sensors

MDPI

*Technical Note*

# Quantifying the Impact of Linear Regression Model in Deriving Bio-Optical Relationships: The Implications on Ocean Carbon Estimations

**Marco Bellacicco [1,2,*], Vincenzo Vellucci [3], Michele Scardi [4], Marie Barbieux [1], Salvatore Marullo [2] and Fabrizio D'Ortenzio [1]**

[1] Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France
[2] Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), 00044 Frascati, Italy
[3] Sorbonne Université, CNRS, Institut de la Mer de Villefranche, IMEV, F-06230 Villefranche-sur-Mer, France
[4] Department of Biology, University of Rome "Tor Vergata", 00133 Rome, Italy
* Correspondence: marco.bellacicco@enea.it

Received: 24 May 2019; Accepted: 8 July 2019; Published: 9 July 2019

**Abstract:** Linear regression is widely used in applied sciences and, in particular, in satellite optical oceanography, to relate dependent to independent variables. It is often adopted to establish empirical algorithms based on a finite set of measurements, which are later applied to observations on a larger scale from platforms such as autonomous profiling floats equipped with optical instruments (e.g., Biogeochemical Argo floats; BGC-Argo floats) and satellite ocean colour sensors (e.g., SeaWiFS, VIIRS, OLCI). However, different methods can be applied to a given pair of variables to determine the coefficients of the linear equation fitting the data, which are therefore not unique. In this work, we quantify the impact of the choice of "regression method" (i.e., either type-I or type-II) to derive bio-optical relationships, both from theoretical perspectives and by using specific examples. We have applied usual regression methods to an in situ data set of particulate organic carbon (POC), total chlorophyll-*a* (TChla), optical particulate backscattering coefficient ($b_{bp}$), and 19 years of monthly TChla and $b_{bp}$ ocean colour data. Results of the regression analysis have been used to calculate phytoplankton carbon biomass ($C_{phyto}$) and POC from: i) BGC-Argo float observations; ii) oceanographic cruises, and iii) satellite data. These applications enable highlighting the differences in $C_{phyto}$ and POC estimates relative to the choice of the method. An analysis of the statistical properties of the dataset and a detailed description of the hypothesis of the work drive the selection of the linear regression method

**Keywords:** linear regression methods; bio-optical properties; BGC-Argo; satellite oceanography

## 1. Introduction

In technical and scientific applications, the linear regression fit is one of the most common models used to establish a relationship between two variables. Two families of statistical methods, i.e., type-I (ordinary least square, OLS) and type-II (e.g., standard major axis, SMA), were developed to perform a linear regression depending on the properties of the data set [1]. In optical oceanography, the rationale behind the choice of a given method for computing a linear regression fit is often an overlooked question, seldom explained or supported by statistical evidence.

This issue, which also pervades other fields of marine science such as fishery ecology, has already been highlighted by Laws et al. (1981) [2]: "the need to use model II (here type-II) regression methods in many applications have long been recognized, but a glance at the current literature will reveal that most biological oceanographers use model I (here type-I) regression methods exclusively

even when model II is clearly needed". Until the 1980s, the lack of widespread statistical software packages could have been a reason that favoured the application of the most common type-I method. However, as reported in Laws et al. (1981) [2] and Innamorati et al. (1990) [3], the need for a more careful choice of the regression models and methods was already clear in the oceanographic community.

Laws et al. (1981) [2] investigated the problem demonstrating that type-II methods should be applied to in situ data which are affected by instrument and sampling uncertainties. Specifically, they mentioned some common applications for which type-II methods are clearly needed, including, but not limited to, estimation of the phytoplankton chlorophyll-to-carbon ratio from chlorophyll vs. particulate carbon relationship [4,5].

The impact of the choice of regression method on optical oceanographic research is not minor. Linear regression models, in fact, are widely used to predict variables that are difficult or expensive to measure from field measurements. For example, the optical particulate backscattering coefficient ($b_{bp}$) is in situ measured or derived from ocean colour imagery. It is at the base of the estimation of the particulate organic carbon (POC) [6–10] and the phytoplankton carbon biomass ($C_{phyto}$) [11–13], both of which are fundamental variables used to constrain and understand the total carbon budget in the ocean [14,15]. Even though there are many works in which regression methods are correctly used and clearly mentioned in the text giving the opportunity to understand and reproduce the work [16–20], there are several cases where no information is provided about the linear regression method used [12,21–25], thus, preventing an evaluation of the impact of the methodology on the derived parameters. The lack of such information is crucial as the use of one method over another can return significantly different estimates on parameters. Indeed, differences due to the application of a sub-optimal regression method, instead, might be considered as errors and propagate if the wrong parameters are then used as inputs for modelling (e.g., empirical algorithms of ocean parameters). For this reason, as McArdle (2003) [26] pointed out "…if the slope, the intercept, or both parameters of the line are important, then care must be taken that the scientific conclusions follow from the data". In other words, the scientific conclusions must be based on the appropriate methodology, i.e., a methodology adapted to the statistical properties of the data set to be analysed.

In this regard, our primary goal is to evaluate the impact of the linear regression model (and methods) in optical and satellite oceanography. To do so, we quantify the differences between the results obtained applying diverse regression methods to the same bio-optical data set, and investigate the consequences of an inappropriate selection. Namely, we used field measurements collected in the north-western Mediterranean Sea at the BOUSSOLE (BOUee pour l'acquiSition d'une Serie Optique à Long termE) site [27–29], during three years of monthly oceanographic cruises (2011–2013) and more than one year (July 2013–November 2014) of Biogeochemical-Argo (aka BGC-Argo) vertical profiles. We applied both type-I and type-II regression methods to determine the coefficients of the linear equations of the total chlorophyll-*a* (TChla)-$b_{bp}$ and $b_{bp}$-POC relationships from discrete samples. Afterwards, we assessed the impact of the derived linear models with to the estimation of $C_{phyto}$ and POC base on the time series of $b_{bp}$ vertical profiles from the BGC-Argo floats and by applying either type-I or type-II regression method. Finally, a similar analysis was also conducted relying on satellite observations, namely $C_{phyto}$ was evaluated, by using 19-years of monthly TChla and $b_{bp}$.

## 2. Data and Methods

### *2.1. Theoretical Background*

Establishing a linear relationship between dependent and independent variables ($y$ and $x$, respectively) requires the computation, through linear regression analysis, of the coefficients of a linear equation, i.e., the slope ($B$) and intercept ($A$):

$$y = B \cdot x + A \quad (1)$$

The independent variable may be either uncontrolled (i.e., whose variability is affected by random phenomena) or controlled by the investigator, whereas the dependent variable has to be random by definition. The properties of both variables should drive the choice of an appropriate method to perform regression analysis, in contrast to linear correlation analysis, which is only aimed at measuring the strength of the linear relationship between two variables, independent of their properties and of any functional or causal link between them [1].

Although linear fitting is mainly used as a means to highlight a linear relationship between a pair of parameters, often the resulting equation is adopted as a model for further computations or analysis. Thus, the selection of the statistical method used to calculate the slope and the intercept of the linear model is of great importance to minimize the uncertainties associated with the dependent variable.

If the objective of linear regression is the interpolation or extrapolation of a data set, then the most common computational method, OLS, is appropriate independent of the properties of the variables [1,30]. This is the default method that most software use for computing and displaying a fitting line onto a scatter plot and the one which most users are familiar with. Nevertheless, other methods might be more appropriate depending on the goal of the analysis and on the properties of the data set. In such a context, the first step is selecting either a type-I or type-II regression method, which depends on whether the relationship between $x$ and $y$ variables is symmetric or asymmetric. This means whether or not the variables can be interchanged without altering the hypothesis/assumptions of the work and the derived parameters. An asymmetric relationship underpins a classical linear regression problem whereby the independent variable is characterized by null or low uncertainties as compared to the dependent one. This is the case, for instance, when the independent variable is fully controlled by the investigator or inherently free from uncertainties. A symmetric relationship, on the other hand, occurs when both variables show comparable uncertainties. Asymmetric relationships require type-I regression, whereas symmetric relationships require a type-II regression method [31]. While OLS is the only method to handle type-I problems, several methods can be adopted with a type-II regression. The OLS needs to be used only if the aim of the analysis is to predict the value of the dependent variable ($y$), given the independent one ($x$). This method minimizes the deviations of $y$ from the fitting line, i.e., those that are relevant to the prediction of unknown $y$ values. Both $x$, $y$ and deviations from the fitting line are instead relevant if the goal of the regression analysis is to assess the slope and/or intercept of the best regression line (see Figure A1 in Appendix A). In this case, OLS is not the most appropriate while type-II methods have to be followed. These type-II methods provide slope and intercept estimates that are, in most cases, significantly different from those obtained through OLS. Type-II methods are: major axis (MA), standard major axis (SMA) and ranged major axis (RMA). Note that the latter acronym is also used for reduced major axis, which is a synonym for SMA [31].

According to [1], MA is the appropriate method when: 1) data distribution is bivariate normal; 2) $x$ and $y$ variables are dimensionless or share the same units, and 3) the error variance is of similar magnitude for the two variables. RMA can handle variables whose units are heterogeneous because data is normalized before computing a MA. Because of this normalization, possible outliers have to be identified and eliminated from the data set, otherwise they could significantly alter the results. In the SMA, the slope is calculated as the ratio of the standard deviation of $y$ to that of $x$ [1]. However, SMA has two drawbacks: it should be computed only when the correlation between $x$ and $y$ and is significant, and its slope cannot be tested for significance.

Differences estimates of the slopes and intercepts obtained by applying different methods depend on the degree of correlation between $x$ and $y$. When $x$ and $y$ are strongly significantly correlated the differences between the three type-II methods based on the major axis are usually small. Nonetheless, all of them differ from OLS regression to a larger extent [1]. Yet, when the correlation coefficient tends to 1, the differences between type-II methods and OLS diminish respectively (see Appendix A).

In the following sections, we are concentrating only on SMA, as a type-II method, and OLS, as type-I, because both are the most widely used in optical oceanography and field sciences (for more

details about their mathematical treatment see Appendix A). Furthermore, MA and RMA cannot be applied to our dataset (i.e., heterogeneous parameters with different units, presence of outliers) [1,26,30,31].

### *2.2. Field and Satellite Measurements*

#### 2.2.1. Cruise Data

The BOUSSOLE project started in 2001, and its activities are developed around a bio-optical buoy located in the deep waters (2440 m) of the Ligurian Sea, one of the sub-basins of the Western Mediterranean Sea [27–29] (http://www.obs-vlfr.fr/Boussole/html/home/home.php). Figure 1 shows the area of study and the location of BOUSSOLE site. The BOUSSOLE site has been visited monthly for buoy servicing, during which 0–400 m casts are performed for the acquisition of conductivity, temperature and pressure (SBE 911plus, SeaBird Inc., Bellevue, WA, USA). Likewise, water samples are collected at 12 depths (400, 200, 150, 80, 70, 60, 50, 40, 30, 20, 10 and 5 m) with 12 L Niskin bottles mounted on a General Oceanic Rosette equipped with an SBE 32 Carousel Water Sampler, and then subsampled into polycarbonate bottles. An independent optical package is coupled to the conductivity, temperature and density (CTD)/Rosette for the acquisition of inherent optical properties. In this study, we used measurements of POC, TChla and $b_{bp}$ collected from October 2011 to December 2013, whose measurement protocols are summarized below:

Particulate organic carbon. Water is filtered through Whatman 25 mm GF/F glass-fiber filters (filtered volume from 2.27 to 5.5 L, depending on samples). Filters were washed beforehand using the soxhlet extraction method with dichloromethane. After filtration, samples were put into petri dishes and stored in liquid nitrogen for the duration of the cruise and then transferred into −80°C freezer until analysis, which took place within 12 months from sampling. Two days before the analysis, the filters were stored in a drying chamber at 50 °C during 1 night and then decarbonated with HCl solution. Finally, the filters were analyzed using a carbon, hydrogen and nitrogen (CHN) analyzer (Perkin Elmer 2400 series II) with the combustion analysis method. The relative uncertainty of the POC was estimated at < 1% from inter-calibration exercises of the analytical platform used here with other French laboratories (L. Coppola personal communication).

Total chlorophyll-*a*. Samples for the determination of phytoplankton pigments were filtered through 25 mm Whatman GF/F (0.7 μm retention capacity), put into petri dishes and stored in liquid nitrogen for the duration of the cruise. They were then transferred into −80 °C freezer until the analysis, which took place within 6–8 months from sampling. Pigments were identified and quantified with the high-performance liquid chromatography (HPLC) technique following [32]. The total chlorophyll-*a* concentration is computed as the sum of the concentrations of chlorophyll-*a*, chlorophyllide-*a*, and divinyl chlorophyll-*a*. Uncertainties in the methodology of the analytical platform used here were evaluated in a series of Round-Robin experiments (SeaHARRE-1 to 5; https://oceancolor.gsfc.nasa.gov/docs/technical/), and is of 6% for TChla used here (report of the analyses by J. Ras and M. Ouhssain).

Particulate backscattering coefficient. The total volume scattering function at 140°, $\beta(140, \lambda)$, was measured with a HydroScat-VI backscattering meter (HOBI Labs) at 6 wavelengths (420, 442, 488, 550, 620, 700 nm). The instrument was deployed within an independent Inherent Optical Properties (IOPs) package mounted below the CTD/Rosette, with the optics field of view looking at nadir. In this study, the down-cast was used to insure the medium was not perturbed during the measurement (i.e., to avoid possible disaggregation of particulate). Thus, the measurements had a maximum time lag of approximately 30 min with the discrete POC and HPLC sampled during the up-cast. The spectral particulate backscattering coefficient, $b_{bp}(\lambda)$, is obtained following [33], with few differences: 1) one dark 0–50 m $\beta(140, \lambda)$ profile was measured, averaged and subtracted from all profiles within each cruise; 2) data were binned around ±0.5 m of each nominal depth (1 m resolution); 3) the total absorption and beam attenuation coefficients used for the $\sigma$ correction [34] were measured respectively with an *a*-Sphere absorption meter (HOBI Labs) and a with Gamma-4 transmissometer (Hobi Labs).

2.2.2. BGC-Argo Floats Data

As part of this study, we used data obtained from two of BGC-Argo profiling floats deployed in the North Western Mediterranean Sea for a total 87 vertical profiles as displayed in Figure 1. The float referenced as WMO (World Meteorological Organization) N°6901496 was deployed in the Ligurian Sea near the BOUSSOLE buoy on 15 July 2013. In March, due to a significant bio-fouling of the optical sensors, this float was recovered, cleaned and deployed on again the 14 March 2014 under the reference of WMO N°6901776. Thus, the two BGC-Argo time series were joined end to end, corresponding to upward casts collected between 16 July 2013 and 25 October 2014.

All casts started from the parking depth at 1000 m at a time that was sufficient for surfacing around noon (local time). Vertical resolution of acquisition was 10 m between 1000 m and 250 m, 1 m between 250 m and 10 m, and 0.2 m between 10 m and the surface. Here, only "noon" casts were used. Following procedures described in [35], the $b_{bp}(700)$ profiles were calibrated, quality-controlled and additionally corrected by removing positive spikes greater than twice the 90th quantiles of the residual signal calculated as the difference between the profile and a median filter (window of 5 dots).

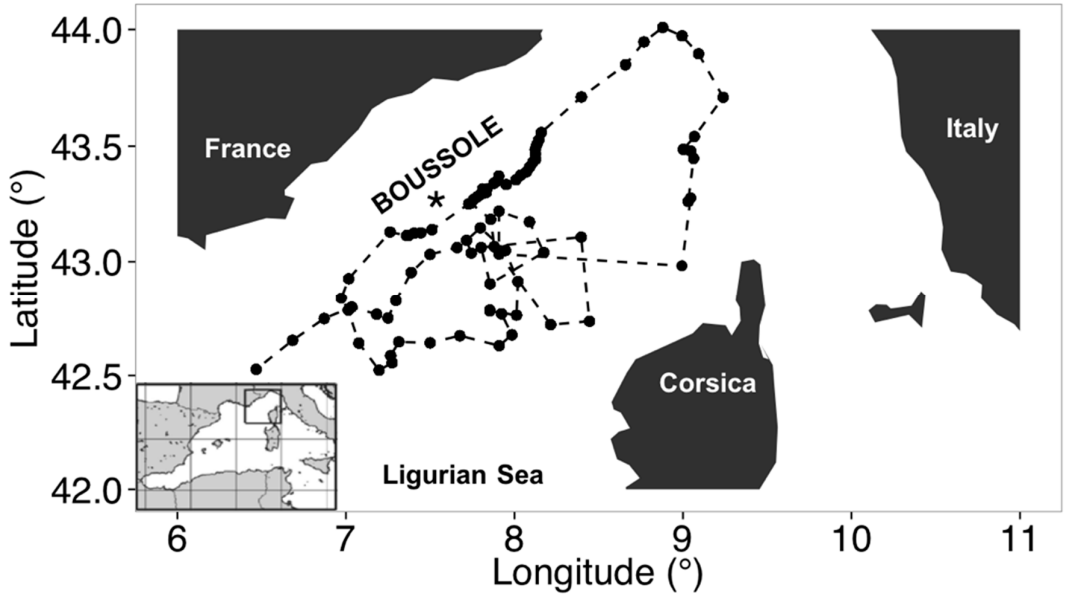


**Figure 1.** The northwestern Mediterranean Sea showing the southern coast of France, the island of Corsica, and the location of the BOUSSOLE buoy in the Ligurian Sea (black star) redrawn from [22]. Black dots are the locations where the float surfaced, while the float trajectory is overlaid in the plot with dashed black line.

2.2.3. Ocean Colour Data

The full ESA OC-CCI v3.0 (European Space Agency Ocean Colour-Climate Change Initiative version 3.0) monthly TChla (mg $m^{-3}$) and $b_{bp}$ ($m^{-1}$; 443nm) data time-series at 4 km resolution for the period 1997–2015 over the global ocean was downloaded from the ESA-CCI website (http://www.esa-oceancolour-cci.org/). ESA-CCI products are the results of the merging between SeaWiFS, MERIS, MODIS-Aqua, and VIIRS time-series [25,36–38]. TChla was estimated with a blending of the OCI (as implemented by NASA, itself a combination of CI and OC4), the OC5 (NASA, 2010) and the OC3 algorithms (http://www.esa-oceancolour-cci.org/?q=webfm_send/684). The Quasi-Analytical Algorithm (QAA) was used to compute $b_{bp}$ [39,40]. The accuracy of the QAA algorithm was demonstrated in several recent studies [12,13,19,23–25,41]. Both datasets were remapped at 100 km resolution, enough to resolve the broader oceanographic scales of variability. In

such a context, monthly TChla and $b_{bp}$ data were selected for the specific area of the northwestern Mediterranean basin to maintain the same domain of the analysis performed by using field measurements (see Figure 1).

### *2.3. Statistics*

In such a context, the following statistical indicators have been used to quantify the impact of regression methods in deriving bio-optical relationships and subsequent biogeochemical parameters (i.e., $C_{phyto}$ and POC):

(i) "*Anomalies*" here defined as the difference between parameters established with OLS and SMA linear regression methods.
(ii) The relative percentage differences (RPD) between parameters computed by the application of the unsuitable and suitable method.

## 3. Results and Discussion

In the following, we use models of $C_{phyto}$ as a function of the TChla -$b_{bp}$ relationship, and POC as a function of $b_{bp}$ as examples to highlight the impact of using a regression method not adapted to the data set on typical bio-optical oceanographic problems.

### *3.1. Total Chlorophyll-a versus Optical Backscattering*

Behrenfeld et al. (2005) [11] proposed the estimation of $C_{phyto}$ based on the relationship between TChla and $b_{bp}(443)$ and applied their model to SeaWiFS ocean color data on a global scale. Bellacicco et al. (2016) [12] revisited the model for regional tuning respective to the Mediterranean Sea and used the 555 nm band instead of 443 nm for $b_{bp}$. Recently, Bellacicco et al. (2018) [13] generalized this approach on a global scale by using $b_{bp}(443)$. The equation for the computation of $C_{phyto}$ is:

$$C_{phyto} = \left[ b_{bp}(\lambda) - b_{bp}^{k}(\lambda) \right] \cdot SF \quad (2)$$

where $\lambda$ is the wavelength. Here, we used $b_{bp}$ at 700 nm for compatibility also with BGC-Argo float measurements. The $b^{k}_{bp}(700)$ is the backscattering coefficient, at 700 nm, of the background fraction of non-algal particles that does not covary with TChla (e.g., heterotrophic bacteria and viruses) [11]. This value corresponds to the $b_{bp}(700)$ when TChla is zero: it is the intercept of the linear fit between the two variables. SF is the scaling factor chosen to give satellite Chl:C values (average value of 0.010) consistent with laboratory results, and also for the average contribution of phytoplankton to total particulate organic carbon (±30%) to be consistent with field estimates. In the original work, SF is equal to 13,000 mg C $m^{-2}$ [11]. Here, taking into account the change of wavelength for $b_{bp}$(700 nm instead of 443 nm) and to remain consistent with [11], we computed, according to in situ data, a SF of 16,455 mg C $m^{-2}$, 26% more with respect to the value of the original work. About the $b^{k}_{bp}(700)$, Bellacicco et al. (2016) [12] demonstrated that $b^{k}_{bp}$ (555) varies both in space and time. However, for sake of simplicity, we considered it to be a constant as in the original work of Behrenfeld et al. (2005) [11]. The main assumption of the model is the good relationship between TChla and $b_{bp}$ [12,13,23]. The first order co-variability between TChla and $b_{bp}$ is expected because phytoplankton cells contain TChla and also act as light backscatterers [13,42–44]. This co-variability also indicates that particles population abundance covaries with phytoplankton biomass, whereas the physiological photoacclimation process plays a secondary role in determining the chlorophyll variations [11,44]. In such a specific context, the underlying hypothesis is that TChla is the independent variable while $b_{bp}$ is the dependent one. There is no likelihood of interchanging the variables for the evaluation of the $b^{k}_{bp}$. Indeed, $b^{k}_{bp}$ is defined as the intercept of the linear regression fit between TChla and $b_{bp}$, and it is the $b_{bp}$ when TChla is equal to 0. The choice of the most appropriate regression method is founded upon which is the dependent variable and which is the independent one. In this case, the main goal is the estimation of a parameter ($b^{k}_{bp}$), allowing for the definition of another parameter ($C_{phyto}$), thus OLS is the preferable method to be applied [11–13].

The use of either OLS or SMA has consequences on the final $C_{phyto}$ biomass estimates that we can compute using in situ data. Furthermore, the intercept of the linear fit (i.e., $b^{k}_{bp}$ coefficient), in fact, has a biological meaning, as being the background contribution of the non-algal particles to the total $b_{bp}$ [11]. Figure 2 shows the TChla-$b_{bp}$ relationship with indicated both slopes and intercepts as computed by applying the two different regression methods. Here, the intercept varies from 5.8 (±0.5) × $10^{-4}$ to 4.5 (±0.3) × $10^{-4}$ $m^{-1}$ when calculated with OLS and SMA, respectively (Figure 2). These values are lower than those reported for the same region (though only surface measurements were used) [12]. This is consistent with a theoretically higher carbon (and its proxy $b_{bp}$) to TChla ratio in more illuminated waters [14].

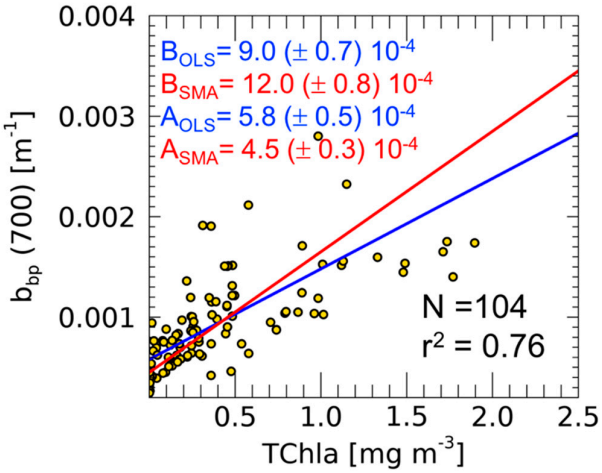


**Figure 2.** Scatter-plot and linear fit (continuous lines) calculated with ordinary least square (OLS) (blue) and standard major axis (SMA) (red) methods in the TChla-$b_{bp}$ relationship at the BOUSSOLE site. For both the coefficients, intercepts (A) and slopes (B), the standard errors are also indicated.

As discussed before, the definition of cause (independent variable) and effect (dependent variable) is thus fundamental and represents the working hypothesis. If one focuses on the relationship between TChla and $b_{bp}$, the goal being their comparison, the SMA (or RMA) has to be applied because it is statistically more robust in the context of the analysis of field measurements as explained in Section 2.1.

To further underline the difference in results when applying OLS or SMA methods, we computed the total $C_{phyto}$ from Equation (2) and by using the $b_{bp}(700)$ 0–400 m profiles collected during the BOUSSOLE cruises. To each profile, we applied the $b^{k}_{bp}(700)$ values (i.e., intercepts) obtained after the application of both OLS and SMA methods (Figure 2). The RPD on $C_{phyto}$ estimation is equal to 23.5% (overestimation of total $C_{phyto}$ using SMA instead of the appropriate OLS method).

Additionally, in order to evaluate how the estimate of $C_{phyto}$ changes when using either methods, we applied the relationships reported in Figure 2 to the time series of $b_{bp}(700)$ vertical profiles from the BGC-Argo dataset. When assessing the integral of $C_{phyto}$ over depth and time, the RPD between $C_{phyto,SMA}$ and $C_{phyto,OLS}$ is 28.7%. In this example, the use of SMA (the less adapted method) leads to an overestimation of $C_{phyto}$.

Furthermore, we conducted the analysis by using 19-years of monthly ocean colour data of TChla and $b_{bp}(443)$ for the period 1997–2015 as shown in Figure 3. As described earlier, the good relationship between TChla and $b_{bp}$ enables defining the $b^{k}_{bp}$ coefficient, a fundamental parameter for the $C_{phyto}$ computation. Figure 3a shows a moderate correlation between TChla and $b_{bp}$ ($r^2$ equal to 0.56) in the northwestern Mediterranean Sea. The correlation implies the reliable estimation of $b^{k}_{bp}$ coefficient by using all the pixels for the period 1997–2015 [12,13]. In such a context, the $b^{k}_{bp}$ is 8.5 (±0.2) × $10^{-4}$ $m^{-1}$ (with the OLS method), a value consistent with the order of magnitude found by a recent work always based on ocean colour data [13]. With the SMA, the $b^{k}_{bp}$ becomes lower with

respect to the computation performed by OLS: 6.3 (±0.3) × $10^{-4}$ $m^{-1}$. Figure 3b shows the subsequent crucial application of this coefficient on the satellite averaged $b_{bp}$ time series for the $C_{phyto}$ computation by using Equation (2) (443 nm instead of 700 nm as a wavelength of reference). Figure 3b shows how both the obtained $C_{phyto}$ time series follow a similar temporal pattern but with different values. Regarding the entire time series, the mean difference between $C_{phyto,SMA}$ (with SMA-based $b^k_{bp}$) and $C_{phyto,OLS}$ (with OLS-based $b^k_{bp}$) is 2.56 mg C $m^{-3}$, that is the 28% of the mean $C_{phyto,OLS}$: $C_{phyto,SMA}$ overestimates $C_{phyto,OLS}$. Therefore, in this specific context, there is a general overestimation of $C_{phyto}$ if one uses the SMA method instead of OLS. This critical point has to be taken into account because of its potential impact in the case of phytoplankton carbon studies on regional and global scales, mostly in ocean carbon budget studies.

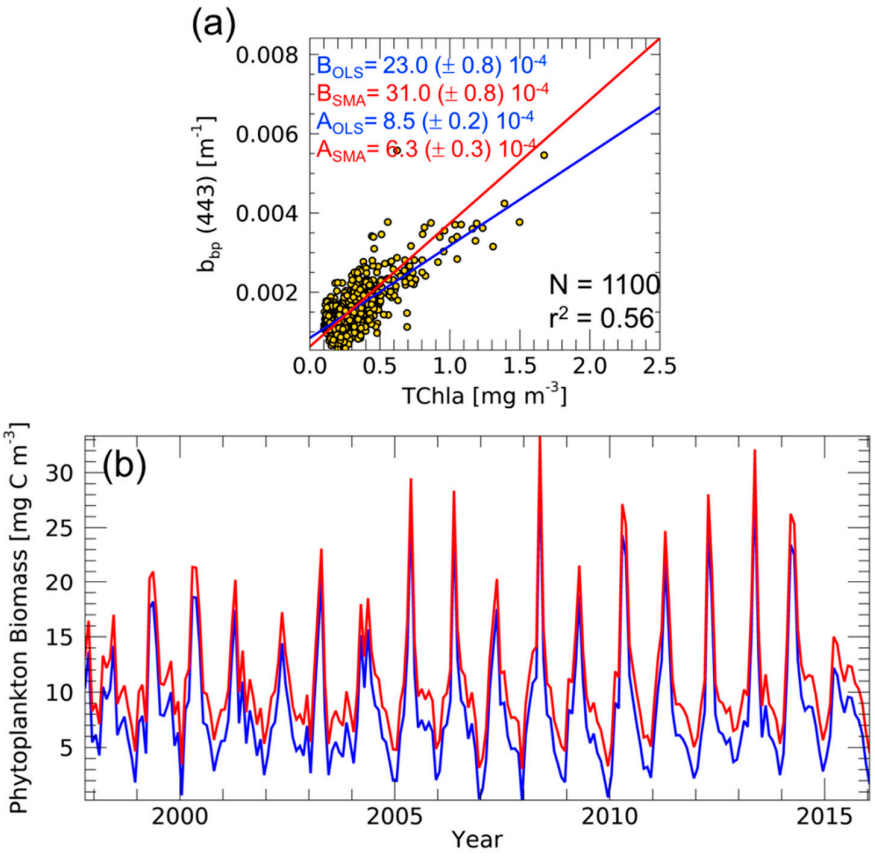


**Figure 3.** Scatter plot between TChla and $b_{bp}$ from ocean colour data in the northwestern Mediterranean Sea with linear fits (continuous lines) calculated with OLS (blue) and SMA (red) methods (**a**). For both the coefficients, intercepts (A) and slopes (B), the standard errors are also indicated. Time series of $C_{phyto}$ (**b**) based on the $b^k_{bp}$ computed by OLS (in blue) and SMA (in red) methods.

### *3.2. Optical Backscattering vs. Particulate Organic Carbon*

The POC is often linearly related to $b_{bp}$ [6–10,23] as follows:

$$POC = B \cdot b_{bp}(\lambda) + A \quad (3)$$

where B is the slope and A is the intercept of the linear regression fit following the general Equation (1). As suggested by Loisel et al., (2001) [6], sub-micrometer particles are efficient backscatterers, such that there is confirmation that the dominant contribution to particulate organic carbon in the ocean is due to sub-micrometer particles that are in sufficiently higher concentrations allowing for dominance of the $b_{bp}$ in oceanic water determining, therefore causing a strong correlation with POC.

The high correlation is caused by the dominance of organic particle concentration in controlling changes in both POC and $b_{bp}$. In such a context, it is possible to interchange POC and $b_{bp}$ for a simple comparison aimed at establishing a relationship between them, i.e., not for optimizing slope and intercept in view of a further application.

If the principal goal is to estimate POC, the appropriate method to be used is type-II regression, owing to methodological uncertainties in both measurements that have to be accounted for.

Several works reported results of a linear regression between POC and $b_{bp}$ from both satellite and in situ data (Table 2 in Thomalla et al., 2017 [10]). Figure 4 shows the established linear relationship between POC and $b_{bp}(700)$ using both SMA and OLS methods. Our estimates of the slope and the intercept, computed using both methods, are consistent with previous results from the Mediterranean Sea [6], Atlantic and Pacific Oceans [7], Southern Ocean [10] and North Atlantic Ocean [23].

Figure 4a shows that the slope and intercept computed using OLS are significantly lower than that computed with SMA (with a higher intercept point). This is a good example of the extent to which the choice of a regression method affects the estimate of the coefficients of the linear fit. As previously presented, these properties depend on minimizing y deviations only, for OLS, or the combination of x and y deviations for SMA. In the case of SMA, reduced uncertainties on the independent variable might balance higher uncertainties on the dependent variable, thus optimizing the overall agreement between the regression line and the data points. It is worth noting that the determination coefficient ($r^2$, i.e., the variance explained by the linear model) as well as the correlation coefficient (r, i.e., a measure of the linear correlation between two variables) are not dependent on the regression method.

To quantify how the estimate of POC changes when using OLS or SMA, we have applied the relationships reported in Figure 4a to a time series of $b_{bp}(700)$ vertical profiles acquired from BGC-Argo floats in the same area sampled to establish the linear models. Figure 4b shows the anomalies of POC as the difference between POC estimated using linear models based on OLS and SMA methods ($POC_{OLS}$ and $POC_{SMA}$ respectively). The anomalies, in general, are weak; however, several areas of large differences have impacted the computation of the POC budget over the time series. In detail, at the end of spring 2014 the largest anomalies are between 5.0 and 20.0 mg $m^{-3}$ in surface waters. In other periods, the anomalies are between −10.0 and +5.0 mg $m^{-3}$. When evaluating the integral of POC along the water column and over time, the RPD between $POC_{OLS}$ and $POC_{SMA}$ is 13.3%, showing the importance of selecting the correct regression method which avoids an incorrect estimate of the POC budget. In this example, the use of OLS (the unsuitable method) causes an overestimation of POC.

In both of the examples analyzed here (i.e., Figures 2 and 4), the differences in slope and intercept estimates obtained from OLS and SMA methods are quite substantial. These differences could be even greater for data sets with a lower correlation between the two variables [2]. This highlights that the selection of a statistical method not fitted to the data set may introduce substantial errors when the derived linear model is used to estimate the dependent variable from a direct or indirect measurement of the independent variable. The same effect is also evident in the empirical algorithms applied to satellite ocean colour imagery [6,8] or on BCG-Argo vertical profiles [10], which therefore should be assessed by applying the appropriate linear regression method.

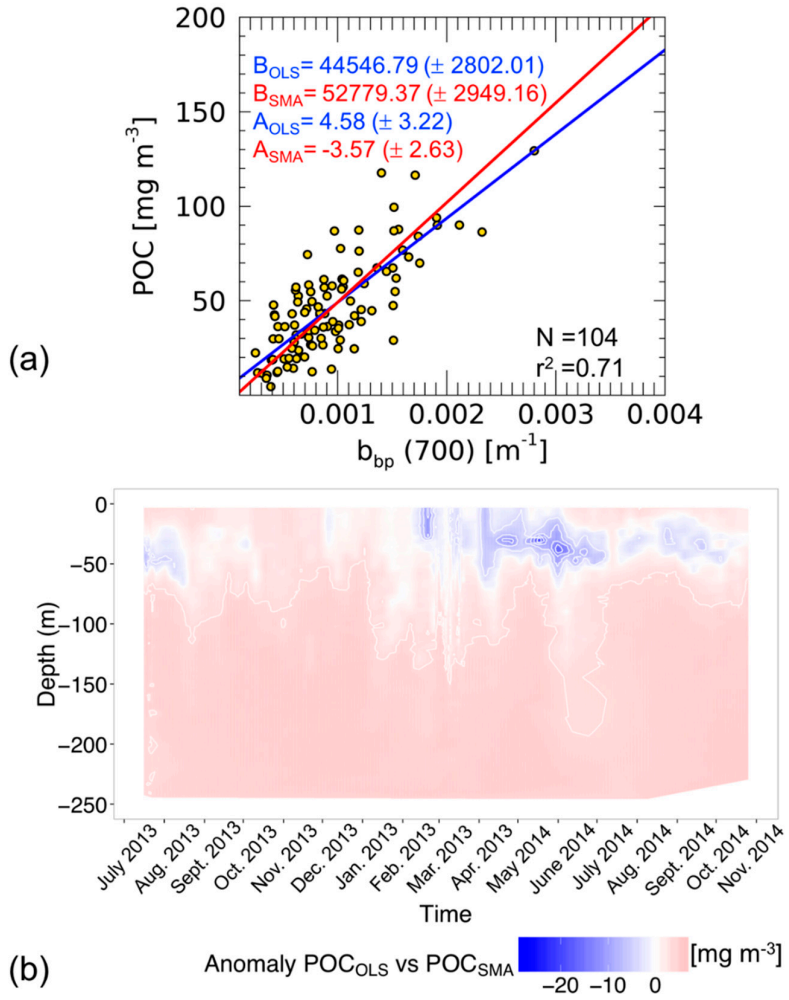


**Figure 4.** Scatter-plot and linear fits calculated with OLS (blue) and SMA (red) methods in the $b_{bp}$-POC relationship at the BOUSSOLE site (**a**). For both the coefficients, intercepts (A) and slopes (B), the standar errors are also indicated. Time series anomalies of particulate organic carbon (POC) derived from BGC-Argo $b_{bp}$ vertical profiles (0–250 m) using OLS and SMA and relationships (**b**).

## 4. Conclusion

In this study, we have used both type-I (i.e., OLS) and type-II (i.e., SMA) methods with bio-optical data collected over three years of monthly oceanographic cruises at the BOUSSOLE site (Figure 1) to derive linear regression coefficients (i.e., slopes and intercepts) that were then applied to BGC-Argo vertical profiles for the estimations of phytoplankton carbon biomass and particulate organic carbon. In addition, a specific analysis using ocean colour data is addressed. The main goal is to quantify the impact of the linear regression methods in satellite optical oceanography. Our analysis has shown that:

- The phytoplankton carbon biomass based on the TChla-$b_{bp}$ relationship needs to be computed using the OLS method due to the asymmetry assumption between the two variables. In such a context, the intercept of the linear fit between TChla and $b_{bp}$, which is necessary to compute the $C_{phyto}$, represents the fraction of $b_{bp}$ that does not co-vary with TChla, confirming that the dependent and independent parameters cannot be interchanged from a theoretical perspective. Only in this specific case, the application of the SMA is unsuitable, as it assumes symmetry of

the parameters. Its application always determines an overestimation of phytoplankton carbon biomass.

- For all linear regression analysis in which the main aim is to compare two parameters (e.g., $b_{bp}$-POC or TChla-$b_{bp}$), the most appropriate method is SMA due to its theoretical symmetry, and because of the uncertainties that affect both variables. It is thus possible to interchange the $x$ and $y$ axes without any impact on the interpretation of the results.

The main outcome of these examples is that the choice of method to determine the coefficients of the linear model significantly impacts $C_{phyto}$ and POC retrievals. The introduction of sizeable errors is a key factor in the carbon budget estimates when linear models are used on a global scale. Indeed, the total $C_{phyto}$:POC ratio utilizing the time series of $b_{bp}(700)$ vertical profiles give an RPD of 13.6% overestimation using $C_{phyto,SMA}$ to $POC_{OLS}$ (the ratio computed using both unsuitable approaches) with respect to the ratio $C_{phyto,OLS}$ to $POC_{SMA}$ (the appropriate methods to be used). It has to be kept in mind that two single relationships are applied to the full time-series of the BGC-Argo floats in the example shown here. It is understood however that spatio-temporal variations of the two relationships exist and could have an impact on the budget estimates. In this work, we thus highlighted the importance of the selection and use of the correct regression method. The choice of the model, and hence the method, has to be done a priori relative to any computation based on the data set properties. Given that, it cannot be overlooked that a fraction of the variation of the data around a linear regression fit can also be due to biogeochemical variability rather than error measurements. This type of error represents that portion of variability unresolved by the fitting function adopted, especially in case of ocean color data, where retrieval models are ofter oversimplified. Furthermore, in case of high correlation between variables, both slope and intercept estimations computed by type-I and type-II regression methods do not show large differences between. Therefore, for a correct application of linear regression methods in optical and satellite oceanography, a deeper study of the relationship between the two variables from a theoretical point of view needs to be performed. In fact, as demonstrated, the influence of the unsuitable method in cases of carbon estimations can be considerable and potentially impactful in the context of global carbon budget studies from space or by using field measurements.

**Author Contributions:** Conceptualization, M.Be., V.V and M.S.; Methodology, M.Be., V.V., M.S.; Writing, Review and Editing, V.V., M.S., M.Ba., S.M., F.D'O. contributed equally.

**Funding:** The BOUSSOLE project is co-funded by the European Space Agency (ESA) and Centre Nationales d'Etudes Spatiales CNES; France). This study was also supported by the following research projects: remOcean (funded by the European Research Council, grant agreement 246777), NAOS (funded by the Agence Nationale de la Recherche in the frame of the French "Equipement d'avenir" program, grant agreement ANR J11R107-F). M.B. conceived this work during his stays at the Laboratoire d'Oceanographie de Villefranche sur Mer (LOV) supported by a postdoctoral fellowship of the (CNES; France). Currently, M.B. has a postdoctoral fellowship by the ESA. This work was supported by the ESA Living Planet Fellowship Project PHYSIOGLOB: Assessing the inter-annual physiological response of phytoplankton to global warming using long-term satellite observations, 2018–2020.

**Acknowledgments:** We thank Melek Golbol and staff of the Villefranche Marine Optics group, for their work in collecting the BOUSSOLE datasets. We also thank David Antoine, Louis Legendre, Nathan Briggs, Antonio Bellacicco for their comments, criticism and suggestions on the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Mathematical Details

In this section, the differences between OLS and SMA methods are described with mathematical detail and from a theoretical point of view.

The linear regression is used to compute the parameters of a first-degree equation relating variables y and x. The equation for a simple linear regression is defined as:

$$\hat{y} = B \cdot x + A \quad \text{(A1)}$$

This corresponds to the equation of a straight line that fitting a cloud of points, on a Cartesian coordinate system, in some optimal way, and establish a model to predict $\hat{y}$ for any value of $x$. $A$ is the estimate of the intercept of the regression line with the $y$ axis and $B$ is the slope of the regression line or regression coefficient. When using this type of regression, one must be aware of the fact that a linear model is imposed on the dataset. In other words, one assumes that the relationship between the two variables may be well described by a straight line and that the vertical dispersion of observed values above and below the line is the result of a random process. The difference between the observed and estimated values along $y$ is:

$$\varepsilon_i = (y_i - \hat{y}) \quad \text{(A2)}$$

for each observation $i$ and may be positive or negative. The term $\varepsilon_i$ is called the residual value of observation $y_i$ after fitting the regression line. If we include the $\varepsilon_i$ term in the Equation (A1), it allows one to describe exactly the ordinate value $y_i$ of each point ($x_i$, $y_i$) of the dataset; $y_i$ is equal to the value $\hat{y}_i$ , as predicted by the regression equation plus the residual $\varepsilon_i$, as follows:

$$y_i = \hat{y}_i + \varepsilon_i = B \cdot x_i + A + \varepsilon_i \quad \text{(A3)}$$

This equation is the linear model of the relationship. The term $\hat{y}_i$ is the predicted or fitted value corresponding to each observation $i$. The model assumes that deviations from the linear relationship are only on the vertical axis (i.e., "errors" $\varepsilon_i$ are only associated to the response variable $y_i$; "errors" associated with the estimation of $x$, $\delta_i$, are equal to zero). The term "error" is the term traditionally used to indicate the deviations of all kinds due to random processes, and including those linked to measurements or methodology.

In simple linear regression, one is looking for the straight line with equation $\hat{y} = B \cdot x + A$ that minimizes the sum of square of the vertical residuals $\varepsilon_i$, between the observed values and the regression line (see Figure A1a). This is the principle of the least squares method. This sum of square residuals, $\Sigma (y_i - \hat{y}_i)^2$ (as in case of the OLS method), offers the advantage of providing a unique solution, which would not be the case if one chose to minimize, for example $\Sigma |y_i - \hat{y}_i|$. On the other hand, the SMA method minimizes the sum of the product of the $x$ and $y$ deviations, $\Sigma (x_i - \hat{x}_i)(y_i - \hat{y}_i)$ which is the equivalent to the area of triangles formed by the deviations of a point from the line in the $x$ and $y$ directions (see Figure A1a). Figure A1b shows how the slopes and intercepts change after the application of OLS and SMA methods.

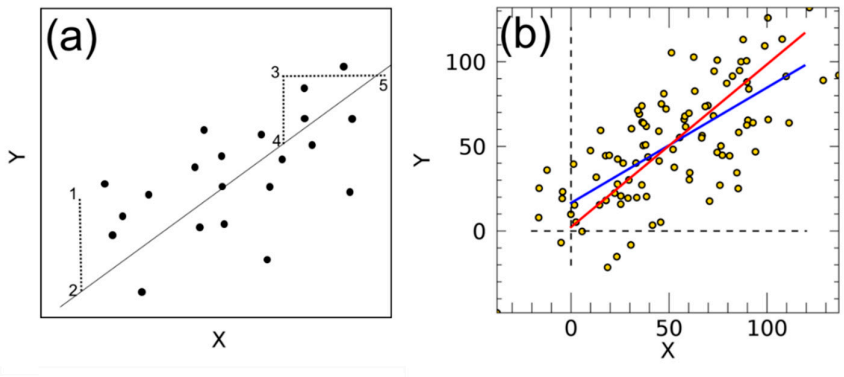


**Figure A1.** For an OLS line, the error is defined as the vertical dispersion of a point from the straight line (distance 1 to 2) and the quantity minimized is the sum of squares of these linear distances. In case of SMA, on the other hand, the error is defined as the area of the triangle 3-4-5 and the quantity minimized is the sum of these area (redrawn from Smith et al., 2009 [30]) (**a**). Scatter plot and linear fits calculated with OLS (blue) and SMA (red) methods by using a syntehtical datasets with a normal distributed error added to both X and Y variables (**b**).

The solution for Equation (A1), satisfying the least square criterion can be found using partial derivatives and is:

$$B_{OLS} = \frac{s_{xy}}{s_x^2} \; ; \; A = \bar{y} - B_{OLS} \cdot \bar{x} \tag{A4}$$

in which $s_{xy}$ and $s_x^2$ are estimates of the covariances and variance, respectively.

In case of SMA, or reduce major axis (e.g., a specific case of type-II model), the slope is computed as:

$$B_{SMA} = \sqrt{\frac{s_y^2}{s_x^2}} = \pm \left(\frac{s_y}{s_x}\right) \tag{A5}$$

This formula is obtained from the classical and more general formula:

$$B = \frac{s_y^2 - \theta s_x^2 + \sqrt{\left(s_y^2 - \theta s_x^2\right)^2 + 4\theta s_{xy}^2}}{2s_{xy}} \tag{A6}$$

where $s_y^2$ and $s_x^2$ are the estimated variances of $y$ and $x$, respectively, $s_{xy}$ is their covariance and $\theta$ is the ratio between the variances of the two errors terms, $\frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$. The formula of the slope of the SMA method is a specific case of the more general Equation (A6). In the SMA slope computation, it is assumed that the error variances $\sigma_\varepsilon^2$ and $\sigma_\delta^2$ of $y$ and $x$, respectively, are proportional to their respective variances $\sigma_x^2$ and $\sigma_y^2$. This means that:

$$\frac{\sigma_\varepsilon^2}{\sigma_y^2} = \frac{\sigma_\delta^2}{\sigma_x^2} \tag{A7}$$

Replacing the variances, $\sigma_\varepsilon^2$ and $\sigma_\delta^2$, with their unbiased estimates, $s_y^2$ and $s_x^2$, gives a value of $\theta$ equal to $\frac{s_y^2}{s_x^2}$.

Since the square root $\sqrt{\frac{s_y^2}{s_x^2}}$ can be positive or negative, the sign of the slope estimate is given by the sign of the Pearson correlation coefficient ($r$) which is the same as that of the covariance $s_{xy}$ in the denominator of the Equation (A7) or numerator of Equation (A5). The $B_{SMA}$ is also the geometric mean of the OLS regression coefficient of $y$ on $x$, thus it can be also defined as:

$$B_{SMA} = \frac{B_{OLS}}{r_{xy}} \; when \; r_{xy} \neq 0 \tag{A8}$$

Therefore, one can compute $B_{SMA}$ using the value of $B_{OLS}$ and $r_{xy}$ provided by an OLS regression algorithm. This equation also shows that when the variables are highly correlated ($r$ tends to 1), $B_{SMA}$ tends to $B_{OLS}$. When they are not, $B_{SMA}$ is always larger than $B_{OLS}$ for positive values of $r$, and smaller for negative values of $r$, in other words, $B_{OLS}$ is always closer to 0 than $B_{SMA}$. For additional details about the theory, mathematical features and a deeper description of regression models and methods, see also [1,26,30,31].

## References

1. Legendre, P.; Legendre, L.F. *Numerical Ecology*; Elsevier: Amsterdam, The Netherlands, 2012; Volume 24.
2. Laws, E.A.; Archie, J.W. Appropriate use of regression analysis in marine biology. *Mar. Biol.* **1981**, *65*, 13–16.

3. Innamorati, M.; Ferrari, I.; Marino, D.; Ribera D'Alcalà, M. Metodi Nell'Ecologia del Plancton Marino. In *Nova Thalassia*; Min. Amb., S.I.B.M. ed., Provincia e Comune, Università di Trieste: Trieste, Italia, 1990; Volume 11.
4. Steele, J.H.; Baird, I.E. Further relations between primary production, chlorophyll, and particulate carbon. *Limnol. Oceanogr.* **1962**, *7*, 42–44.
5. Banse, K. Determining the carbon-to-chlorophyll ratio of natural phytoplankton. *Mar. Biol.* **1977**, *41*, 199–212.
6. Loisel, H.; Bosc, E.; Stramski, D.; Oubelkheir, K.; Deschamps, P.Y. Seasonal variability of the backscattering coefficient in the Mediterranean Sea based on Satellite SeaWiFS imagery. *Geophys. Res. Lett.* **2001**, *28*, 4203–4206.
7. Stramski, D.; Reynolds, R.A.; Babin, M.; Kaczmarek, S.; Lewis, M.R.; Röttgers, R.; Claustre, H. Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans. *Biogeosciences* **2008**, *5*, 171–201.
8. Neukermans, G.; Loisel, H.; Mériaux, X.; Astoreca, R.; McKee, D. In situ variability of mass-specific beam attenuation and backscattering of marine particles with respect to particle size, density, and composition. *Limnol. Oceanogr.* **2012**, *57*, 124–144.
9. Cetinić, I.; Perry, M.J.; Briggs, N.T.; Kallin, E.; D'Asaro, E.A.; Lee, C.M. Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment. *J. Geophys. Res. Ocean.* **2012**, doi:10.1029/2011JC007771.
10. Thomalla, S.J.; Ogunkoya, G.; Vichi, M.; Swart, S. Using optical sensors on gliders to estimate phytoplankton carbon concentrations and chlorophyll-to-carbon ratios in the Southern Ocean. *Front. Mar. Sci.* **2017**, *4*, 34.
11. Behrenfeld, M.J.; Boss, E.; Siegel, D.A.; Shea, D.M. Carbon-based ocean productivity and phytoplankton physiology from space. *Glob. Biogeochem. Cycles* **2005**, *19*, doi:10.1029/2004GB002299.
12. Bellacicco, M.; Volpe, G.; Colella, S.; Pitarch, J.; Santoleri, R. Influence of photoacclimation on the phytoplankton seasonal cycle in the Mediterranean Sea as seen by satellite. *Remote Sens. Environ.* **2016**, *184*, 595–604.
13. Bellacicco, M.; Volpe, G.; Briggs, N.; Brando, V.; Pitarch, J.; Landolfi, A.; Colella, S.; Marullo, S.; Santoleri, R. Global distribution of non-algal particles from ocean color data and implications for phytoplankton biomass detection. *Geophys. Res. Lett.* **2018**, *45*, 7672–7682.
14. Halsey, K.H.; Jones, B.M. Phytoplankton strategies for photosynthetic energy allocation. *Annu. Rev. Mar. Sci.* **2015**, *7*, 265–297.
15. Evers-King, H.; Martinez-Vicente, V.; Brewin, R.J.; Dall'Olmo, G.; Hickman, A.E.; Jackson, T.; Roy, S. Validation and Intercomparison of Ocean Color Algorithms for Estimating Particulate Organic Carbon in the Oceans. *Front. Mar. Sci.* **2017**, *4*, 251.
16. Briggs, N.; Perry, M.J.; Cetinić, I.; Lee, C.; D'Asaro, E.; Gray, A.M.; Rehm, E. High-resolution observations of aggregate flux during a sub-polar North Atlantic spring bloom. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **2011**, *58*, 1031–1039.
17. Mélin, F.; Vantrepotte, V.; Clerici, M.; D'Alimonte, D.; Zibordi, G.; Berthon, J.F.; Canuti, E. Multi-sensor satellite time series of optical properties and chlorophyll-a concentration in the Adriatic Sea. *Prog. Oceanogr.* **2011**, *91*, 229–244.
18. Zibordi, G.; Berthon, J.F.; Mélin, F.; D'Alimonte, D. Cross-site consistent in situ measurements for satellite ocean color applications: The BiOMaP radiometric dataset. *Remote Sens. Environ.* **2011**, *115*, 2104–2115.
19. Brewin, R.J.; Sathyendranath, S.; Müller, D.; Brockmann, C.; Deschamps, P.Y.; Devred, E.; Groom, S. The Ocean Colour Climate Change Initiative: III. A round-robin comparison on in-water bio-optical algorithms. *Remote Sens. Environ.* **2015**, *162*, 271–294.
20. Martínez-Vicente, V.; Evers-King, H.; Roy, S.; Kostadinov, T.S.; Tarran, G.A.; Graff, J.R.; Röttgers, R. Intercomparison of ocean color algorithms for picophytoplankton carbon in the ocean. *Front. Mar. Sci.* **2017**, *4*, 378.
21. Pitarch, J.; Volpe, G.; Colella, S.; Krasemann, H.; Santoleri, R. Remote sensing of chlorophyll in the Baltic Sea at basin scale from 1997 to 2012 using merged multi-sensor data. *Ocean Sci.* **2016**, *12*, 379–389.
22. Antoine, D.; Siegel, D.A.; Kostadinov, T.; Maritorena, S.; Nelson, N.B.; Gentili, B.; Guillocheau, N. Variability in optical particle backscattering in contrasting bio-optical oceanic regimes. *Limnol. Oceanogr.* **2011**, *56*, 955–973.

23. Graff, J.R.; Westberry, T.K.; Milligan, A.J.; Brown, M.B.; Dall'Olmo, G.; van Dongen-Vogels, V.; Behrenfeld, M.J. Analytical phytoplankton carbon measurements spanning diverse ecosystems. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **2015**, *102*, 16–25.
24. Pitarch, J.; Bellacicco, M.; Volpe, G.; Colella, S.; Santoleri, R. Use of the quasi-analytical algorithm to retrieve backscattering from in-situ data in the Mediterranean Sea. *Remote Sens. Lett.* **2016**, *7*, 591–600.
25. Constantin, S.; Doxaran, D.; Constantinescu, Ș. Estimation of water turbidity and analysis of its spatio-temporal variability in the Danube River plume (Black Sea) using MODIS satellite data. *Cont. Shelf Res.* **2016**, *112*, 14–30.
26. McArdle, B.H. Lines, models, and errors: Regression in the field. *Limnol. Oceanogr.* **2003**, *48*, 1363–1366.
27. Antoine, D.; Chami, M.; Claustre, H.; D'Ortenzio, F.; Morel, A.; Bécu, G.; Scott, A.J. *BOUSSOLE: A Joint CNRS-INSU, ESA, CNES, and NASA Ocean Color Calibration and Validation Activity*; NASA: Washington, DC, USA, 2006.
28. Antoine, D.; D'Ortenzio, F.; Hooker, S.B.; Bécu, G.; Gentili, B.; Tailliez, D.; Scott, A.J. Assessment of uncertainty in the ocean reflectance determined by three satellite ocean color sensors (MERIS, SeaWiFS and MODIS-A) at an offshore site in the Mediterranean Sea (BOUSSOLE project). *J. Geophys. Res. Ocean.* **2008**, doi:10.1029/2007JC004472.
29. Antoine, D.; Guevel, P.; Desté; J.F.; Bécu, G.; Louis, F.; Scott, A.J.; Bardey, P. The "BOUSSOLE" buoy-a new transparent-to-swell taut mooring dedicated to marine optics: Design, tests, and performance at sea. *J. Atmos. Ocean. Technol.* **2008**, *25*, 968–989.
30. Smith, R.J. Use and misuse of the reduced major axis for line-fitting. *Am. J. Phys. Anthropol.* **2009**, *140*, 476–486.
31. Sokal, R.R.; Rohlf, F.J. *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd ed.; W. H. Freeman and Company: New York, NY, USA, 1995; Volume 887.
32. Ras, J.; Claustre, H.; Uitz, J. Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: Comparison between in situ and predicted data. *Biogeosciences* **2008**, *5*, 353–369.
33. Kheireddine, M.; Antoine, D. Diel variability of the beam attenuation and backscattering coefficients in the northwestern Mediterranean Sea (BOUSSOLE site). *J. Geophys. Res. Ocean.* **2014**, *119*, 5465–5482.
34. Maffione, R.A.; Dana, D.R. Instruments and methods for measuring the backward-scattering coefficient of ocean waters. *Appl. Opt.* **1997**, *36*, 6057–6067, doi:10.1364/ao.36.006057.
35. Organelli, E.; Claustre, H.; Bricaud, A.; Barbieux, M.; Uitz, J.; D'Ortenzio, F.; Dall'Olmo, G. Bio-optical anomalies in the world's oceans: An investigation on the diffuse attenuation coefficients for downward irradiance derived from Biogeochemical Argo float measurements. *J. Geophys. Res. Ocean.* **2017**, doi:10.1002/2016JC012629.
36. Mélin, F.; Vantrepotte, V.; Chuprin, A.; Grant, M.; Jackson, T.; Sathyendranath, S. Assessing the fitness-for-purpose of satellite multi-mission ocean color climate data records: A protocol applied to OC-CCI chlorophyll-a data. *Remote Sens. Environ.* **2017**, *203*, 139–151.
37. Mélin, F.; Sclep, G. Band shifting for ocean color multi-spectral reflectance data. *Opt. Express* **2015**, *23*, 2262–2279.
38. Sathyendranath, S.; Brewin, R.J.W.; Jackson, T.; Melin, F.; Platt, T. Ocean-colour products for climate-change studies: What are their ideal characteristics? *Remote Sens. Environ.* **2017**, *203*, 125–138.
39. Lee, Z.; Carder, K.L.; Arnone, R.A. Deriving inherent optical properties from water color: A multiband quasi-analytical algorithm for optically deep waters. *Appl. Opt.* **2002**, *41*, 5755–5772.
40. Lee, Z. Update of the Quasi-Analytical Algorithm (QAA_v6). 2014. Available online: http://www.ioccg.org/groups/Software_OCA/QAA_v6_2014209.pdf (accessed on 8 July 2019).
41. Mélin, F.; Berthon, J.F.; Zibordi, G. Assessment of apparent and inherent optical properties derived from SeaWiFS with field data. *Remote Sens. Environ.* **2005**, *97*, 540–553.
42. Dall'Olmo, G.; Westberry, T.K.; Behrenfeld, M.J.; Boss, E.; Slade, W.H. Significant contribution of large particles to optical backscattering in the open ocean. *Biogeosciences* **2009**, *6*, 947–967.
43. Dall'Olmo, G.; Boss, E.; Behrenfeld, M.J.; Westberry, T.K. Particulate optical scattering coefficients along an Atlantic Meridional Transect. *Opt. Express* **2012**, *20*, 21532–21551.

44. Siegel, D.A.; Behrenfeld, M.J.; Maritorena, S.; McClain, C.R.; Antoine, D.; Bailey, S.W.; Eplee, R.E., Jr. Regional to global assessments of phytoplankton dynamics from the SeaWiFS mission. *Remote Sens. Environ.* **2013**, *135*, 77–91.