $c_p$ measurements in natural waters, including the potential to develop new instrumentation to estimate *Chl-a* in situ.

## 2. Methods

### 2.1. Data description

To train the deep neural network a large dataset of particulate beam attenuation coefficients ($c_p$) and absorption-based chlorophyll-a (chl-ACS) estimates was used. $c_p$ and particulate absorption data collected during numerous expeditions and processed as previously described [9,11] were extracted from the NASA SeaBASS archive (Table 1, [25]). Due to missing chl-HPLC or missing $c_p$ data, only the following legs of the Tara Oceans expedition were used: AbuDhabi-Masqat, Alger-Barcelone, Ba-Ushuaia, Beyruth-Alexandrie, Ct-Rio, Easterisland-Guayaquil, Honolulu-SanDiego, Madagascar-Mayotte, Male-StBrandon, Masqat-Mumbay, Mauritius-Reunion, Mayotte-CT, Moorea-Honolulu, Mumbay-Male, Panama-Savannah, Reunion-Madagascar, Rikitea-Papeete, Rio-Ba, Sandiego-Panama, Savannah-NewYork, Sharm-Jeddah, Tanger-Alger and Valparaiso-EasterIsland.

The line-height method applied to the chlorophyll-a absorption peak in the red spectral region was used to obtain a nominal estimate of *Chl-a* from the ACS data (chl-ACS):

$$\text{chl-ACS}_n = [a_p(676) - 39/65 a_p(650) - 26/65 a_p(714)]/0.014, \tag{1}$$

where 0.014 $m^2$ mg Chl-a is a nominal estimate of the chlorophyll-a specific absorption coefficient at 676 nm [8].

These chl-ACS$_n$ estimates were validated and de-biased by comparing them to concurrent discrete high-performance liquid chromatography estimates of total chlorophyll-a (chl-HPLC; estimated from the sum of monovinyl chl-a, divinyl chl-a and chlorophyllide a). Specifically, estimates of chl-ACS$_n$ were interpolated in time onto coincident chl-HPLC samples taken near the surface ($\leq 5$ m). Treating each expedition separately, we debiased the chl-ACS$_n$ estimates by first computing the relative residuals between each chl-ACS and chl-HPLC measurement:

$$\text{rel\_res} = \text{chl-ACS}_n/\text{chl-HPLC} - 1 \tag{2}$$

and then used their median, $\delta$, to remove the bias from the initial estimates of chl-ACS as follows:

$$\text{chl-ACS} = \text{chl-ACS}_n(1 - \delta). \tag{3}$$

**Table 1. Information on the datasets used to train and validate the neural network, as well as summary statistics of the chl-ACS datasets before debiasing: "N. HPLC" and "N. ACS" are the number of *Chl-a* values in the HPLC and ACS datasets, respectively; "Range" is the range of *Chl-a* in the ACS dataset; $\delta$ and $\sigma$ are the median and the robust standard deviation of the relative residuals between chl-ACS and chl-HPLC, respectively, before we subtracted the bias from each relationship (Eq. (3)). *a* and *b* are estimates of the parameters fitted to the equation chl-HPLC = $aL^b$, where $L = 0.014 \times$ chl-ACS$_n$ (see Eq. (1)). Numbers in brackets after *a* and *b* are their 95% confidence intervals.**

| Expedition | N. HPLC | N. ACS | Range (mg/m$^3$) | $\delta$ | $\sigma$ | $a$ | $b$ |
|---|---|---|---|---|---|---|---|
| Tara Oceans | 105 | 186342 | 0.01 – 7.19 | 0.14 | 0.69 | 180 (350) | 1.19 (0.34) |
| Tara Polar | 86 | 60173 | 0.04 – 8.88 | -0.045 | 0.40 | 50 (56) | 0.92 (0.21) |
| NAAMES1 | 39 | 13568 | 0.19 – 4.27 | 0.00 | 0.18 | 89 (175) | 1.05 (0.39) |
| NAAMES2 | 37 | 17836 | 0.14 – 9.28 | -0.16 | 0.18 | 84 (61) | 1.00 (0.17) |
| NAAMES3 | 41 | 15952 | 0.07 – 2.29 | -0.01 | 0.14 | 35 (53) | 0.88 (0.28) |
| NAAMES4 | 42 | 14743 | 0.21 – 4.26 | -0.10 | 0.14 | 49 (59) | 0.90 (0.26) |
| NAB08 | 20 | 839 | 0.43 – 4.64 | -0.06 | 0.10 | 22 (24) | 0.71 (0.26) |
| TAO2012 | 104 | 1222 | 0.04 – 0.42 | 0.052 | 0.12 | 92 (69) | 1.05 (0.12) |
| AMT19 | 16 | 814 | 0.03 – 0.72 | -0.14 | 0.073 | 77 (27) | 1.00 (0.05) |
| AMT22 | 215 | 32015 | 0.00 – 1.31 | 0.042 | 0.14 | 64 (16) | 1.00 (0.04) |
| AMT26 | 81 | 34518 | 0.01 – 2.25 | 0.022 | 0.073 | 68 (38) | 0.97 (0.09) |
| Tot. train | | 35787 | 0.00 – 9.28 | | | | |
| Tot. valid | | 128007 | 0.00 – 9.23 | | | | |

This de-biasing was performed because HPLC is currently considered as the "gold standard" for estimating *Chl-a*. Therefore, any bias found in the chl-ACS$_n$ estimates (when compared to chl-HPLC) was attributed to imperfections in the absorption line-height method.

Once de-biased, the chl-ACS and corresponding $c_p$ were median filtered (in the time dimension), with a window size of 30 minutes, to remove any high-frequency noise in the data. The median filtered data which were less than 3 times the robust standard deviation of the filtered data were kept, which removed 13% of the initial data points across all the cruises. Finally, negative chl-ACS values were removed.

To make our results comparable to previous studies [26,27], we also fitted for each expedition the power law chl-HPLC = $aL^b$, where $L = 0.014 \times$ chl-ACS$_n$. The resulting exponents were hardly ever significantly different from 1 (Table 1), indicating that the relationships between chl-HPLC and $L$ were linear. These results are in disagreement with previous findings reporting a non-linear relationship between chl-HPLC and particulate absorption in the red spectral region [26] and suggest that the ACS red-absorption data used in this study were not affected by the "package effect".

## 2.2. Data division

The $c_p$ and chl-ACS datasets described in section 2.1 were divided into three separate datasets: training, validation and independent validation. All the datasets in Table 1, excluding Tara Oceans, were divided into the training and validation datasets. The training dataset was used to train the model. The validation dataset was used to validate the model once it had been trained, however, as the data in the validation dataset came from the same expeditions as the training dataset, it was not considered truly independent. To independently validate our model, the Tara Oceans dataset was used, which consisted of data from an expedition in areas which were not included in the training dataset. The spatial and temporal distribution of the datasets can be seen in Figs. 2 and 3.
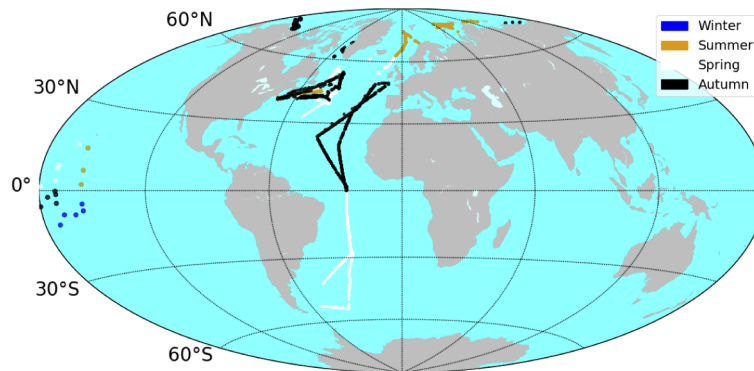
**Fig. 2.** Map of the spatial and temporal distribution of the training and validation datasets. Reported seasons are respective of each hemisphere.

The training and validation datasets were divided by applying a stratified sampling: the initial dataset, excluding the Tara Oceans expedition, was divided into 500 log-scaled bins of chl-ACS between 0.004 mg/m$^3$ and 9.28 mg/m$^3$. Then, based on the observed maximum number of samples available in the bin with the fewest samples, samples were randomly extracted from each bin to generate the training dataset. Note that the bins at the extremes of the *Chl-a* range had fewer than 100 points, because fewer measurements were available. Finally, to ensure that each bin was represented in the validation dataset, when the number of samples within a bin was less than 60, a single randomly selected sample was moved from the training dataset into the validation dataset.

## 2.3. *Deep neural network*

### 2.3.1. Data augmentation

Data augmentation is a key step needed to maximise the performance of deep neural networks and it aims at increasing the amount and diversity of the input data. Data augmentation can involve extracting additional features from the input data (e.g., gradients) or data normalisation.

Initially, neural networks were fitted to the data produced in sections 2.1 and 2.2. However, we found that data augmentation could improve the performance of the models.

The initial data were in the form of $c_p$ spectra (from 620 to 710 nm, interpolated every 2 nm) and the corresponding chl-ACS values. To minimise the training time, we extracted one $c_p$ value from three wavelength ranges defined as $\lambda_1$ (664 - 670 nm), $\lambda_2$ (682 - 688 nm) and $\lambda_3$ (704 - 710 nm) (shown in Fig. 1). These ranges were selected through experimentation, but also keeping in mind the main characteristics of spectral $c_p$ around the red absorption peak (Fig. 1). The width of the spectral ranges was limited by computational costs. A spectral width of 6 nm (4 different values for each specific range) was found to be suitably representative of the spectral features, while not requiring too many additional computational resources. The experimentation consisted of training models using several ranges around the spectral features and observing their bias and precision, until the wavelength ranges above were reached. Thus the initial training data set contained in each row the three $c_p$ values selected from the three wider $\lambda_1$, $\lambda_2$ and $\lambda_3$ ranges as well as their corresponding chl-ACS value.

This initial training dataset was then augmented as follows. The absolute value of $c_p$ varies due to different factors besides *Chl-a* [15], while the imprint of chlorophyll-a absorption on $c_p$ results in specific spectral features (Fig. 1). Therefore the gradients and ratios between the three
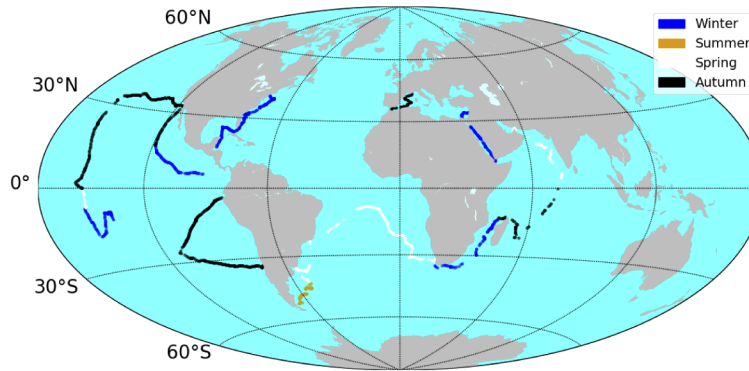
**Fig. 3.** Map of the spatial and temporal distribution of the independent validation dataset. Seasons are respective of hemisphere.

selected $c_p$ values were also computed and used as inputs. Gradients were calculated as follows:

$$\nabla = \frac{c_p(\lambda_i) - c_p(\lambda_j)}{\lambda_i - \lambda_j} \tag{4}$$

Using these gradients and ratios, a new input vector was formed for the neural network. The input vector consisted of 12 features: three $c_p$ values, and the three gradients and three ratios between the $c_p$ values, as well as the three wavelengths at which the $c_p$ values were measured.

To ensure equal weighting of all features when training the model: all 12 features were normalised using Z-score normalisation:

$$\text{Z-score} = \frac{x_i - \mu}{\sigma}, \tag{5}$$

where $x_i$ is the i$^{th}$ value of a given feature $x$, $\mu$ is the mean value of $x$ and $\sigma$ is its standard deviation.

Finally, we repeated the above procedure for all possible combinations of $\lambda_1$, $\lambda_2$ and $\lambda_3$ in the wavelength ranges mentioned above. Thus we were able to train the network on multiple wavelength combinations and assess which wavelength triplet resulted in the most accurate estimates of *Chl-a* (lowest bias and precision of the relative residuals).

### 2.3.2. Network architecture

The deep neural network architecture used was a Multi Layer Perceptron (MLP) which was shown to produce strong regression results and was previously successfully applied to chlorophyll retrievals [28–30]. A MLP is composed of an input layer, hidden layers and an output layer. The MLP used in this study was designed with two hidden layers each consisting of 2048 nodes (this number of nodes was reached through experimentation, Fig. 4). To reach this architecture, we tested, using the bias and precision of the relative residuals as optimising metrics, other architectures including an encoder or random forest regressor as well as different numbers of neurons (from 32 to over 2048 per layer).

In a neural network, the activation function of a node defines the output from the node, based on the input. Activation functions allow non-linear relations in the neural network, permitting the translation of complex data into output values. Without these functions neural networks would only be capable of linear mapping from inputs to outputs. The activation function employed in this study was the rectified linear activation function $\max(0, x)$, where $x$ is the input to the neuron [31].
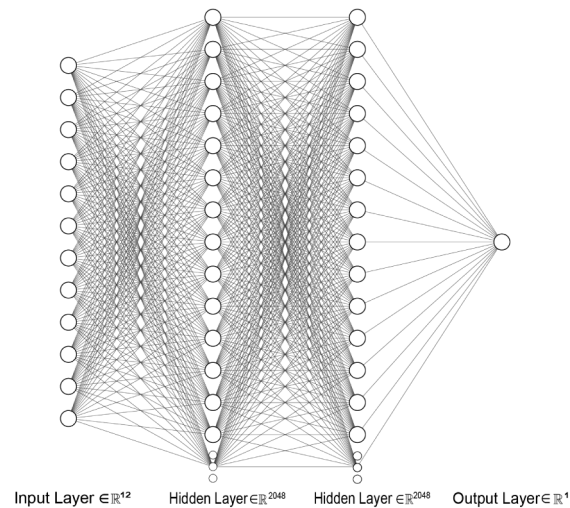
**Fig. 4.** Multi-Layer Perceptron architecture used in this study.

To improve the performance and resilience of the model to outliers, an ensemble of the nine best performing MLPs was employed to generate predictions. Each of these MLP models individually predicted *Chl-a* values (chl-CP), but here we report the median of the nine models. An additional important advantage of using this ensemble prediction is that it allowed us to provide estimates of the prediction uncertainty. This uncertainty was computed as the robust standard deviation of the nine predicted chl-CP values divided by the square root of nine, in order to provide a "robust" equivalent to the standard error of the mean. While each of the ensemble models has the same architecture (Fig. 4), their results vary due to the random sampling of the training dataset (section 2.2), as well as because the initial weights of each MLP were randomly assigned (see next section).

### 2.3.3. Network tuning

Each MLP was trained by means of the optimiser function called RMSprop [32], which was selected due to its speed of convergence [32]. The weights in a model are the coefficients that parameterise the functions linking inputs, nodes and output. Initially these weights are randomised, then RMSprop applies iterative changes to the weights to attempt to minimise the value of the loss function, in this case the Huber loss function [33]:

$$L_\alpha(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \leq \alpha \\ \alpha|y - f(x)| - \frac{1}{2}\alpha^2 & \text{otherwise}, \end{cases} \tag{6}$$

where $y$ is the true or observed value (chl-ACS), $f(x)$ is the predicted value (chl-CP) and $\alpha$ a tuning parameter. The Huber-loss function attempts to achieve a balance between the Mean Absolute Error (MAE) and Mean Squared Error (MSE) loss functions. The MAE is more robust to outliers, whereas the MSE produces more accurate models [34]. In the Huber loss function, $\alpha$ determines which part of the loss function is used based on the error, defined as $|y - f(x)|$ (see Eq. (6)). An $\alpha$ value of 1.5 was selected for this model, which meant that if the error in the model was less than 1.5 mg/m$^3$ then the MAE loss function was applied, and else the MSE was employed. This $\alpha$ was selected through experimentation and found as the best balance between robustness against outliers, that were especially important at lower chl-ACS concentrations, and accuracy.

The model was tuned with the training dataset for 200 epochs, where one epoch is defined as one forward and backward training iteration of the network. Early stopping with a "patience" of 70 was used, which stopped the tuning if an improvement on the current best loss function was not reached for 70 consecutive epochs.

## 3. Results

### 3.1. Data results

Figure 5 shows that there was a strong correlation between debiased chl-ACS and chl-HPLC. The samples were also well distributed across a wide range of chlorophyll-a values. This result confirmed that chl-ACS was suitable to be used as accurate *Chl-a* values when training and validating the deep neural network model.
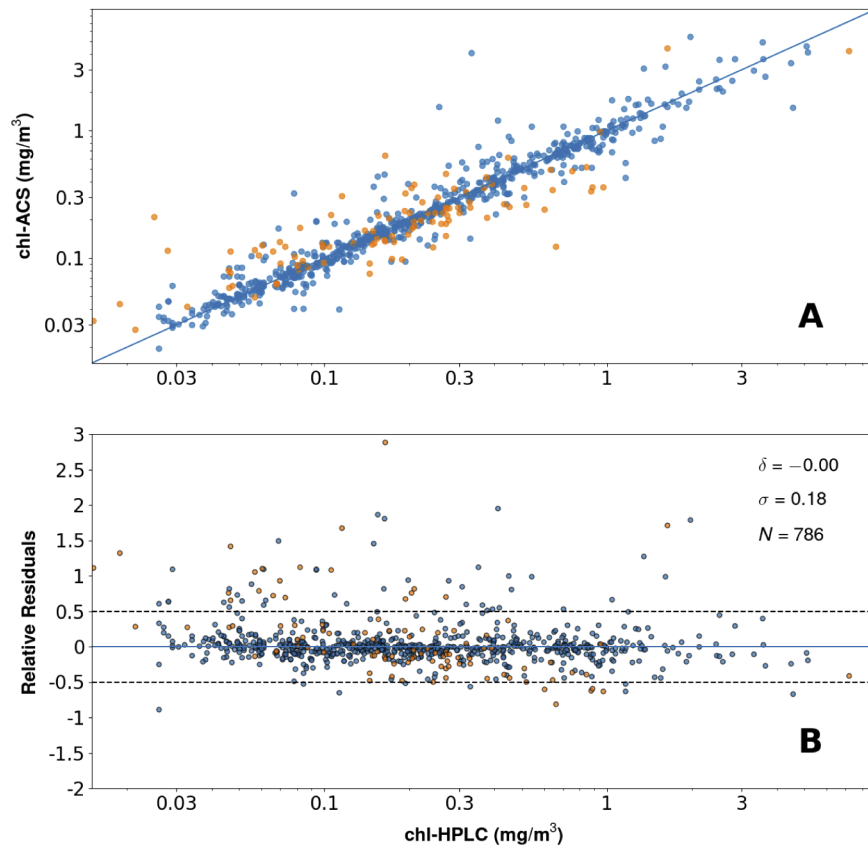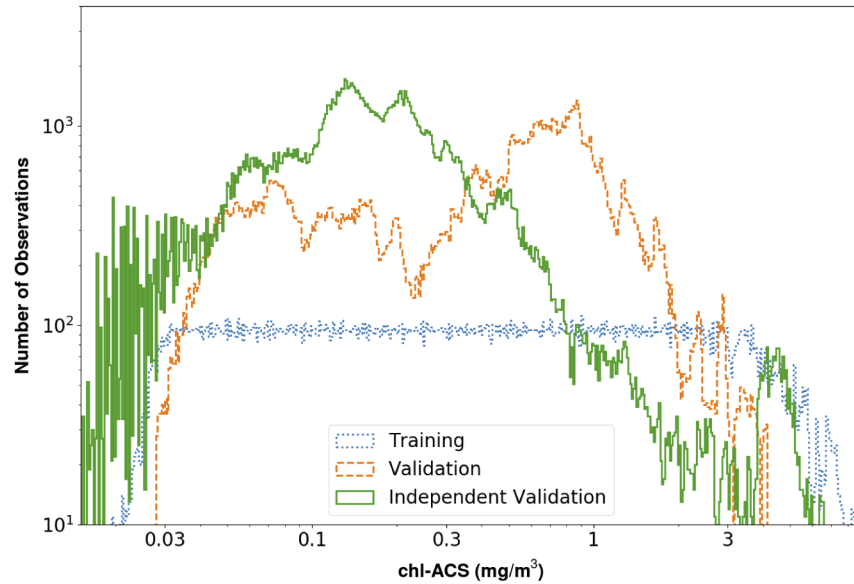


**Fig. 5.** (A) Scatter plot between debiased chl-ACS vs. chl-HPLC and (B) the corresponding relative residuals computed as chl-ACS/chl-HPLC−1. $\delta$ represents the median (i.e., the bias of the relationship) and $\sigma$ the robust standard deviation (i.e., the precision) of the relative residuals. Compiled using data from all the cruises in Table 1 with orange points representing the Tara Oceans expedition.

The distributions of chl-ACS for the three datasets (training, validation and independent validation) resulting from the stratified sampling are presented in Fig. 6. The training dataset was uniformly distributed between about 0.03 and 3 mg/m³ with fewer data outside these limits. The validation dataset contained a large number of samples with chlorophyll-a values less than 2

mg/m$^3$. The independent-validation dataset contained relatively few samples above 1 mg/m$^3$, but a significant number of them below 0.03 mg/m$^3$.



**Fig. 6.** Distributions of the training, validation and independent-validation datasets.
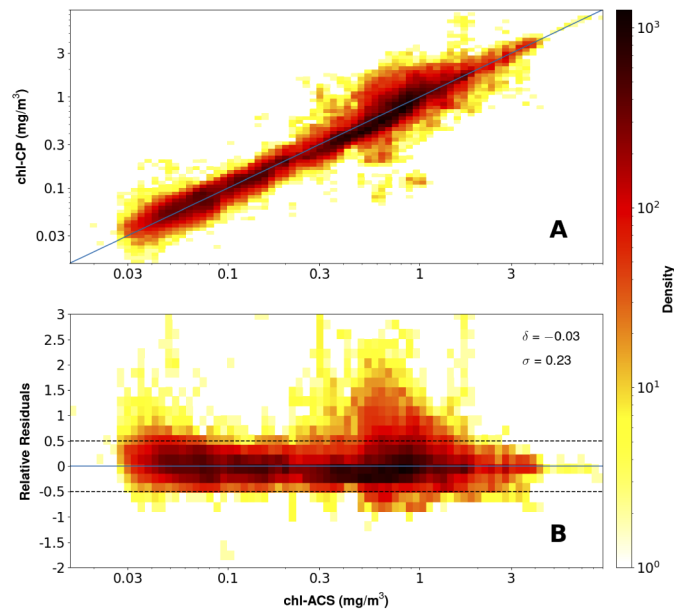


**Fig. 7.** Bi-dimensional histograms of (A) the relationship between predicted chl-CP and measured chl-ACS and (B) the corresponding relative residuals, computed as chl-ACS/chl-HPLC−1, for the validation dataset. $\delta$ represents the bias of the relationship (the median of the relative residuals) and $\sigma$ shows its precision (the robust standard deviation). Horizontal dashed lines mark the ±50% residuals.

The training and validation data were mostly collected from spring to autumn in the Atlantic and Arctic (Fig. 2), although few samples in these datasets were collected in the equatorial Pacific and during winter. The independent validation data were collected in all seasons and in oceanic regions, such as the Indian ocean and the Red Sea, for which no training data were available (Fig. 3).

## 3.2. Validation dataset results

The neural-network was successful at predicting *Chl-a* across a wide range of conditions. A strong relationship was found between chl-ACS in the validation dataset and the 9 neural-networks selected for the ensemble. Figure 7 shows the predicted chl-CP using $\lambda_1 = 670$ nm, $\lambda_2 = 688$ nm, $\lambda_3 = 710$ nm for one of the models selected for the ensemble. The prediction bias (-3%) and precision (23%) remained relatively consistent across the whole range of chl-ACS. The results of only one model rather than the ensemble are presented here as each of the 9 models in the ensemble had a different training-validation split as described in section 2.2. The results of the
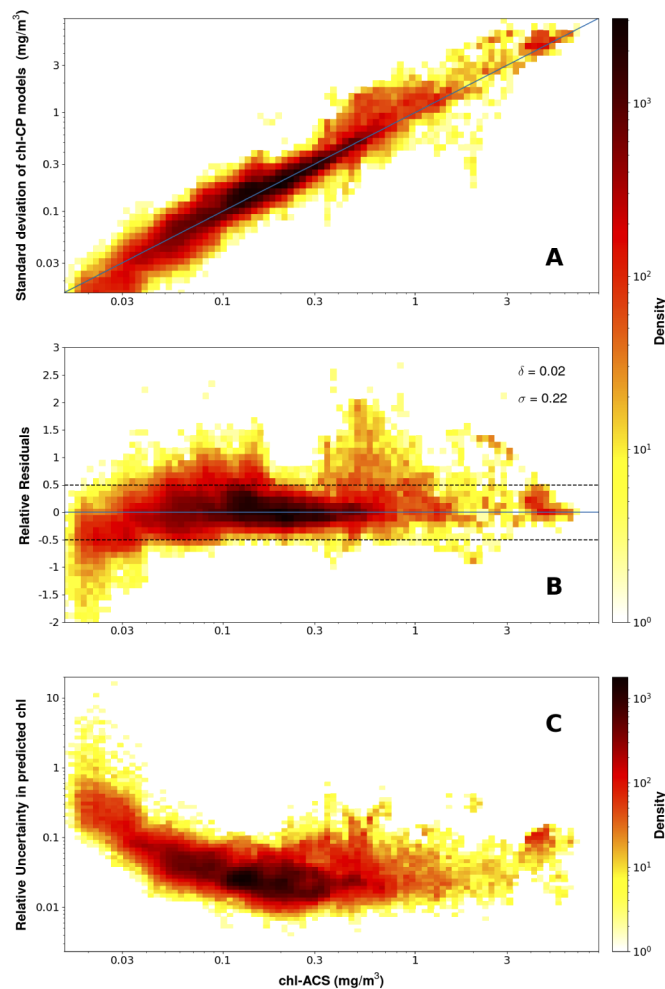


**Fig. 8.** (A) and (B) As for Fig. 7, but for the independent validation and using the median value from the 9 models forming the ensemble. (C) Relative prediction uncertainty estimated from the variability within the ensemble.

remaining 8 models which formed the ensemble closely resembled the results in Fig. 7, with a median bias of -7% and median precision of 23% across the 9 models.

### 3.3. Independent-validation dataset results

Ensemble predictions (using $c_p$ at $\lambda_1 = 670$ nm, $\lambda_2 = 688$ nm, $\lambda_3 = 710$ nm) also compared favourably to observations for the independent-validation dataset (with +2% bias and a precision of about 22%). This result suggests that the ensemble can predict *Chl-a* regardless of the season or region in which $c_p$ is sampled (Fig. 8). However, the accuracy of the model appeared to degrade at lower chl-ACS concentrations (<0.05 mg/m$^3$). On the other hand, the model was still able to accurately predict chl-CP at chlorophyll-a concentrations >3 mg/m$^3$.

The relative prediction uncertainty, computed from the precision of the ensemble models, was <20% from the highest chl-ACS concentration down to about 0.05 mg m$^{-3}$ (Fig. 8(C)), indicating that at these concentrations model predictions could be trusted. However, at chlorophyll levels below 0.05 mg m$^{-3}$ the uncertainty approached 100%, suggesting lower accuracy for these predictions, consistent with the validation data (Fig. 8(B)).

Figure 9 presents the relative bias and relative precision obtained for the independent-validation dataset when training the ensemble with different $\lambda_1$ and $\lambda_2$ combinations, while keeping $\lambda_3$ fixed at 710 nm. Even in the narrow ranges of wavelengths we selected, there were specific pairs of $\lambda_1$ and $\lambda_2$ that resulted in significantly more accurate predictions. Specifically combinations of longer $\lambda_1$ and $\lambda_2$ appeared to produce less biased results. On the other hand, Fig. 9(B) shows that, regardless of $\lambda_1$ and $\lambda_2$, the relative precision of the models remained consistently around 0.22.
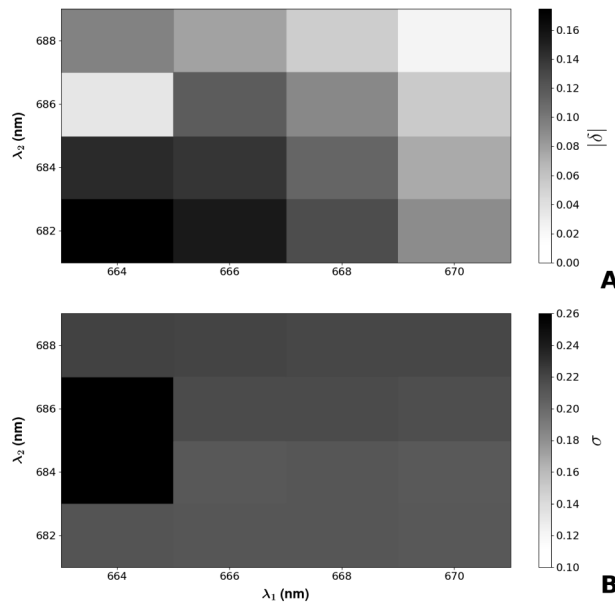


**Fig. 9.** Relative bias (in absolute value, A) and relative precision (B) of the relative residuals computed for the independent-validation dataset for multiple $\lambda_1$ and $\lambda_2$ combinations, while keeping $\lambda_3$ fixed at 710 nm.

## 4. Discussion

### 4.1. Use of a neural network

Devising a mechanistic model to predict chlorophyll-a concentration (*Chl-a*) from three spectral channels of the particulate beam attenuation ($c_p$) is challenging because, besides the absorption

by chlorophyll-a, multiple other factors simultaneously affect $c_p$. We attempted this task with a deep neural network and were successfully able to predict chl-CP with a minimal bias. We believe the success of this method is due to the ability of the network to disentangle from the $c_p$ spectra the information about the absorption by *Chl-a*. This ability most likely arises from the flexibility of the neural network and the very large training dataset that we assembled.

The model ensemble was based on randomising the initial weights of each model and on random sampling of the training dataset. This ensemble prediction allowed us to estimate the uncertainty of our predictions as well as to maximise the information contained in the training data, without biasing the model towards specific concentrations of chl-ACS (section 2.2).

### 4.2.   Regional and seasonal distribution of independent-validation dataset

The results from the independent-validation dataset in Fig. 8 show that the network also performed well in geographic regions it had not been exposed to during training (Fig. 3). Therefore, the network was successfully trained to avoid regional or seasonal biases. Nonetheless, the results presented for the independent-validation dataset (Fig. 8) were somewhat noisier, (i.e., with a larger bias at lower $Chl - a$ concentrations) than those for the validation dataset (Fig. 7). We hypothesise that an important reason for this discrepancy may be that the independent-validation dataset was collected during the Tara Oceans expedition. During this expedition, the HPLC samples were collected by personnel who had received minimal training in water sampling and filtration and sometimes samples were shipped in sub-optimal conditions to the laboratory that conducted the HPLC analyses (E. Boss, personal communication, 2020). In addition, 2 liters of water were sampled for HPLC analyses [27], but this water volume is probably too low to accurately determine *Chl-a* in oligotrophic waters. Furthermore, Tara Polar crossed many land sources which could have resulted in very different $c_p$ spectra from increased mineral concentrations. This line of reasoning could help in explaining the larger noise in Fig. 8 and the poorer statistics for Tara Oceans and Tara Polar in Table 1. Importantly, we do not mean to discredit the monumental efforts spent collecting and sharing the precious data on the Tara expeditions: we are simply trying to provide a plausible explanation for the larger noises observed for these datasets.

### 4.3.   Performance of model at lower chl-ACS concentrations

The model bias increased at lower chl-ACS concentrations (Figs. 7 and 8). The bias in Fig. 7 is interesting, because the network should have been more robust when validated against data from datasets that it was trained on. This bias could be due to the intrinsic noise of the ACS absorption measurements and to the lack of features in the spectral $c_p$ at very low chl-ACS concentrations (Fig. 1). As the chl-ACS concentration decreases the $c_p$ curve becomes smoother and loses the features that the neural network relies on to predict chl-CP. In addition, in clear waters the chl-ACS estimates approach their background noise, which is determined by the precision of the $a_p$ data [11]. The relatively higher bias in Fig. 8 at very low concentration of *Chl-a* is due to samples from the South Pacific subtropical gyre that presents the clearest waters on the planet and specific phytoplankton assemblages. This extremely low *Chl-a* values and ocean region were missing from the training database. This drawback could potentially be improved upon by re-training the models using additional independent measurements collected in these ultra-oligotrophic waters. The increase in prediction uncertainty at the same low *Chl-a* (Fig. 8(C)) is also most likely a result of the phenomena described above. As a result of these factors, the variability among model predictions increased at the lower chlorophyll concentrations causing the increase in uncertainty. This demonstrates that associating an uncertainty with every chl-CP prediction is important to decide the extent to which the predicted value can be trusted.

### 4.4.   Performance of model compared to absorption

A further test was performed to compare the performance of the attenuation-based chl-CP estimates against the absorption-based chl-ACS. For this test the chl-ACS was not de-biased. The chl-CP and chl-ACS were independently validated against HPLC measurements taken from the Tara Oceans expedition. chl-ACS achieved a bias of +14% and precision of 69%, whereas chl-CP achieved a bias of +1% and precision of 68%. Despite this difference in biases, due to the relatively large variability of the Tara Oceans dataset (Table 1 and Fig. 5(B)), we can only conclude that the two methods did not differ considerably, likely because the chl-CP method was trained using the chl-ACS data.

## 5.   Conclusion

A deep neural network was successfully trained to accurately predict *Chl-a* using spectral particulate beam-attenuation coefficients. Additional training data could be used to improve the accuracy of the model, especially at the extremes of the *Chl-a* range. Our study opens a new way to exploit measurements of the particulate beam-attenuation coefficient.

## Disclosures

The authors declare no conflicts of interest.

## Data Availability

The dataset containing all data on latitude, longitude, time, $c_p$, and chl-ACS can be found at: https://zenodo.org/record/3953725#.XxmOtZ5KiUk

The code and instructions required to estimate *Chl-a* from $c_p$ measurements using the ensemble of tuned neural networks can be found at: https://github.com/SebastianGraban/CMAP

## References

1. L. Van Heukelem and C. S. Thomas, "Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments," J. Chromatogr. A **910**(1), 31–49 (2001).

2.  S. B. Hooker, H. Claustre, J. Ras, L. Van Heukelem, J.-F. Berthon, C. Targa, D. van der Linde, R. Barlow, and H. Session, "The first SeaWiFS hplc analysis round-robin experiment (SeaHARRE-1)," *NASA Technical Memorandum - SeaWIFS Postlaunch Technical Report Series no.14* pp. 40–42 (2000).
3.  C. S. Yentsch and D. A. Phinney, "Spectral fluorescence: an ataxonomic tool for studying the structure of phytoplankton populations," J. Plankton Res. **7**(5), 617–632 (1985).
4.  H. Claustre, K. S. Johnson, and Y. Takeshita, "Observing the global ocean with biogeochemical-argo," Annu. Rev. Mar. Sci. **12**(1), 23–48 (2020).
5.  C. S. Roesler and A. H. Barnard, "Optical proxy for phytoplankton biomass in the absence of photophysiology: Rethinking the absorption line height," Methods Oceanogr. **7**, 79–94 (2013).
6.  J. Wollschlager, R. Röttgers, W. Petersen, and K. H. Wiltshire, "Performance of absorption coefficient measurements for the in situ determination of chlorophyll-a and total suspended matter," J. Exp. Mar. Biol. Ecol. **453**, 138–147 (2014).
7.  R. F. Davis, C. C. Moore, J. R. V. Zaneveld, and J. M. Napp, "Reducing the effects of fouling on chlorophyll estimates derived from long-term deployments of optical instruments," J. Geophys. Res. **102**(C3), 5851–5855 (1997).
8.  E. S. Boss, R. Collier, G. Larson, K. Fennel, and W. S. Pegau, "Measurements of spectral optical properties and their relation to biogeochemical variables and processes in Crater Lake, Crater Lake National Park, OR," Hydrobiologia **574**(1), 149–159 (2007).
9.  G. Dall'Olmo, T. K. Westberry, M. J. Behrenfeld, E. Boss, and W. H. Slade, "Significant contribution of large particles to optical backscattering in the open ocean," Biogeosciences **6**(6), 947–967 (2009).
10. T. K. Westberry, G. Dall'Olmo, E. Boss, M. J. Behrenfeld, and T. Moutin, "Coherence of particulate beam attenuation and backscattering coefficients in diverse open ocean environments," Opt. Express **18**(15), 15419–15425 (2010).
11. W. H. Slade, E. Boss, G. Dall'Olmo, M. R. Langner, J. Loftin, M. J. Behrenfeld, C. Roesler, and T. K. Westberry, "Underway and moored methods for improving accuracy in measurement of spectral particulate absorption and attenuation," J Atmos Ocean Tech **27**(10), 1733–1746 (2010).
12. G. Dall'Olmo, E. Boss, M. J. Behrenfeld, T. K. Westberry, C. Courties, L. Prieur, M. Pujo-Pay, N. Hardman-Mountford, and T. Moutin, "Inferring phytoplankton carbon and eco-physiological rates from diel cycles of spectral particulate beam-attenuation coefficient," Biogeosciences **8**(11), 3423–3439 (2011).
13. G. Dall'Olmo, E. Boss, M. Behrenfeld, and T. Westberry, "Particulate optical scattering coefficients along an Atlantic Meridional Transect," Opt. Express **20**(19), 21532–21551 (2012).
14. R. J. W. Brewin, G. Dall'Olmo, S. Pardo, V. van Dongen-Vogels, and E. S. Boss, "Underway spectrophotometry along the Atlantic Meridional Transect reveals high performance in satellite chlorophyll retrievals," Remote. Sens. Environ. **183**, 82–97 (2016).
15. H. Loisel and A. Morel, "Light scattering and chlorophyll concentration in case 1 waters: A reexamination," Limnol. Oceanogr. **43**(5), 847–858 (1998).
16. D. Stramski, G. Rosenberg, and L. Legendre, "Photosynthetic and optical-properties of the marine chlorophyte dunaliella-tertiolecta grown under fluctuating light caused by surface-wave focusing," Mar. Biol. **115**(3), 363–372 (1993).
17. H. C. van de Hulst, *Light scattering by small particles* (Wiley, New York, 1957).
18. C. F. Bohren and D. R. Huffman, *Absorption and scattering of light by small particles* (Wiley, New York, 1983).
19. A. Bricaud, A. Morel, and L. Prieur, "Optical efficiency factors of some phytoplankters," Limnol. Oceanogr. **28**(5), 816–832 (1983).
20. D. Stramski, A. Morel, and A. Bricaud, "Modeling the light attenuation and scattering by spherical phytoplanktonic cells - a retrieval of the bulk refractive-index," Appl. Opt. **27**(19), 3954–3956 (1988).
21. H. Houskeeper, D. Draper, R. Kudela, and E. Boss, "Chlorophyll absorption and phytoplankton size information inferred from hyperspectral particulate beam-attenuation," Appl. Opt. (2020).
22. E. Boss, N. Haëntjens, S. Ackleson, B. Balch, A. Chase, G. Dall'Olmo, S. Freeman, Y. Liu, J. Loftin, W. Neary, N. Nelson, M. Novak, W. Slade, C. Proctor, and P. Tortell, *Inherent Optical Property Measurements and Protocols: Best Practices for the Collection and Processing of Ship-Based Underway Flow-Through Optical Data.*, vol. Volume 4.0 (IOCCG, 2019).
23. E. Boss, M. Picheral, T. Leeuw, A. Chase, E. Karsenti, G. Gorsky, L. Taylor, W. Slade, J. Ras, and H. Claustre, "The characteristics of particulate absorption, scattering and attenuation coefficients in the surface ocean: Contribution of the tara oceans expedition," Methods Oceanogr. **7**, 52–62 (2013).
24. M. J. Behrenfeld, R. H. Moore, C. A. Hostetler, J. Graff, P. Gaube, L. M. Russell, G. Chen, S. C. Doney, S. Giovannoni, H. Liu, C. Proctor, L. M. Bolaños, N. Baetge, C. Davie-Martin, T. K. Westberry, T. S. Bates, T. G. Bell, K. D. Bidle, E. S. Boss, S. D. Brooks, B. Cairns, C. Carlson, K. Halsey, E. L. Harvey, C. Hu, L. Karp-Boss, M. Kleb, S. Menden-Deuer, F. Morison, P. K. Quinn, A. J. Scarino, B. Anderson, J. Chowdhary, E. Crosbie, R. Ferrare, J. W. Hair, Y. Hu, S. Janz, J. Redemann, E. Saltzman, M. Shook, D. A. Siegel, A. Wisthaler, M. Y. Martin, and L. Ziemba, "The north atlantic aerosol and marine ecosystem study (NAAMES): Science motive and mission overview," Front. Mar. Sci. **6**, 122 (2019).
25. P. J. Werdell, S. Bailey, G. Fargion, C. Pietras, K. Knobelspiesse, G. Feldman, and C. McClain, "Unique data repository facilitates ocean color satellite validation," Eos Trans. AGU **84**(38), 377 (2003).
26. A. Bricaud, M. Babin, A. Morel, and H. Claustre, "Variability in the chlorophyll-specific absorption-coefficients of natural phytoplankton - analysis and parameterization," J. Geophys. Res. **100**(C7), 13321–13332 (1995).

27. P. J. Werdell, C. W. Proctor, E. Boss, T. Leeuw, and M. Ouhssain, "Underway sampling of marine inherent optical properties on the Tara Oceans expedition as a novel resource for ocean color satellite data product validation," Methods Oceanogr. **7**, 40–51 (2013).

28. C. Brockmann, R. Doerffer, M. Peters, S. Kerstin, S. Embacher, and A. Ruescas, "Evolution of the C2RCC Neural Network for Sentinel 2 and 3 for the Retrieval of Ocean Colour Products in Normal and Extreme Optically Complex Waters," in *Living Planet Symposium*, vol. 740 of ESA Special Publication (2016), p. 54.

29. E. Alcântara, C. P. de Andrade, A. C. Gomes, N. Bernardo, A. F. Carmo, T. Rodrigues, and F. Watanabe, "Performance analysis of the c2rcc processor in estimate the water quality parameters in inland waters using olci/sentinel-3a images," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, (2018), pp. 9300–9303.

30. R. Sauzède, H. Claustre, C. Jamet, J. Uitz, J. Ras, A. Mignot, and F. D'Ortenzio, "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.: Oceans **120**(1), 451–470 (2015).

31. V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, (2010), pp. 807–814.

32. T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural networks for machine learning **4**, 26–31 (2012).

33. P. J. Huber, "Robust estimation of a location parameter," Ann. Math. Statist. **35**(1), 73–101 (1964).

34. T. Chai and R. R. Draxler, "Root mean square error (RMSE or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," Geosci. Model Dev. **7**(3), 1247–1250 (2014).