1 **ESTIMATION OF PHYTOPLANKTON PIGMENTS FROM OCEAN-COLOR**

2 **SATELLITE OBSERVATIONS IN THE SENEGALO-MAURITANIAN REGION BY**

3 **USING AN ADVANCED NEURAL CLASSIFIER**

4 By

6 Khalil Yala[1], N'Dye Niang[2], Julien Brajard[1,4], Carlos Mejia[1], Maurice Ouattara[2], Roy El

7 Hourany[1], Michel Crépon[1] and Sylvie Thiria[1,3]

10 [1] IPSL/LOCEAN, Sorbonne Université (Université Paris6, CNRS, IRD, MNHN), 4 Place

11 Jussieu, 75005 Paris, France

12 [2] CEDRIC, CNAM, 292 rue Saint Martin, 75003 Paris, France

13 [3] UVSQ, F-78035, Versailles, France

14 [4] Nansen Center, Thormøhlensgate 47, 5006, Bergen, Norway

16 Corresponding author: Michel Crepon (crepon@locean-ipsl.upmc.fr)

19 **ABSTRACT**

20 We processed daily ocean-color satellite observations to construct a monthly climatology of

21 phytoplankton pigment concentrations in the Senegalo-Mauritanian region. Thanks to the difficulty of

22 the problem, we proposed a new method. Our new method primarily consists in associating, in well-

23 identified clusters, similar pixels in terms of ocean-color parameters and in situ pigment concentrations

24 taken from a global ocean database. The association is carried out using a new Self Organized Map

25 (2S-SOM). Its major advantage is to allow taking into account the specificity of the optical properties

26 of the water by adding specific weights to the different ocean color parameters and the in situ

27 measurements. In the retrieval phase, the pigment concentration of a pixel is estimated by taking the

28 pigment concentration values associated with the 2S-SOM cluster presenting the ocean-color satellite

29 spectral measurements, which are the closest to those of the pixel under study according to some

30 distance. The method was validated by using a cross-validation procedure. We focused our study on

31 the fucoxanthin concentration, which is related to the abundance of diatoms. We showed that the

32 fucoxanthin starts to develop in December, presents its maximum intensity in March when the

33 upwelling intensity is maximum, extends up to the coast of Guinea in April and begins to decrease in

34 May. The results are in agreement with previous observations and recent in situ measurements. The

35 method is very general and can be applied in every oceanic region.

38 **1 - INTRODUCTION**

40 Phytoplankton are the basis of the ocean food web and consequently drive the ocean productivity.

41 They also play a fundamental role in climate regulation by trapping atmospheric carbon dioxide (CO_2)

42 through gas exchanges at the sea surface, and consequently lowering the rate of anthropogenic increase

43 in the atmosphere of CO_2 concentration by about 25% (*Le Quéré et al, 2018*). With the growing interest

44 in climate change, one may ask how the different phytoplankton populations will respond to changes

45 in ocean characteristics (temperature, salinity, acidity) and nutrient supply, which presents an

46 important societal impact with respect to both climate and fisheries, with a possible effect on fish

47 grazing phytoplankton via the marine food chain.

48 Methods for identifying phytoplankton have greatly progressed during the last two decades.

49 Phytoplankton were first described by microscopy. Microscopy is time consuming and is unable to

50 identify picoplankton. Imaging flow cytometry (IFC) has renewed microscopic methods, thanks to the

51 speed at which they are able to characterize phytoplankton in a water sample (IOCCG report n°15,

52 2014). An alternative method is the analysis of seawater samples by high-performance liquid

53 chromatography (HPLC) which is widely used to categorize broad phytoplankton groups such as PFT

54 or PSC (*Jeffreys et al*, 1997, *Brewin et al,* 2010, *Hirata et al,* 2011). HPLC enables the identification

55 of 25 to 50 pigments within a single analysis, which is much easier and faster to conduct than

56 microscopic observations (*Sosik, H.M et al,* 2014*)*. Each phytoplankton group is associated with

57 specific diagnostic pigments, and a conversion formula, the so-called "Diagnostic Pigment Analysis"

58 can be derived to estimate the percentage of each group from the pigment measurements (*Vidussi et*

59 *al,* 2001; *Uitz et al*, 2010). HPLC measurements are now recognized as the standard for calibrating

60 and validating satellite-derived chlorophyll-a concentration and for mapping groups of phytoplankton

61 (IOCCG report n°15, 2014).

62 The use of satellite ocean color sensor measurements has permitted to map the ocean surface at a daily

63 frequency. Satellite sensors measure the sunlight, at several wavelengths, backscattered by the ocean.

64 The downwelling sunlight interacts with the seawater through backscattering and absorption in such a

65 manner that the upwelling radiation transmitted to the satellite ('water-leaving' reflectance) contains

66 information related to the composition of the seawater. The light transmitted to the satellite depends

67   on the phytoplankton cell shape (backscattering), its pigments (absorption), the dissolved matter (e.g.

68   CDOM).

69   This upwelling radiation, the so-called remotely sensed reflectance $\rho_w(\lambda)$, is determined by the spectral

70   absorption a and backscattering (b_b (m^{-1})) coefficients of the ocean (pure water and various particulate

71   and dissolved matters) using the simplified formulation (*Morel* and *Gentili*, 1996):

72

73    $$\rho_w(\lambda) = G\, b_b\,(\lambda)/(a(\lambda) + b_b(\lambda)) \qquad (1)$$

74

75   where (a (m^{-1})) is the sum of the individual absorption coefficients of water, phytoplankton pigments,

76   colored dissolved organic matter, and detrital particles, (b_b (m^{-1})) depends on the shape of the

77   phytoplankton species. G is a parameter mainly related to the geometry of the situation (sensor and

78   solar angles) but also to environmental parameters (wind, aerosols).

79   In the open ocean far from the coast (in case-1 waters), the light seen by the satellite sensor mainly

80   contains information on phytoplankton abundance and diversity. Ocean-color measurements have

81   been first used intensively to estimate chlorophyll-*a* concentration (*chl-a* in the following) in the

82   surface waters of the ocean, marginal seas and lakes. (*Longhurst et al.,* 1995; *Antoine et al.,* 1996;

83   *Behrenfeld and Falkowski*, 1997; *Behrenfeld et al.,* 2005; *Westberry et al.*, 2008).

84   It has been shown that it is also possible to extract additional information such as phytoplankton size-

85   classes (PSC) by using some relationship between chlorophyll concentration and PSC (*Uitz et al.*, 2006;

86   *Ciotti and Bricaud*, 2006; *Hirata et al.*, 2008; *Mow and Yoder,* 2010). These algorithms try to establish

87   a relationship between the *chl-a* concentration and the *chl-a* concentration fractions associated with

88   each of the three PSC. Some of them (*Uitz et al*, 2006; *Aiken et al.,* 2009) break-down the *chl-a*

89   abundance into several ranges for each of which a specific relationship is computed. Others (*Brewin*

90   *et al*, 2010; *Hirata et al*, 2011) are based on a continuum of *chl-a* abundance. Studies have also been

91   done to estimate the phytoplankton groups (PFT) by taking into account spectral information

92   (*Sathyendranath et al.,* 2004, *Alvain et al.*, 2005, 2012; *Hirata et al.*, 2011; *Ben Mustapha et al,* 2013;

93   *Farikou et al*, 2015). This is of fundamental interest to the understanding of the phytoplankton behavior

94   and to modeling its evolution.

95   Due to highly non-linear relationship linking the multispectral ocean color measurements with the

96   pigment concentrations, we proposed a neural network clustering algorithm (2S-SOM) able to deal

97   with multi variables linked by complex relationships. The 2S-SOM algorithm is well adapted to this

98   complex task by weighting the different inputs. The clustering algorithm was calibrated on a restricted

99   database composed of remote sensed observations co-located with measurements taken in the global

100   ocean.

In the present paper, we propose the retrieval of the major pigment concentrations from satellite ocean color multi-spectral sensors in the Senegalo-Mauritanian upwelling, which is an oceanic region off the coast of West Africa where a strong seasonal upwelling occurs (Figure 1).

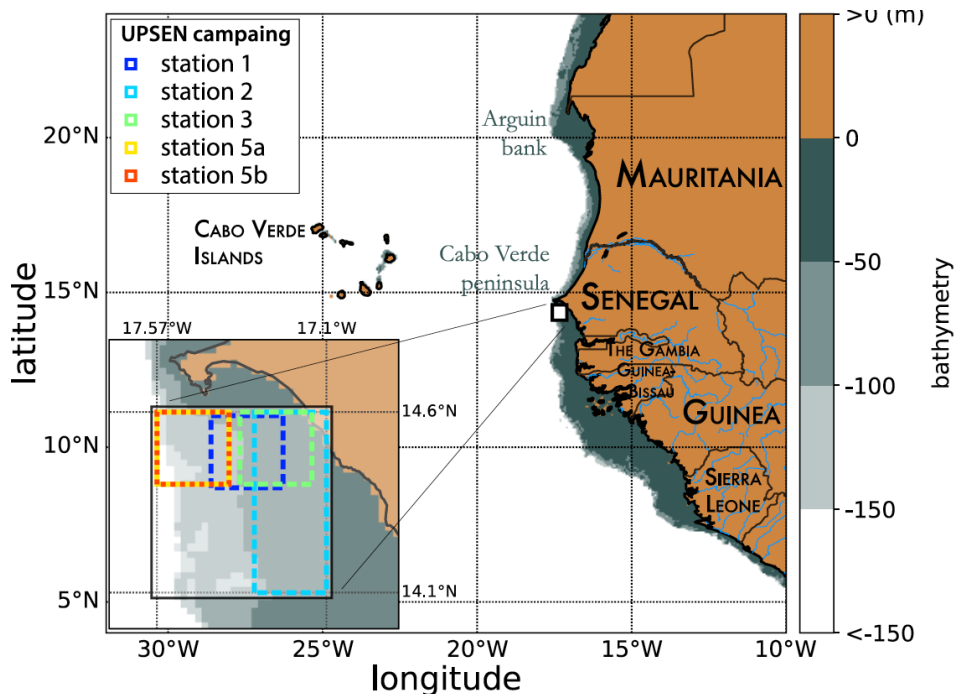



Figure 1: *Mauritania and Senegal coastal topography. The land is in brown and the ocean depth is represented in meters by the color scale on the right side of the figure. The UPSEN stations are shown at the bottom left cartoon of the figure.*

The Senegalo-Mauritanian upwelling is one of the most productive eastern boundary upwelling systems (EBUS) with strong economic impacts on fisheries in Senegal and Mauritania. Since the region has been poorly surveyed in situ, we have chosen to extract pertinent biological information from ocean-color satellite measurements. The region has been intensively studied by analysis of SeaWiFS ocean-color data and AVHRR sea-surface temperature as reported in *Demarcq* and *Faure* (2002), *Sawadogo et al.* (2009), *Farikou et al.* (2013, 2015), *Ndoye et al,* (2014) and more recently by *Capet et al,* (2017) with in situ observations.

The paper is articulated as follows: in section 2, we present the data we used (in situ and remote sensing observations). The mathematical aspect of the clustering method (2S-SOM) is detailed in section 3. In section 4 we present the methodological results. The spatio-temporal variability of the fucoxanthin and chl-a concentration in the Senegalo-Mauritanian upwelling region are presented in section 5, as well as the results of the oceanic UPSEN campaigns. In section 6 we discuss the results and the method. A conclusion is presented in section 7.

124 **2- MATERIALS**

125

126 In this study we used three distinct datasets: the first was used to calibrate the method, the second to

127 conduct a climatological analysis of the Senegalo-Mauritanian upwelling region and the third was

128 obtained during the oceanographic UPSEN campaign. These datasets are composed of satellite remote

129 sensing observations and in-situ measurements.

130

131 *2.1 The calibration data base (DPIG)*

132 The calibration database (DPIG) comprises in situ pigment measurements co-located with satellite

133 ocean-color observations done by the SeaWiFS (Sea-viewing, Wide-Field-of-view Sensor).

134 This DPIG is composed of 515 matched satellite observations and in situ measurements made in the

135 global ocean (mainly in the North Atlantic and the equatorial ocean; *Ben Mustapha et al.*, 2014). The

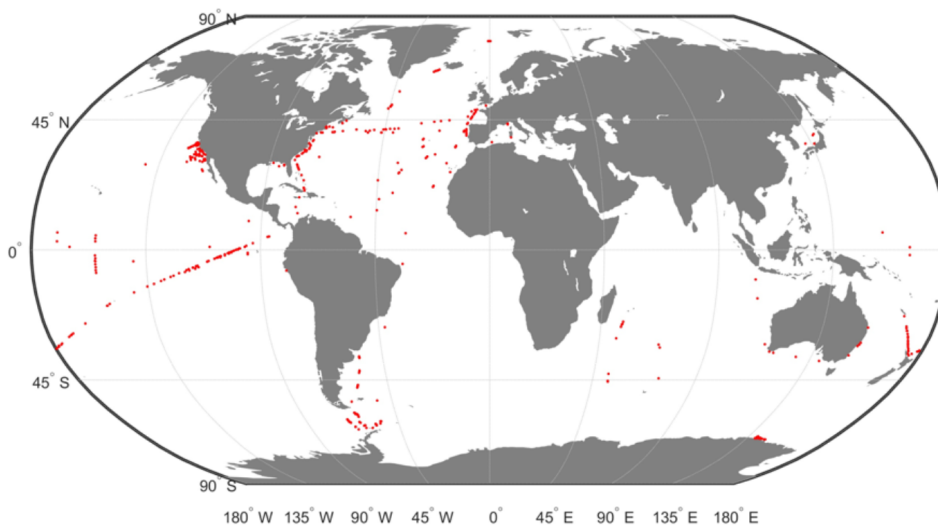
136 match-up criteria were quite severe: we used satellite pixel situated at a distance less than 20km from

137 the in situ measurement in a time window of +/- 12h. The geographic distribution of the 515 coincident

138 in situ and satellite measurements is shown in Fig. 2. Matchup procedure between in situ and satellite

139 observation is a crucial question to estimate remote sensing algorithms. If the parameters of the

140 procedure are too severe, the number of collocated data is

141

142

143

144 Figure 2: *Geographic positions of the 515 in situ and satellite collocated measurements of the*
145 *DPIG database.*

146

147 dramatically decreasing. If the parameters are too large, it is the accuracy of the matching, which is

148 decreasing. We accordingly chose some compromise. Usually people use a matchup window of 3X3 pixels

149 (*Alvain et al,* 2005) which corresponds to a distance less than 20km between the satellite pixel and in

150 situ measurement, since we deal with level 3 satellite observations whose pixel is of the order of 9X9km.

151 This criterium refers to the typical length of ocean variability (*Levy et al*, 2012; *Levy,* 2003)

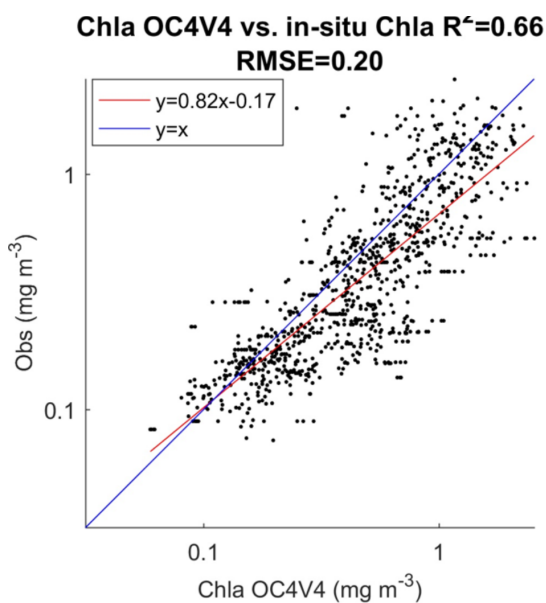
152

153 In Figure 3 we present the R^2 coefficient between the in situ *chl-a* a and the SeaWiFS *chl-a* a computed

154 by using the OC4V4 algorithm (*O'Reilly et al,* 2001) for the DPIG collocated observations. We remark

155 that the two measurements are in good agreement at global scale. Each data of DPIG is a vector

156

157

158

159

160 Figure 3: *Dispersion diagram of DPIG chl-a computed from the SeaWiFS observations using the*

161 *OC4V4 algorithm versus in situ chl-a. The coefficient of vraisemblance R^2 and the RMSE (Root Mean*

162 *Square Error) were computed in in mg m^{-3}*

163

164 having 17 components (five ocean reflectance ($\rho_w(\lambda)$ and *Ra(λ)* at five wavelengths (412, 443, 490,

165 510 and 555nm)*,* SeaWiFS *chl-a,* five in situ pigment ratios and in situ *chl-a* concentration). The in

166 situ *chl-a* a concentration ranges between 0.007 and   3. mg m^{-3} (see Table 1).

167 The five *Ra(λ)* are defined following *Alvain et al,* (2012 :

168 $\qquad$ *Ra(λ) = $\rho_W(\lambda)$/ ρ_{Wref}(λ, chl-a)* $\qquad\qquad\qquad$ (2)

169 where the parameter $\rho_{wref}(\lambda, chl_a)$ is an average reflectance depending on the *chl-a* concentration only

170 which was computed according to the procedure reported in *Farikou et al*, 2015. *Ra(λ)* is a non-

171   dimensional parameter which is independent of the *chl-a* abundance and sensitive to the secondary

172   pigments only (*Alvain et al* , 2012).

173

174   The DPIG database thus provides information on the existing links between the pigment composition

175   and the SeaWiFS measurements. The pigment composition are defined by the pigment ratios which

176   are non-dimensional variables of the form in the present study:

177   Pigment Ratio=DP/T*chl-a*   (3)

178   which is defined as the ratio of the diagnostic pigment (DP) versus the total *chl-a*

179   (T*chl-a = chl-a* +divinyl *chl-a*, according to *Alvain et al.*, 2005).

180

181   The pigments of the DPIG and their statistical characteristics are given in Table 1. The statistical tests

182   presented in Figure 3 (R^2 and RMSE) and in Table 1 (MEAN, STD, MIN, MAX) were computed in

183   mg m^{-3}.

184

	Rdiviny a	Rperid	Rfuco	R19HF	Rzeax	Chloro *in situ*
Mean	0.1414	0.0272	0.1248	0.1859	0.1696	0.5292
STD	0.1584	0.0196	0.0971	0.0996	0.2063	0.5720
MIN	0.0037	0.0035	0.0053	0.0066	0.0027	0.007
MAX	0.8889	0.2027	0.8514	0.7654	1.5574	2.9980

185

186

187   Table 1: *Pigments of the DPIG and their statistical characteristics: :STD (Standard Deviation), MIN*
188   *(minimum value), MAX (maximum value).*

189

190   ***2.2 The Senegalo-Mauritanian upwelling satellite data (DSAT)***

191   The satellite dataset we processed to retrieve the pigment concentration consist of five $\rho_w(\lambda)$ and five

192   *Ra(λ)* at five wavelengths (412, 443, 490, 510 and 555nm), and the SeaWiFS *chl-a* concentration

193   observed in the Senegalo-Mauritanian upwelling region (8°N-24°N, 14°W-20°W; Figure 3) during 11

194   years (1998-2009) by SeaWiFS. This data set is here below denoted *DSAT*.

195   The satellite observations ($\rho_w(\lambda)$ and *chl-a* concentration) were provided by NASA with a resolution

196   of nine kilometers. Due to the presence of Saharan dusts in this region, very few estimations of satellite

197    $\rho_w(\lambda)$ and in situ *chl-a* were available, and some satellite estimations of *chl-a* could present strong over-

198   estimations (*Gregg et al*, 2004). For this reason, we reprocessed the $\rho_w(\lambda)$ and *chl-a* data with an

199   atmospheric correction algorithm developed specifically for Saharan dust (*Diouf et al,* 2013,

200   http://poac.locean-ipsl.upmc.fr) in order to improve the satellite observations.

201

2.3 The UPSEN database

Recently, some HPLC measurements were made in the Senegalo-Mauritanian region during two oceanographic cruises (UPSEN campaigns) of the oceanographic ship "Le Suroit" from 7 to 17 March 2012 and from 5 to 26 February 2013 as reported in *Ndoye et al,* (2014); *Capet et al,* (2017). The goal was to study the dynamics and the biological variability of the Senegalo-Mauritanian upwelling. During these campaigns, in-situ HPLC measurements were carried out. We expected to be able to co-locate them with the ocean-color VIIRS (Visible Infra-red Imaging Radiometer Suite) sensor observations whose wavelengths are close to those of the SeaWiFS. Unfortunately, we were only able to process satellite observations made on 21 February 2013 due to the presence of clouds and Saharan aerosols the other days. We processed the satellite observations provided by the VIIRS sensor at four wavelengths (443, 490, 510, 555 nm) for pixels in the vicinity of the ship stations (within a distance of 20km) and observed in a time window of +/- 12h, and for which the satellite *chl-a* was less than 3 mg m^{-3}, which is the limit of validity of our method imposed by the range of *chl-a* observed in DGIP (mean of 0.52 mg m^{-3}). Only five stations off Cabo Verde peninsula fitted these requirements (see Figure 1 for their positions).

3 - THE PROPOSED METHOD (2S-SOM)

Classification methods were applied for retrieving geophysical parameters from large databases in several studies including weather forecasting (*Lorenz*, 1969; *Kruizinga and Murphy*, 1983), short-term climate prediction (*Van den Dool*, 1994), downscaling (*Zorita and von Storch*, 1999), reconstruction of oceanic pCO$_2$ (*Friedrichs and Oschlies.,* 2009), and of *chl-a* concentration under clouds (*Jouini et al*, 2013). In the present study, we used a new neural network classifier, which is an extension of the SOM algorithms.

3-1 The SOM clustering

The SOM algorithms (*Kohonen,* 2001) constitute powerful nonlinear unsupervised classification methods. They are unsupervised neural classifiers, which have been commonly used to solve environmental problems (*Cavazos,* 1999; *Hewitson et al,* 2002; *Richardson et al,* 2003; *Liu et al,* 2005, 2006; *Niang et al,* 2003, 2006; *Reusch et al,* 2007). The SOM aims at clustering vectors $z_i \in \mathbb{R}^N$ of a multidimensional database $\textbf{\textit{D}}$. Clusters are represented by a fixed network of neurons (the SOM map), each neuron c being associated with the so-called referent vector $w_c \in \mathbb{R}^N$ representing a cluster. The self-organizing maps are defined as an undirected graph, usually a rectangular grid of size $p \times q$. This

232 graph structure is used to define a discrete distance (denoted by δ) between two neurons of the $p \times q$

233 rectangular grid which presents the shortest path between two neurons. Each vector z_i of $\boldsymbol{D}$ is assigned

234 to the neuron whose referent $\boldsymbol{w_c}$ is the closest, in the sense of the Euclidean distance: $\boldsymbol{w_c}$ is called the

235 projection of the vector z_i on the map. A fundamental property of a SOM is the topological ordering

236 provided at the end of the clustering phase: close neurons on the map represent data that are close in

237 the data space. The estimation of the referent vectors $\boldsymbol{w}_c$ of a SOM and the topological order is achieved

238 through a minimization process in which the referent vectors $\boldsymbol{w}$ are estimated from a learning data set

239 (The DPIG data base in the present case). The cost function is shown in Annex:

240 The SOMs have frequently been used in the context of completing missing data (*Jouini et al*, 2013),

241 so the projected vectors z_i may have missing components. Under these conditions, the distance between

242 a vector $z_i \in \boldsymbol{D}$ and the referent vectors $\boldsymbol{w_c}$ of the map is the Euclidean distance that considers only the

243 existing components (the Truncated Distance or *TD* hereinafter).

244

245 ***3-2 The 2S-SOM Classifier***

246 In the present case, we used the 2S-SOM algorithm, which is a modified version of the SOM, very

247 powerful in the case of a large number of variables. It automatically structures the variables having

248 some common characters into conceptually meaningful and homogeneous blocks. The 2S-SOM takes

249 advantage of this structuration of $\boldsymbol{D}$ and the variables into different blocks, which permits an automatic

250 weighting of the influence of each block and consequently of each variable. The block weighting

251 facilitates the clustering procedure by considering the most pertinent variables. The vectors of DPIG

252 defined in section 2 can be decomposed in four blocks. The essence of this decomposition in blocks is

253 that each of the 17 components of the DPIG vectors gathered information with a different physical

254 influence in the classification phase. The composition of each block is done as follows:

255 ***First Block*** (B1) comprises the five pigment in-situ concentration ratios (divinyl chlorophyll-a,

256 peridinin, fucoxanthin, 19'hexanoyloxyfucoxanthin, zeaxanthin concentration ratios). The pigment

257 ratios are defined in Eq. 3.

258 ***Second Block*** (B2) comprises the water-leaving reflectance $\rho_w(\lambda)$ at the five SeaWiFS wavelengths

259 ***Third Block*** (B3) comprises the five $Ra(\lambda)$,

260 ***Fourth Block*** (B4) comprises two variables: The in situ and the SeaWiFS *chl-a* concentrations.

261

262 The 2S-SOM is able to deal with a large quantity of variables, choosing those that are the most

263 significant for the classification and neutralizing those which are the least significant. This is done by

264   estimating weights on the blocks and the variables. We fully describe the 2S-SOM algorithm in Annex.

265   In the following we use a simplified version of 2S-SOM in which only the blocks are weighted.

266

267   ### *3.3 The calibration phase*

268   Similarly to the standard SOM, the 2S-SOM is determined through a learning phase by using a more

269   complex cost function (see Annex) that estimate for each neuron, in addition to the referent vector, a

270   weight (α) for each block. For a neuron c, we define the weights α_{cb} of each block b ($b = 1....4$). .

271   At the end of the calibration phase, each element z_i of the dataset DPIG is associated with a referent

272   w_c whose components are partitioned into four blocks. In the present study, the 2S-SOM map is

273   represented by a two-dimensional (9x18=162) grid that represents the partition of the DPIG dataset

274   into different classes. Each class provided by the 2S-SOM is associated with a so-called referent vector

275   w_c with $c \in \{1.....162\}$. The size of the map has been determined by using the procedure provided by

276   the SOM software available at : http://www.cis.hut.fi/projects/somtoolbox/download/.

277

278   ### *3.4 The Pigment retrieval*

279   In the second phase, which is an operating phase, we estimated the pigment concentration ratios of a

280   pixel PX_m from its satellite ocean-color sensor observations only. The 11 ocean color satellite

281   observations (5 $\rho_w(\lambda)$, 5 $Ra(\lambda)$, and *chl-a*) of pixel PX_m were projected onto the 2S-SOM using the

282   Truncated Euclidian Distance (section 3.1). We select the neuron c associated with a referent vector

283   whose the 11 ocean-color parameters are the closest to those observed by the satellite sensor. The

284   pigment ratios of PX_m are those associated with the neuron c. At the end of the assignment phase, each

285   pixel PX_m of a satellite image is associated with a referent vector w_c, which has 6 pigment

286   concentration ratios among its 17 components. The flowcharts of the method (2S-SOM learning and

287   pigment retrieval) are presented in Figure 4.

288

289

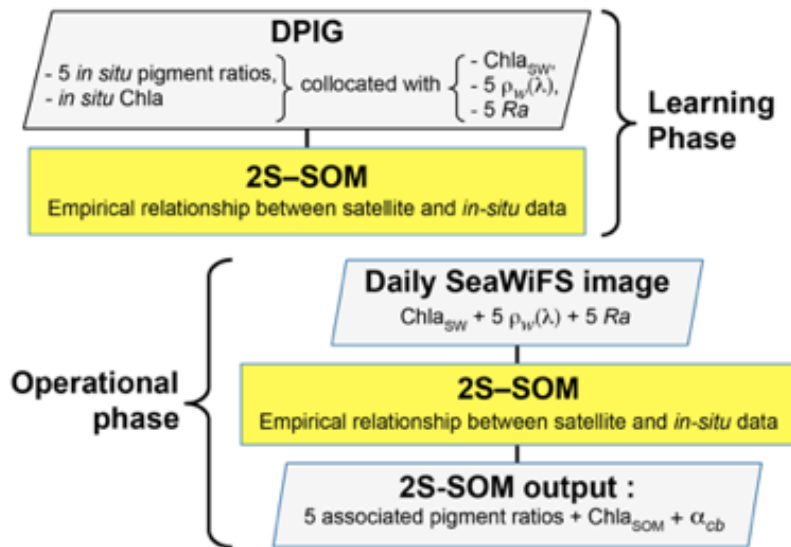



Figure 4: *Flowchart of the method: top panel - Learning phase; bottom panel – operational phase which consists in pigment retrieval and the determination of the α_{cb} block parameters.*

4 - METHODOLOGICAL RESULTS

4-1 Statistical validation of the method

The validation of the method was focused on the retrieval of the fucoxanthin ratio, which is a characteristic of diatoms, but the same procedure could be applied to any pigment. The hyper-parameter μ (see Annex) was optimized in order to retrieve that ratio, while η was set constant since only the block were weighted in the present study. Due to the small amount of data in the DPIG, we estimated the accuracy of the fucoxanthin retrieval by a cross-validation procedure, which is a powerful procedure in statistics. The principle is the following: we learned 30 2S-SOM using 30 different learning datasets L_i constituted of 90% of DPIG taken at random, and then computed statistical estimator on the retrieved quantities using 30 test datasets (10% of DPIG). The algorithm was as follows:

i=1 …. 30

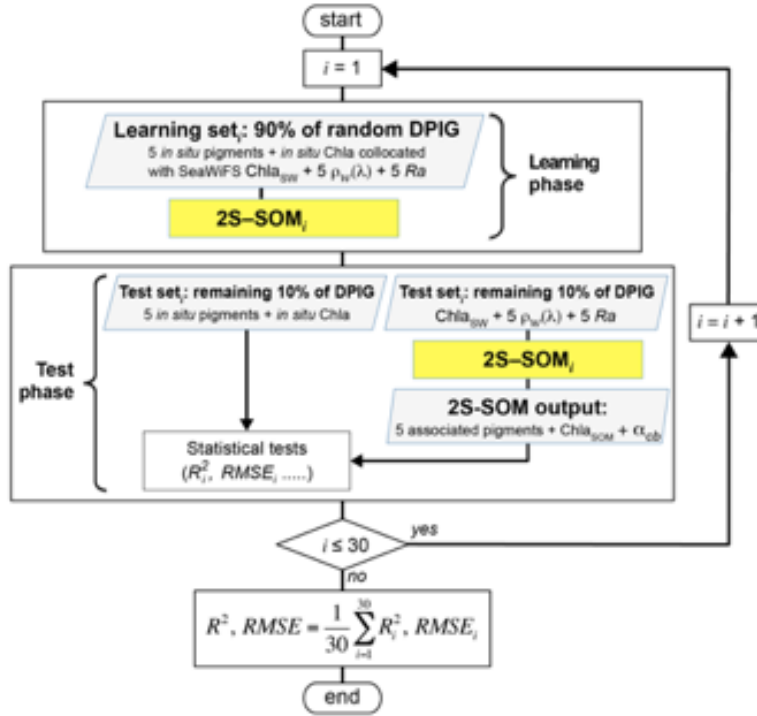
1. determination at random of a learning dataset L_i (90% of DPIG) and a test dataset TL_i (10% of DPIG)
2. training of a 2S-SOM map M_i using L_i (see section 3.2 and 3.3).
3. Validation using TL_i according to the procedure described in section 3.4
4. Estimation of the $RMSE_i$ and $R^2{}_i$ on TL_i between the estimated and observed fucoxanthin ratios

end

314   Computation of the mean RMSE and R^2 $\left(R^2, \text{RMSE} = \frac{1}{30} \sum_{i=1}^{I=30} R^2 i, RMSEi\right)$

315

316   The flowchart of the cross-validation procedure is presented in Figure 5.

317

318
319

320   Figure 5: *Flowchart of the cross-validation procedure for 30 partitions of the DPIG database.*

321

322   Statistical parameters (R^2 coefficients, RMSE and P-values) of the cross validation between the DPIG

323   in situ pigments and the pigments given by the 2S-SOM averaged for the 30 2S-SOM realizations,

324   which are presented in table 2, show the good performance of the method.

325

326

	R^2	RMSE [MG M^{-3}]	PVAL
CHLA SOM	0.84	0.22	0.001
DVCHLA	0.60	0.02	0.001
FUCO	0.87	0.02	0.001
PERID	0.81	0.01	0.001

327

328   Table 2: *Statistical parameters (R^2 coefficients, RMSE and P-values) of the cross validation between*
329   *the DPIG in situ pigments and the pigments given by the 2S-SOM averaged for the 30 2S-SOM*
330   *realizations*
331

332

4-2 Analysis of the topology of the 2S-SOM

As explained in sections 3-2 and 3-3, the referent vector components ($w_c \in R^{17}$), which are estimated during the learning phase, are partitioned in four blocks B1, B2, B3 and B4. The hyper parameters μ was tuned in order to favor the accuracy of the retrieval of the fucoxanthin ratio. We recall that all the pigment ratios are estimated during the calibration phase, but in the present paper attention was focused on the fucoxanthin ratio when selecting the parameter μ. In Figure 6, we

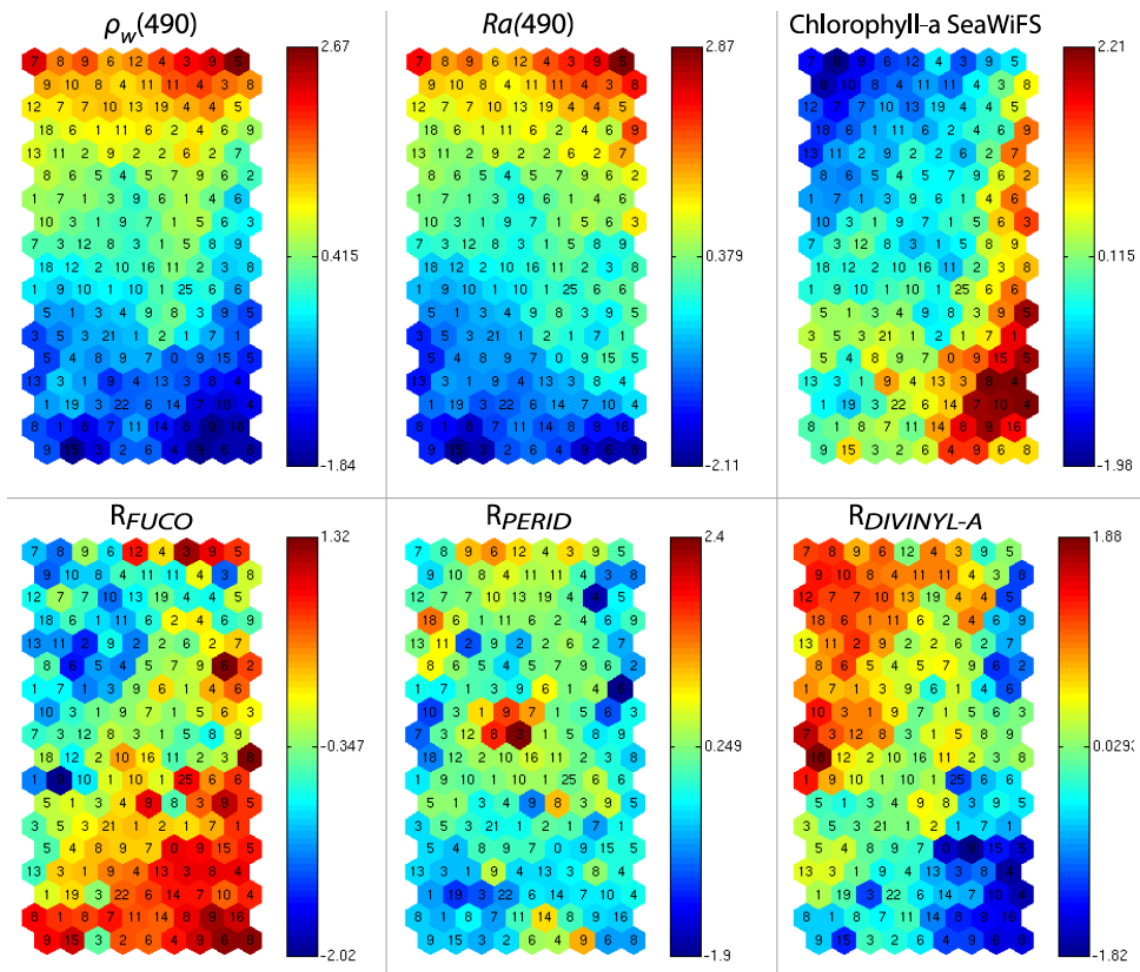



Figure 6: *2S-SOM Map. From left to right and top to bottom, values of the referent vectors for ρ_w(490), Ra(490), SeaWiFS chl-a, and fucoxanthin, peridinin, divinyl Ratios. The number in each neuron indicates the amount of DPIG data captured at the end of the learning phase, the values indicated by the color bars are centered-reduced and non-dimensional values.*

present six of the referent vector components of the 2S-SOM map. These components are ρ_w(490), *Ra(490)*, SeaWiFS *chl-a*, and the ratios of fucoxanthin, which is a specific diatom pigment, and of

349 *peridinin* and *divinyl*. They exhibit a coherent topological order, the components having close values

350 being close together on the topological map. The remaining eleven components (not shown) exhibit

351 the same coherent topological order. One can observe a very good topological order for the fucoxanthin

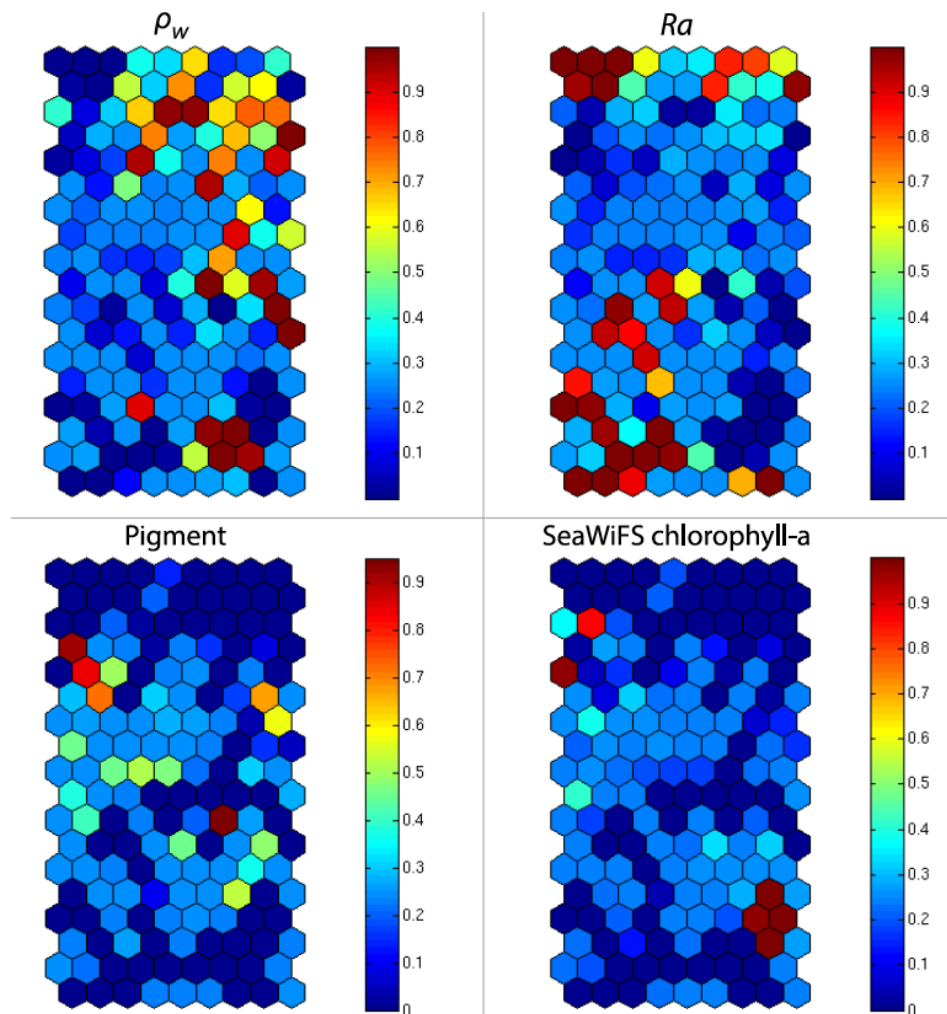
352 ratio that was favored by the determination of the hyperparameter μ. Moreover, the bottom right region

353 in the 2S-SOM map (Figure 6) may correspond to the diatoms with a good confidence since high

354 fucoxanthin is associated with high chlorophyll concentration and low peridinin. This is endorsed in

355 section 5 by looking at the geographical location of the different pigment concentrations (figures 8, 10,

356 11). Another important remark is that the value of each component presents a large range of variation

357

358

359

360 Figure 7: *2S-SOM map. Weights (α_{cb}) of the four block parameters determined at the end of the learning*
361 *phase; from left to right and top to bottom: ρ_w, Ra, Pigment, SeaWifs chl-a. The color bars show the % of*
362 *the weight estimated by 2S-SOM, a value of 1 or 0 indicating that the data in the neuron are assembled with*
363 *respect to that block only.*

364

365 of the same order as the range of variation found in the DPIG variables. It means that the 2S-SOM

366 map has captured most of the variability of the dataset.

367 Figure 6 shows a strong link between the values of the referent vectors for fucoxanthin and *chl-a* (high

368 fucoxanthin and *chl-a* values, at the bottom right of the 2S-SOM) while fucoxanthin is high and *chl-a*

369 low for the referent vectors at the bottom left of the 2S-SOM. Additional information will be provided

370 by the *Ra(490)* values when the fucoxanthin is less closely linked to the chlorophyll.

371 Besides, for each neuron, the 2S-SOM provides a weight for each block (α_{cb}) and each variable (β_{cbj}).

372 For a given neuron *c* the weights (α_{cb}) of the blocks are normalized, their sum being 1. A value of 1

373 for one block (and therefore a value of 0 for the other blocks) indicates that the data in the neuron are

374 gathered with respect to that block only because there is too much noise in the variables in the other

375 blocks. By examining the weights on the map, one can see which block most influences the link

376 between the satellite measurements and the pigment ratios.

377 In Figure 7, we present the α_{cb} values estimated during the learning phase of the 4 blocks (B1, B2, B3,

378 B4). For some neurons, only the blocks related to the reflectance and the reflectance ratio are used for

379 the definition of the neuron, while the weights for the two other blocks (pigments and *chl-a*) are null,

380 indicating that for these neurons, in situ observations and SeaWiFS *chl-a* are more noisy than the

381 reflectance. These neurons correspond to very small *chl-a* concentrations, which are estimated with

382 large error. Besides, we remark that high α values for *chl-a* corresponds to high *chl-a* concentration

383 values (bottom right of the *chl-a* panel in figure 7 and figure 6 respectively). For these cases, the

384 clustering assembled data that mainly depend on *chl-a* concentration.

385

386

387 **5 - GEOPHYSICAL RESULT**

388

389 In the present study, we apply the 2S-SOM (section 3), which explicitly makes a weighted use of the

390 data according to their specificity (ocean-color signals or in situ observations) to retrieve the

391 fucoxanthin concentration from remote sensed data in the Senegalo-Mauritanian upwelling region

392 where in situ measurements are lacking. According to the good results of the cross-validation method

393 as shown in section 4.1, we expect that the 2S-SOM will provide pertinent results in a region which

394 has been poorly surveyed.

395

396

5-1 The pigment estimation from SeaWiFS observations in the Sénégalo-Mauritanian upwelling region

We decoded the DSAT database (section 2-3) using the 2S-SOM for 11 years (1998-2009) of SeaWiFS data observed in the Senegalo-Mauritanian upwelling region (8°N-24°N, 14°W-20°W). This study was done according to the retrieval phase described in section 3.4. For each day, we projected the 11 SeaWiFS observations (5 $\rho_w(\lambda)$, 5 $Ra(\lambda)$ and *chl-a*) of each pixel PX_m on the 2S-SOM. At the end of the assignment phase, each pixel of a satellite image was associated with 6 pigment concentration ratios. The underlying assumption is that the link between the remote sensing information and the pigment ratios of a pixel is this provided by the selected referent w_c. Thanks to the topological order provided by the 2S-SOM, we expect that the best neurons chosen during the retrieval would give accurate concentration ratios. In Figures 8, 10 and 11 we present the fucoxanthin concentration ratio restitution for three different days and the associated SeaWiFS Chlorophyll images (1 and 6 January, and 28 February 2003). Due to the limited size of the DPIG, the range of the ratio learned for the

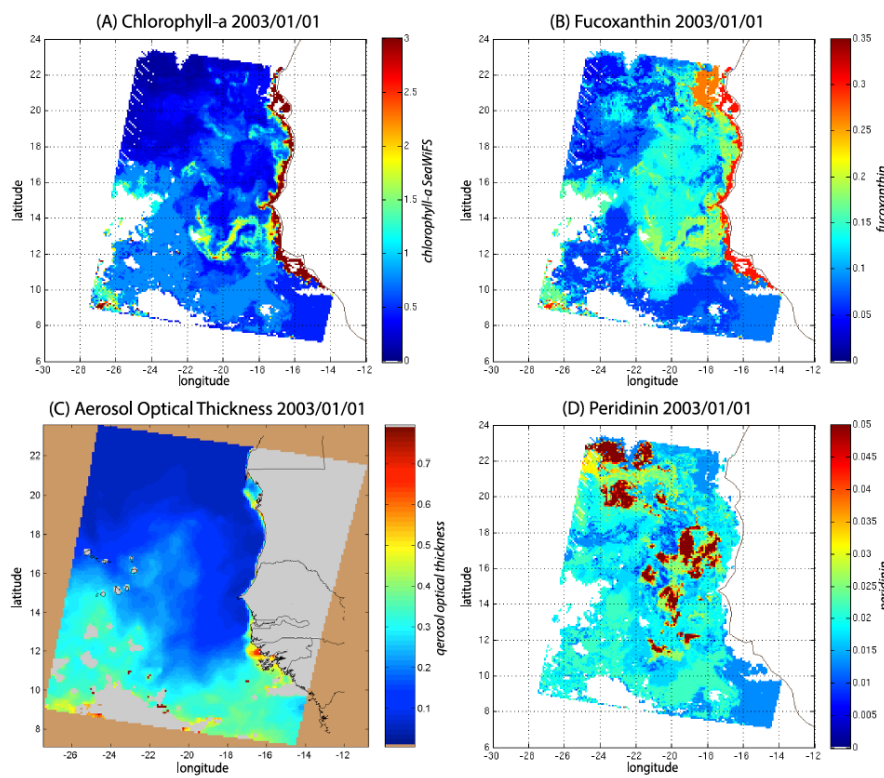
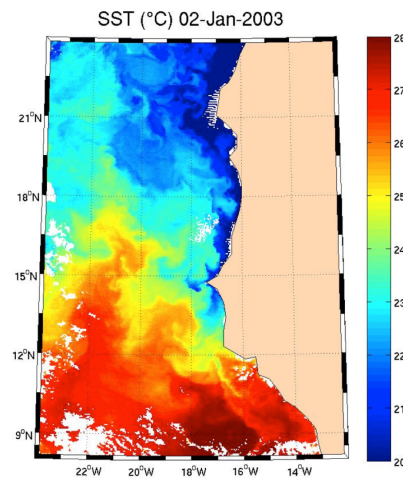



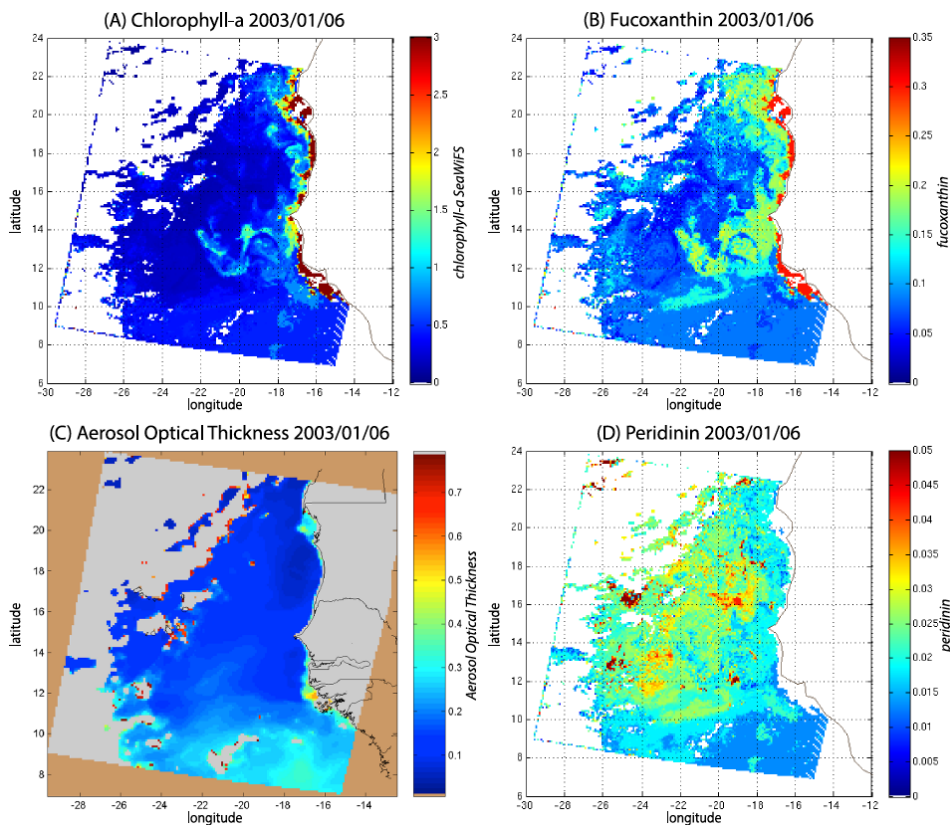
Figure 8: *A) chl-a concentration, (B) fucoxanthin ratio, (C) aerosol optical thickness, (D) peridinin for 1 January 2003. Panels (B) and (D) show that a second-order information was retrieved, which is correlated with the chl-a concentration (A) but not equivalent. The aerosol optical thickness (C) does not seem to contaminate the estimated parameters (fucoxanthin and peridinin ratios).*

418 the fucoxanthin is between 0.3% and 20% with a mean of 10% and the *chl-a* content is between 0.5

419 mg m⁻³ and 3 mg m⁻³. The statistical estimator we used cannot extrapolate what has not been learned,

420

421

422 Figure 9: *SST for 2 January 2003. Note the well-marked upwelling (cold temperature) north of 13°N.*

423

425

426 Figure 10: *(A) chl-a concentration, (B) fucoxanthin ratio, (C) aerosol optical thickness, (D) peridinin for 6*

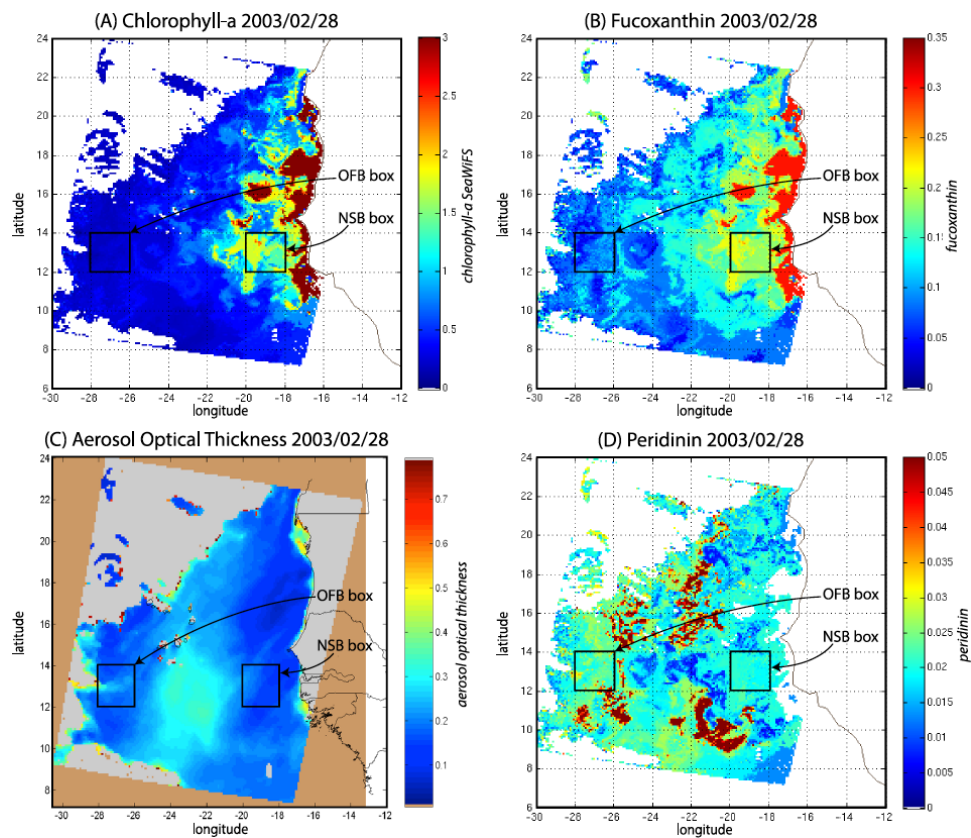
427 *January 2003. Panels (B) and (D) show that a second-order information was retrieved, which is correlated*

428 *with the chl-a concentration (A) but is not equivalent. It is found that the aerosol optical thickness (C) does*

429 *not contaminate the estimated parameters (fucoxanthin and peridinin ratios).*

430   and for that raison we flagged the pixels in the SeaWiFS images that have a *chl-a* concentration greater

431   than 3. mg m^{-3}.

432

433

434

435   Figure 11: *(A) chl-a concentration, (B) fucoxanthin ratio, (C) aerosol optical thickness, (D) Peridinin for*
436   *28 February 2003. Panels (B) and (D) show that a second order information was retrieved, which is*
437   *correlated with the chl-a concentration (A) but is not equivalent. It is found that the aerosol optical*
438   *thickness (C) does not contaminate the estimated parameters (fucoxanthin and peridinin ratios). The*
439   *position of the NSB and OFB boxes are figured out by black square boxes*

440

441   Regarding the images obtained for 1 January 2003 in the Senegalo-Mauritanian region

442   (Fig 8A, B, C, D), we observe that the *chl-a* (Fig 8A) is very high at the coast and decreases offshore

443   in accordance with the upwelling intensity as shown in the SST image (Fig 9). Moreover, we observed

444   a persistent well-marked *chl-a* pattern south of the Cap Vert peninsula in form of a "W", which is the

445   signature of a baroclinic Rossby wave (*Sirven et al*, 2019).

446   Except in the southern part of the region, the AOT (Aerosol Optical Thickness) is low, which means

447   that the atmospheric correction of the reflectance is quite small, which gives confidence in the ocean-

448   color data products. The fucoxanthin concentration is maximum at the coast and decreases offshore as

449   does the *chl-a* concentration, in agreement with the works of *Uitz et al.,* (2006, 2010). Fucoxanthin

450   presents coherent spatial patterns. Peridinin concentration is somewhat complementary to that of

fucoxanthin, with the low fucoxanthin concentration area corresponding to high peridinin concentration area (northern part of Figs 8B, D). This behavior is also observed in Figure 10 (6 January 2003) and in Figure 11 (28 February, 2003) endorsing the analysis shown in Figure 8.

For 28 February, we selected two square box regions (Fig. 11), one near the coast (NSB, long [-20°, -18°], lat [12°,14°]) and the other about 800 km offshore (OFB, long [-28°, -26°], lat [12°,14°]). NSB waters correspond to upwelling waters while OFB waters correspond to oligotrophic waters. We projected the eleven ocean color parameters of the NSB and OFB pixels on the 2S-SOM map.

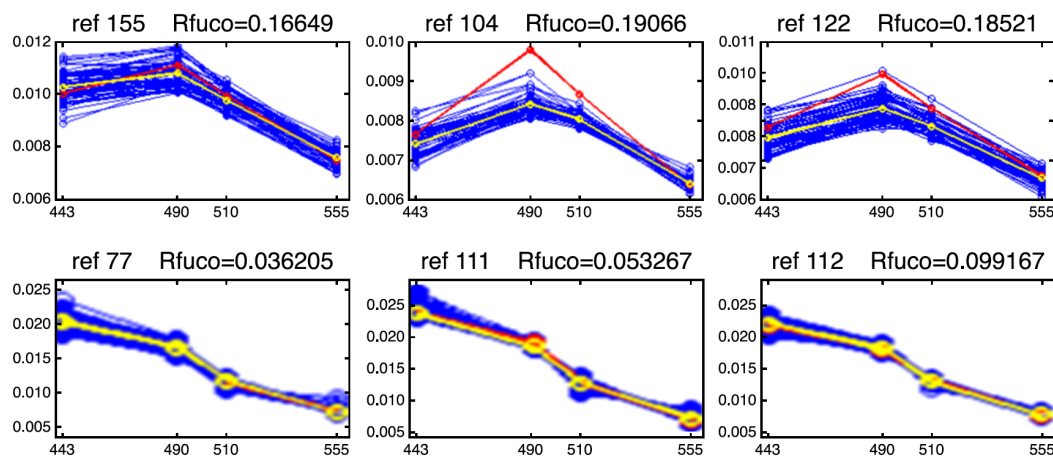



Figure 12: *Reflectance spectra (in blue) captured the 28 February by six neurons whose referent vector spectra are in yellow: top line, for pixels in the NSB region (long. [-20°, -18°], lat. [12°, 14°]); bottom line, for pixels in the OFB region (long. [-28°, -26°], lat. [12°, 14°]).*

Figure 12 presents the reflectance spectra (in blue) captured by three neurons of the 2S-SOM corresponding to pixels located in the NSB region (*top line*) and those captured by three neurons corresponding to pixels located in the OFB region (*bottom line*). The reflectance spectra of the associated referent vectors w are in yellow. The satellite reflectance spectra match the referent vector spectra; moreover the fucoxanthin ratio varies inversely with the mean value of the spectrum: the higher the fucoxanthin ratio, the smaller the mean value of the spectrum. The pigment concentration is greater near the coast.

We note a strong difference between the shape and the intensity of the near-shore (NSB) and offshore (OFB) spectra. The OFB spectra present mean values higher than those of the NSB spectra. This is due to the fact that NSB spectra were observed in a region where diatoms are abundant, as shown by

the high value of fucoxanthin concentration in this region (Figs 8, 10, and 11), which is a proxy for diatoms along with higher *chl-a* concentration. In Figure 12, we note the lower values of the coastal spectra at 443 nm, which can be interpreted as a predominant effect of spectral absorption by phytoplankton pigments and CDOM. The different spectra are close together in the OFB region and more disperse in the NSB region. This can be explained by the fact that the OFB region corresponds to Case-1 waters while the NSB region waters are close to Case-2 waters and are influenced by the variability of near shore process like turbidity or presence of dissolved matters, and dynamical instabilities.

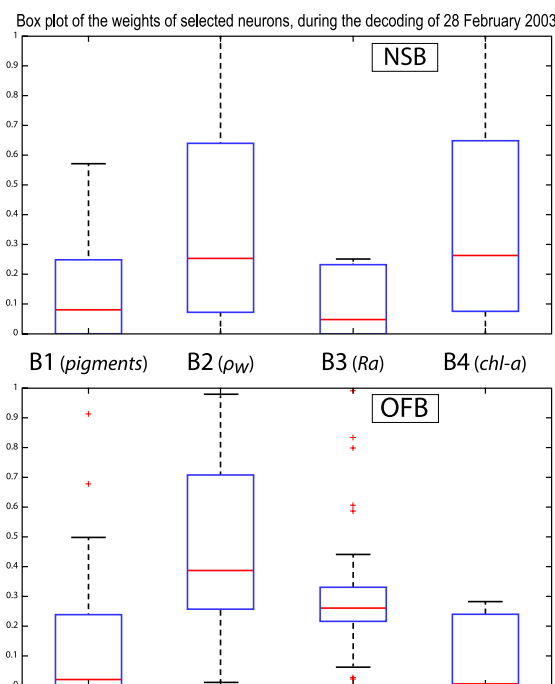



Figure 13: *Box plot of the weights of the selected neurons during the decoding of the 28 February data. From left to right, weights of blocks B1, B2, B3, B4. Top panel, in the NSB region (long. [-20°, -18°], lat. [12°, 14°]); bottom panel, in the OFB region (long. [-28°, -26°], lat. [12°, 14°]).*

We analyzed the weights of the blocks for the neurons selected in the analysis of the costal (NSB) and offshore (OFB) boxes. Figure 13 presents the box plot of the weight α_{cb} corresponding to the neurons belonging to the four blocks (B1, B2, B3, B4), with the constrain that the sum of the weights of a neuron is 1; a weight α larger than 0.25 indicates the predominance of a block in the learning for the classification (see section 3.5). It is clear that the weights for pixels near the coast (Fig 13, top panel) are different from those for offshore pixels (Fig. 13, bottom panel). As already mentioned in section 4.3 and also shown in Figure 7, the weights of the 2S-SOM play a significant role in the 2S-SOM

501 topology and consequently in the pigment retrieval. The weights of blocks B1 and B4 that take into

502 account the influence of the pigment ratios and the chlorophyll content in the retrieval are very low for

503 the offshore (OFB) oligotrophic region and more important for the coastal (NSB) region. The weights

504 of the blocks B2 and B3, which take into account the influence of the reflectance ($\rho_w(\lambda)$, $Ra(\lambda)$),

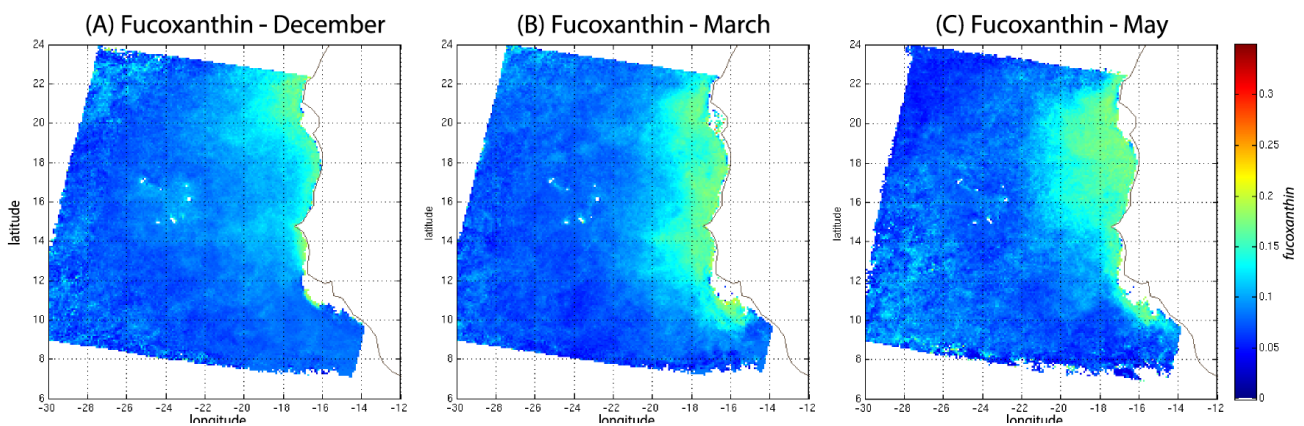
505 dominate for the offshore regions. In coastal waters, the weights of all the blocks are used, with a

506 smaller influence of B3, which is associated with R_a. This gives information on the role played by the

507 different variables on the classification in waters having different phytoplankton concentration and

508 composition. Besides it shows the automatic adaptation of the 2S-SOM to the environment in order to

509 optimize the clustering efficiency with respect to a classical SOM.

510

511

512 Figure 14: *Monthly fucoxanthin concentration averaged for an 11- years (1998-2009) for December*
513 *(A), March (B) and May (C).*

514

515 In order to study the seasonal variability of the fucoxanthin concentration with some statistical

516 confidence in the Senegalo-Mauritanian upwelling region, we constructed a monthly climatology for

517 an 11-year period (1998–2009) of the SeaWiFS observations by summing the daily pixels of the month

518 under study. The resulting climatology is presented in Figure 14 for December (Fig. 14a), March (Fig.

519 14b), and May (Fig 14c), which correspond to the most productive period (Fig. 14c). The fucoxanthin

520 concentration, and consequently the associated diatoms, presents a well-marked seasonality.

521 Fucoxanthin starts to develop in December North of 19°N, presents its maximum intensity in March

522 when the upwelling intensity is maximum, extends up to the coast of Guinea (12°N) in April and

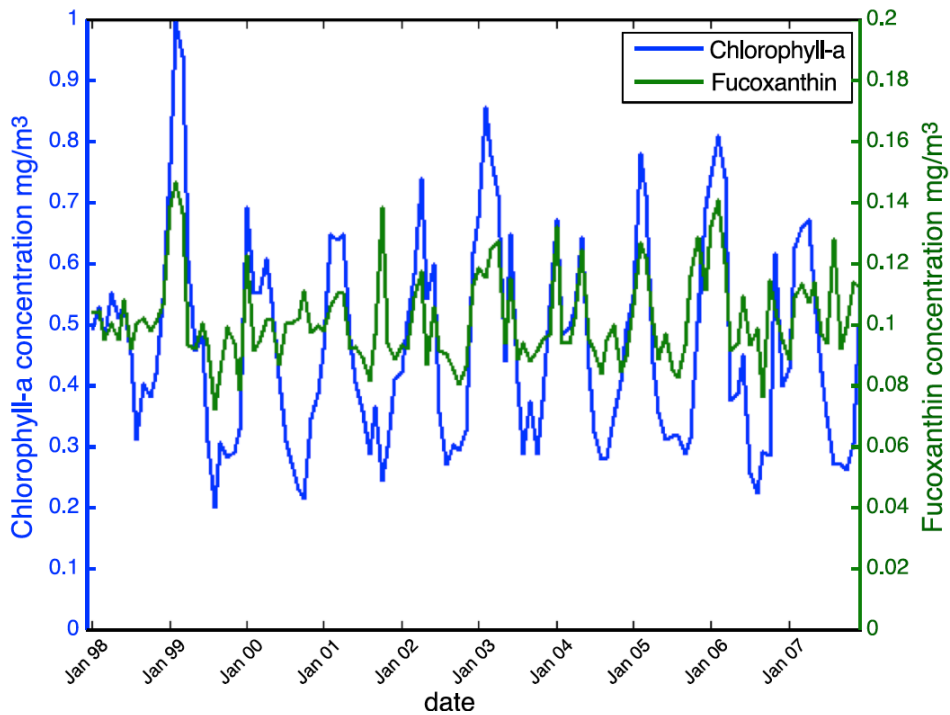
523 begins to decrease in May where it is observed north of Cabo Verde peninsula (15°N) in agreement

524 with the observations reported by *Farikou et al,* (2015) and *Demarcq and Faure,* (2000).

525 Figure 15 shows the fucoxanthin (in green) and the *chl-a* (in blue) concentrations computed from

526 satellite observations for an 11-year period of SeaWiFS observations in the NSB region. There is a

527 good correlation in phase between these two variables but not in amplitude (a good coincidence of

528 peak occurrence but weak correlation in peak amplitude) showing that the relationship between
529

530
531

532 Figure 15: . *chl-a (in blue) and fucoxanthin (in green) concentrations for near-shore pixels (in the NSB*
533 *region).*

534

535 fucoxanthin and *chl-a* is complex as mentioned by *Uitz et al*, (2006). In particular, there is a weak peak
536 in fucoxanthin in October 2001, which is not correlated with a *chl-a* peak.

537

5-2 Analysis of the UPSEN campaigns

539 Figure 16 shows, for every UPSEN stations 1, 2, 3, 5a and 5b (see figure 1 for their geographical
540 position), the averaged in-situ UPSEN spectrum (in blue), the referent spectrum (in red) of the 2S-
541 SOM neuron captured by the collocated satellite VIIRS sensor observations. The referent spectrum
542 is the mean of the different spectra captured by that neuron during the learning phase. Among
543 these different spectra, there is one (black curve in figure 16) which is the closest to the UPSEN
544 spectrum. Obviously, the black curve is closer to the blue curve than the red one which is flatten
545 due to the averaging process. These three spectra are close together showing the good functioning
546 of the 2S-SOM.

547

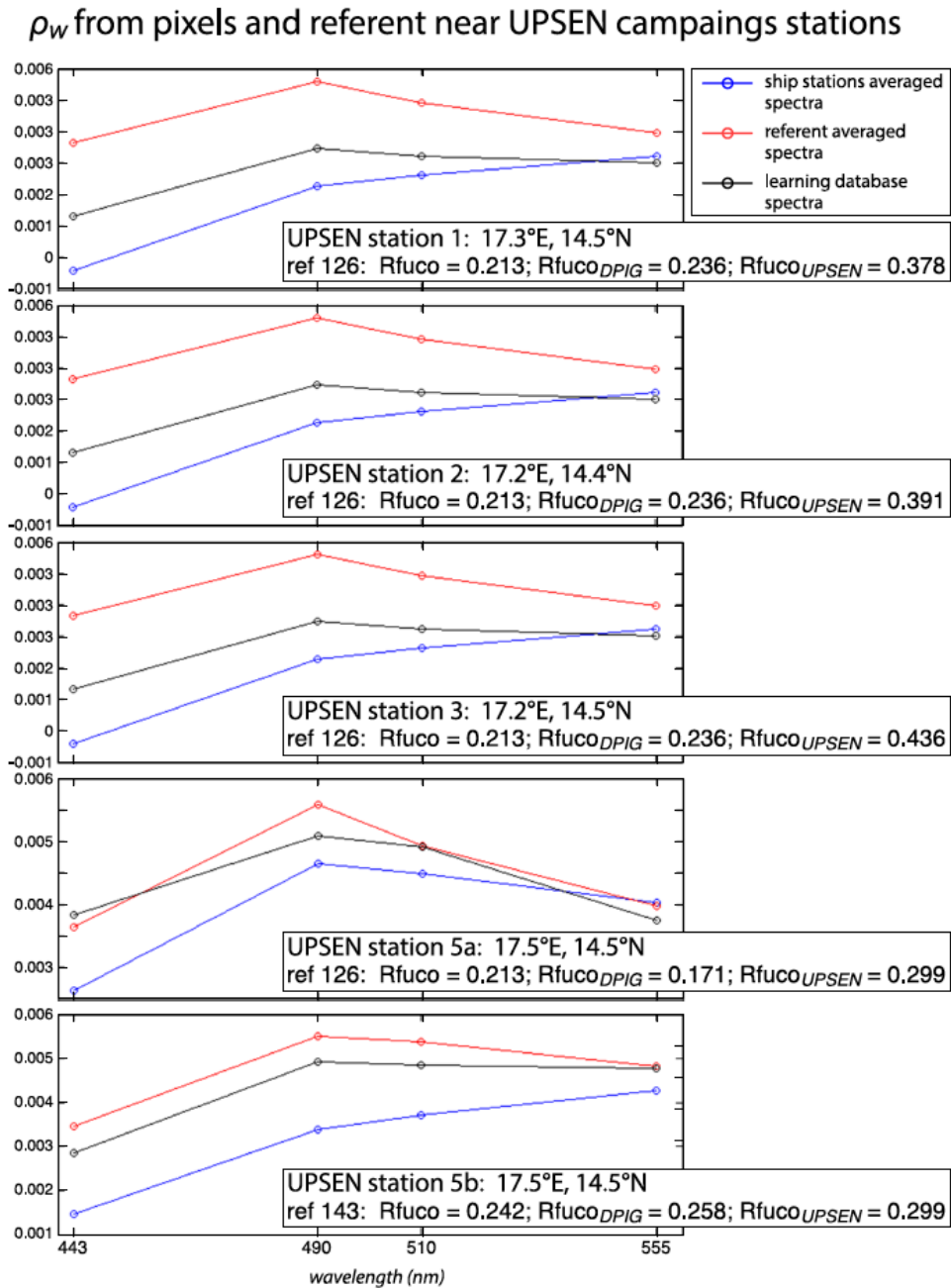



Figure 16: *For ship stations 1, 2, 3, 5a and 5b, we show the averaged spectrum of the in situ spectra of the UPSEN station in blue; the spectrum of the referent vector (in red) of the 2S-SOM neuron, which has captured the closest satellite observations to the UPSEN station; among the different spectra constituting the referent spectrum, the spectrum of the learning database (DGIP) that is the closest to the averaged satellite spectra is shown in black. In the rectangular cartoons, we show the position of the UPSEN station, the number of the neuron of the 2S-SOM which has captured the satellite observation, the Rfuco of the referent vector, the $Rfuco_{DGIP}$ of the closest DGIP and the in situ $Rfuco_{UPSEN}$*

561 Their shapes are close to these observed in the NSB region (Figure 12) but their intensity is lower
562 meaning that their waters are more absorbing than the NSB waters due to a higher pigment
563 concentration. In fact, the UPSEN stations were located close to the coast (figure 1) in the Hann bight
564 south off the Cap Verde peninsula, which is very rich in phytoplankton pigments. In table 3, we present
565 the fucoxanthin ratios associated with the referent vectors (Rfuco$_{2S\text{-}SOM}$), the closest DPIG fucoxanthin-
566 ratios captured by the neuron of the referents and the fucoxanthin-ratios measured during the UPSEN
567 campaign. We note that the fucoxanthin ratios of the in-situ measurements are in the range of the DPIG
568 (see table 1), which allows a good functioning of the 2S-SOM estimator. The pigment ratios obtained
569 from ocean-color observations through the 2S-SOM are close to pigment concentrations measured at
570 the ship stations, which confirms the validity of the method we have developed. We remark that the
571 best 2S-SOM estimate of fucoxanthin ratio with respect to the UPSEN in-situ measurement is given
572 at station 5b which is the farthest off the coast. These results endorse the climatological study of the
573 Senegalo-Mauritanian upwelling region we have done with the 2S-SOM (section 5.1).

574
575

UPSEN STATION	REFERENT N°	RFUCO 2S-SOM	RFUCO DPIG	RFUCO UPSEN
STAT 1 17.3 E 14.5 N	126	0.213	0.236	0.378
STAT 2 17.2 E 14.4 N	126	0.213	0.236	0.391
STAT 3 17.2 E 14.5 N	126	0.213	0.236	0.436
STAT 5A 17.5 E 14.5 N	126	0.213	0.171	0.299
STAT 5B 17.5 E 14.5 N	143	0.242	0.258	0.295

576
577

578 Table 3: *For ship stations 1, 2, 3, 5a and 5b of the UPSEN campaigns, we show the referent captured*
579 *by the VIIRS observations, the fucoxanthin-ratio associated with this referent (Rfuco-2S-SOM), the*
580 *fucoxanthin-ratio of the closest DPIG fucoxanthin-ratio captured by the neuron of the referent and the*
581 *fucoxanthin-ratio measured in situ during the UPSEN campaign*
582
583

584 The 2S-SOM method gives pigment concentrations that are close to those obtained by in situ
585 observations. The method could be applied to a large variety of other parameters in the context of
586 studying and managing the planet Earth. The major constraint to obtaining accurate results is to deal
587 with a learning data set that statistically reflects all the situations encountered in the observations
588 processed. Due to its construction, the method cannot be used to find values beyond the range of the
589 learning data set.

590

591

6 - DISCUSSION

593

Machine learning methods are powerful methods to invert satellite signals as soon as we have adequate database to support the calibration. Several technics have been used for retrieving biological information from ocean color satellite observations. First, studies employed multilayer perceptrons (MLP), which are a class of neural networks suitable to model transfer function (*Thiria* et al, 1993). *Gross* et al, (2000, 2004) retrieved *chl-a* concentration from SeaWiFS, *Bricaud* et al, (2006) modeled the absorption spectrum with MLP, *Raitsos* et al, 2008 and *Palacz* et al, 2013 introduced additional environmental variables in their MLPs such as SST in the retrieval of PSC/PFT from SeaWiFS, which improved the skill of the inversion. Another suitable procedure was to embed NN in a variational inversion, which is a very efficient way when a direct model exists (*Jamet* et al, 2005; *Brajard* et al, 2006a,b; *Badran* et al, 2008). Statistical analysis of absorption spectra of phytoplankton and of pigment concentrations were conducted by *Chazottes* et al, (2006, 2007), by using a SOM.

In the present study, due to the fact that the learning dataset was quite small (515 elements), we used an unsupervised neural network classification method, which is an extension of the SOM method well adapted to dealing with a small database whose elements are very inhomogeneous. We clustered available satellite ocean-color reflectance at five wavelengths and their derived products, such as chlorophyll concentration, and the associated in situ pigment ratios.

The major points of this study are as follows:

- The clustering was carried out by developing a new neural classifier, the so-called 2S-SOM, which presents several advantages with respect to the classical SOM. As in the SOM, we defined clusters that assemble vectors, which are close together in terms of a specified distance. This classifier was learned from a worldwide database (DPIG) whose vectors are ocean-color parameters observed by satellite multi-spectral sensors and associated pigment concentrations measured in situ. In the operational phase, SeaWiFS images are decoded, allowing the estimation of the pigment concentration ratios. The major advantage of 2S-SOM with respect to the classical SOM is to cluster variables having similar physical significance in blocks having specific weights. The weights attributed to the four blocks are computed during the learning phase and vary with the quality of the variables and with respect to their location on the ocean (near the coast or offshore). This permits to modulate the variable influence in the cost function, which makes the clustering more informative than that provided by the SOM. The block decomposition provides useful scientific information. For offshore, the weight analysis allowed us to show that more influence is given to the reflectance ratios $Ra(\lambda)$ and less to the *chl-a* and pigment concentrations; on the contrary near the coast the weights

625   indicate a more active use of the pigment composition and the *chl-a* concentration. Therefore, the

626   resulting 2S-SOM clustering therefore at best takes into account the information that belongs to the

627   specific water content.

628   - The 2S-SOM decomposes the DPIG into a large number of significant ocean-color classes allowing

629   reproduction of the different possible situations encountered in the dataset we analyze. Besides, we

630   assume that the relationship between the pigment concentration and the remote sensed ocean-color

631   observations is independent on the location, which is justifiable since the relationship depends on the

632   optical properties of ocean waters through well-defined physical laws which are region-independent.

633   This also endorses the fact that we used a global database to retrieve pigments in a definite region.

634   On the contrary, the different phytoplankton species vary from one region to another making the

635   relationship between pigment ratio and phytoplankton species strongly depending on the region. This

636   justifies the fact we focused our study on the pigment retrieval rather than on the PSC or PFT, as

637   mentioned above. Moreover, most of the recent phytoplankton in situ identifications have been made

638   using pigment measurements with the HPLC method (*Hirata et al*, 2011). It is therefore more natural

639   to retrieve the pigment concentrations, which is the quantity we measured, than the associated PSC

640   or PFT, which are estimated from the pigment observations through complex non-linear and region-

641   dependent algorithms (*Uitz et al*, 2006). Due to the characteristics of the DPIG, the method can

642   retrieve pigment concentration patterns over a large range ($0.02 - 2$ mg m^{-3}).

643   - We were able to analyze the pigment concentration in the Senegalo-Mauritanian region by processing

644   satellite ocean color observations with the 2S-SOM. We found an important seasonal signal of

645   fucoxanthin concentration with a maximum occurring in March. We evidenced a large offshore

646   gradient of fucoxanthin concentrations, the near shore waters being richer than the offshore ones. We

647   showed that the offshore region waters correspond to Case-1 waters, while the near shore waters are

648   close to Case-2 waters and are influenced by the variability of near shore process like turbidity, or

649   the presence of dissolved matters. The UPSEN measurements show that the pigment ratios of the

650   Senegalo-Mauritanian region are in the range of the DPIG database used to calibrate the method,

651   which justifies the use of the 2S-SOM algorithm to investigate this region.

652   - We used daily satellite observations to construct a monthly climatology of pigment concentrations

653   of the Senegalo-Mauritanian upwelling region, which has been poorly surveyed by oceanic cruises.

654   Due to the highly non-linear character of the algorithms for determining the pigment concentrations

655   from satellite measurements, it is mathematically more rigorous to apply these algorithms to daily

656   satellite data and to average this daily estimate for the climatology period under study, than to

657   estimate them from the satellite data climatology, as many authors have done (*Uitz et al., 2010*;

658   *Hirata et al.,* 2011). We found that Fucoxanthin starts developing in December North of 19°N,

659   presents its maximum intensity in March when the upwelling intensity is maximum, extends up to

660   the coast of Guinea (12°N) in April and begins to decrease in May

661

662   Another important aspect of our study concerns the validity of our results. The 2S-SOM method has

663   been validated by focusing the retrieval accuracy on the fucoxanthin ratio, by using a cross-validation

664   procedure. These results were qualitatively confirmed by two other independent studies.

665   - We first applied a cross validation procedure (see section 4.1), which is powerful technique for

666   validating models (*Kohavi,* 1995; *Varma* and *Simon*, 2006). We learned 30 different 2S-SOM using

667   30 different learning dataset determined at random from the DPIG dataset (each learning dataset

668   representing 90% of DPIG) and 30 test datasets (10% of DPIG). By averaging the results, we found

669   that the 2S-SOM method retrieves the fucoxanthin concentration with a good score (see the

670   statistical parameters in table 2) which confirms the pertinence of the method.

671   - We then found that our fucoxanthin climatology is in agreement with in situ observations of

672   phytoplankton reported in *Blasco et al.* (1980) in March to May 1974 off the coast of Senegal during

673   the JOINT I experiment. These authors analyzed 740 water samples collected with Niskin bottles

674   at 136 stations extending along a line at 21°40'N (in the northern part of the studied region) from 0

675   to 100 km offshore. The samples were taken at several depths (mostly at 100, 50, 30, 15, 5 m).

676   Phytoplankton cells were counted and identified by the Utermohl inverted microscope technique

677   (*Blasco,* 1977). These authors found that diatoms reach their maximum concentration in April–May

678   and are the most abundant group in that period, whereas the other cells predominate in March.

679   Similar microscope observations have been reported in the ocean area south of Dakar by *A. Dia*

680   (1985) during several ship surveys in February–March 1982–1983.

681   - Our method is also in agreement with the monthly eleven years climatology presented in *Farikou et*

682   *al,* (2015) who used a modified PHYSAT method to retrieve the *PFT* in the Senegalo-Mauritanian

683   region.

684   - The pigment concentrations provided by the 2S-SOM from the VIIRS sensor observations are in

685   qualitative agreement with the in-situ measurements done at five stations during the two UPSEN

686   campaigns in 2012 and 2013, showing that the method is able to function in waters where the

687   pigment concentrations are quite high (fucoxanthin ratios of the order 0.4).

688

689

690

691

692

693 **7 - CONCLUSION**

694

695 We developed a new neural network clustering method, the so-called 2S-SOM algorithm to retrieve

696 phytoplankton pigment concentration from satellite ocean color multi spectral sensors. The 2S-SOM

697 algorithm is a SOM specifically designed to deal with a large number of heterogeneous components

698 such as optical and chemical measurements. The major advantage of 2S-SOM with respect to the

699 classical SOM is to cluster variables having similar significance in blocks having specific weights.

700 The weights attributed to the blocks during the learning phase vary with the quality of the variables in

701 the classification. This permits to modulate the variable influence in the cost function, which makes

702 the clustering more informative than that provided by the SOM. Besides, the block weighting provides

703 useful information on the functioning of the classification by permitting to identify the variables which

704 control it. It also allows us to better understand the dynamics of the phytoplankton communities.

705 The 2S-SOM method is efficient and rapid as soon as the calibration is done, since it uses elementary

706 algebraic operations only. The 2S-SOM method is like a piecewise regression that takes advantage of

707 the unsupervised classification of the SOM. We decomposed the DPIG database into quite a large

708 number of partitions ($9\mathbf{x}8=162$) when comparing our study to other studies (*Uitz et al*, 2006, 2012).

709 The validity of the method has been controlled through a cross validation procedure and confirmed by

710 three qualitative studies. Statistical parameters (R^2 coefficients, RMSE and P-values) of the cross-

711 validation between the DPIG in situ pigments and the pigments given by the 2S-SOM averaged for the

712 30 2S-SOM realizations presented in table 2, show the good performance of the method. It must be

713 noticed that the performance mainly depends on the size of the learning set used to calibrate the 2S-

714 SOM. This set must include all the situations encountered in the pigment retrieval. The larger the

715 learning set, the better the method performs. Due to its generic character and its flexibility, the method

716 could be used to determine a large variety of measures done with satellite remote sensing

717 observations.

718 In this work, the method was applied to study the seasonal variability of the fucoxanthin concentration

719 in Senegalo-Mauritanian upwelling region. We showed a large offshore gradient of fucoxanthin, the

720 higher concentration being situated near the shore. We were able to construct a monthly climatology

721 for an 11-year period (1998–2009) of the SeaWiFS observations by summing the daily pixels of the

722 month under study in a region which was poorly surveyed by oceanic cruises. The fucoxanthin

723 concentration, and consequently the associated diatoms, present a well-marked seasonality (Figure 10).

724 Fucoxanthin starts developing in December North of 19°N, presents its maximum intensity in March

725 when the upwelling intensity is maximum, extends up to the coast of Guinea (12°N) in April and

726 begins to decrease in May where it is observed north of Cabo Verde peninsula (15°N), in agreement

727 with the observations reported by *Farikou et al,* (2015) and *Demarcq and Faure*, (2000). The UPSEN

728 campaign results endorse the validity of the study of the Senegalo-Mauritanian upwelling region done

729 with the 2S-SOM.

730

Acknowledgments

The study was supported by the projects CNES-TOSCA 2013-2014 and 2014-2015. The water-leaving reflectances were obtained from the SeaWiFS daily reflectances, $\rho obsTOAw(\lambda)$, provided by the NASA/GSFC/DAAC observed at the top of the atmosphere (TOA) and processed with the SOM-NV algorithm (Diouf et al., 2013) from 1998 to 2010. They are available at the web site: http://poacc.locean-ipsl.upmc.fr/. The DPIG data base was kindly provided by Dr. S. Alvain. We thank Dr. Alban Lazar and Dr. E. Machu for providing in situ data measured during the UPSEN experiments as well as stimulating discussions for their interpretation. We also thank Ray Griffiths for editing the manuscript.

740
741

References

Alvain S, Moulin C., Dandonneau Y. and Breon F. M. : Remote sensing of phytoplankton groups in case-1 waters from global SeaWiFS imagery. Deep-Sea Res. Part1, V 5**2** (11), pp 1989-2004, 2005.

Alvain, S. Loisel H. and Dessailly D. : Theoretical analysis of ocean color radiances anomalies and implications for phytoplankton group detection. Optics Express, V **20** (2), 2012.

Antoine D., André J. M. , Morel A. : Oceanic primary production : Estimation at global scale from satellite (Coastal Zone Color Scanner) chlorophyll. Global Biogeochem Cy. V **10**, pp 57-69, 1996.

Badran F., Berrada M. , Brajard J., Crepon M. , Sorror C., Thiria S., Hermand J.P. , Meyer M., Perichon L., Asch M. : Inversion of satellite ocean colour imagery and geoacoustic characterization of seabed properties : Variational data inversion using a semi-automatic adjoint approach J. Marine Systems, V 69, pp 126-136, 2008

Behrenfeld M. J., Boss E., Siegel D.A., Shea D.M. : Carbon-based ocean productivity and phytoplankton physiology from space. Global Biogeochem. Cy. V 19**,** GB1006, doi:10.1029/2004GB002299, 2005

Behrenfeld M. J., and Falkowski P.G. : Photosynthetic rates derived from satellite base chlorophyll concentration. Limnol. Oceanogr, V **42,** pp 1-20, 1997

Ben Mustapha Z. S., Alvain S. , Jamet C., Loisel H. and Desailly D. : Automatic water leaving radiance anomalies from global SeaWiFS imagery: application to the detection of phytoplankton groups in open waters. Remote Sens. Environ., vol 146, pp 97-112, 2014.

Blasco D. : Red tide in the upwelling region of Baja California. Limnol. Oceanogr. vol 22, pp 255-263, 1977

Blasco D., Estrada M. and Jones B. : Relationship between the phytoplankton distribution and composition and the hydrography in the northwest African upwelling region, near Cabo Corbeiro. Deep-Sea Res. , vol 27A, pp 799-821, 1980.

Bracher A., Bouman HA, Brewin RJW, Bricaud A, Brotas V, Ciotti AM, Clementson L, Devred E, Di Cicco A, Dutkiewicz S, Hardman-Mountford NJ, Hickman AE, Hieronymi M, Hirata T, Losa SN, Mouw CB, Organelli E, Raitsos DE, Uitz J, Vogt M and Wolanin A : Obtaining Phytoplankton Diversity from Ocean Color: A Scientific Roadmap for Future Development. Front. Mar. Sci. 4:55. doi: 10.3389/fmars.2017.00055, 2017

Brajard J., Jamet C., Moulin C. and Thiria S. : Atmospheric correction and oceanic constituents retrieval with a neuro-variational method. Neural Networks, Vol 19(2), p178-185, 2006

774   Brajard J., Jamet C., Moulin C. and Thiria S : Neurovariational inversion of ocean color images. J.
775   Atmos. Space Res. Vol 38, n 2, pp 2169-2175, 2006

776   Brewin R. J. W., Hardman-Mountford N. J., Lavender S. J., Raitsos D. E., Hirata T., Uitz J., et al. :
777   An inter-comparison of bio-optical techniques for detecting dominant phytoplankton size class
778   from satellite remote sensing. Remote Sens. Environ. 115, 325–339. doi:
779   10.1016/j.rse.2010.09.004, 2011

780   Brewin R. J. W., Sathyendranath S., Hirata, T., Lavender, S.J., Barciela, R., Hardman-Montford, N.J :
781   A three-component model of phytoplankton size class for the Atlantic Ocean. Ecol. Model. vol **22,**
782   pp 1472-1483, 2010.

783   Bricaud A., Mejia C. , Blondeau Patissier D. , Claustre H., Crepon M. and Thiria S. : Retrieval of
784   pigment concentrations and size structure of algal populations from absorption spectra using
785   multilayered perceptrons. Applied Optics Mars 2007 vol 46 n°8., 2006

786   Capet X., Estrade, P., Machu, E., Ndoye, S. et al. : On the Dynamics of the Southern Senegal
787   Upwelling Center: Observed Variability from Synoptic to Superinertial Scales : J. Phys. Oceanogr.
788   vol **47** (1), pp 155-180, 2017

789   Cavazos T. : Using Self-Organizing Maps to Investigate Extreme Climate Events: An Application to
790   Wintertime Precipitation in the Balkans. J. Climate, vol **13**, 1718–1732, 2000.

791   Chazotte A., Crepon M., Bricaud A., Ras J. and Thiria S. : Statistical analysis of absorption spectra
792   of phytoplankton and of pigment concentrations observed during three POMME cruises using a
793   neural network clustering method. Applied Optics, 46 (18), 3790-3799, 2007

794   Chazottes A., Bricaud A., Crepon M. and Thiria S. : Statistical analysis of a data base of absorption
795   spectra of phytoplankton and pigment concentrations using self-organizing maps. Appl. Opt. 45,
796   8102-8115, 2006

797   Ciotti A. and Bricaud A. : Retrievals of a size parameter for phytoplankton and spectral light absorption
798   by colored detrital matter from water-leaving radiances at SeaWiFS channels in a continental shelf
799   region off Brazil. Limnol. Oceangr. Methods, vol **4**, pp 237-253, 2006.

800   Demarcq H. and Faure V. : Coastal upwelling and associated retention indices from satellite SST.
801   Application to Octopus vulgaris recruitment. Oceanografica Acta, vol **23**, pp 391-407, 2000.

802   Dia A. Biomasse et biologie du phytoplancton le long de la petite côte sénégalaise et relations avec
803   l'hydrologie. Rapport interne N°44 du CRODT, Réf: 0C000798, 1981-1982. On line on the web
804   site:http://www.sist.sn/gsdl/collect/publi/index/assoc/HASH2127.dir/doc.pdf

805   Diouf D., Niang A., Brajard J., Crepon M. and Thiria S. : Retrieving aerosol characteristics and sea-
806   surface chlorophyll from satellite ocean color multi-spectral sensors using a neural-variational
807   method. Remote Sens. Environ. **vol 130**, pp 74-86, 2013.

808 Farikou O., Sawadogo S., Niang A., Brajard J., Mejia C., Crépon M. and Thiria S. : Multivariate
809   analysis of the Sénégalo-Mauritanian area by merging satellite remote sensing ocean color and SST
810   observations. J. Environ. Earth Sci. vol **5** (12), pp 756-768, 2013

811 Farikou O., Sawadogo S., Niang A., Diouf D., Brajard J., Mejia C., Dandonneau Y., Gasc G., Crepon
812   M., and Thiria S. : Inferring the seasonal evolution of phytoplankton groups in the Sénégalo-
813   Mauritanian upwelling region from satellite ocean-color spectral measurements, J. Geophys. Res.
814   Oceans, vol **120**, pp 6581-6601, 2015.

815 Friedrich T. and Oschlies A. : Basin-scale pCO2 maps estimated from ARGO float data : A model
816   study, J. Geophys. Res.*, vol **114**, C10012, doi: 10. 1029/2009JC005322, 2009.

817 Gordon H. R. : Atmospheric correction of ocean color imagery in the Earth Observing System era. J.
818   Geophys. Re. Atmospheres, vol **102**(D14), pp 17081-17106, 1997.

819 Hewitson B.C. and Crane R. G. : Sef organizing maps : application to synoptic climatology. Climate
820   research, vol **22**, pp 13-26, 2002

821 Gross L., Frouin R., Dupouy C., Andre J. M. and Thiria S. : Reducing biological variability in the
822   retrieval of chlorophyl_a concentration from spectral marine reflectance. Applied Optics, Vol. 43
823   Issue 20 pp. 4041, 2004

824 Gross L., Thiria S., Frouin R., Mitchell B.G : Artificial neural networks for modeling transfer
825   function between marine reflectance and phytoplankton pigment concentration J. Geophys. Res.
826   Vol 105,no.C2, pp3483-3949, february 15, 2000.

827 Hirata T. , Aiken J., Hardman-Mountford N., Smyth T. J. and Barlow R.G. : An absorption model to
828   determine phytoplankton size classes from satellite ocean color, Remote Sens. Environ. vol **112**, pp
829   3153-3159, 2008.

830 Hirata T. , Hardman-Mountford N.J., Brewin R.J.W., Aiken J., Barlow R., Suzuki K., Isada T., Howell
831   E., Hashioka T., Noguchi-Aita M. and Yamanaka Y. : Synoptic relationships between surface
832   chlorophyll-*a* and diagnostic pigments specific to phytoplankton functional types. Biogeosciences,
833   vol **8** (2): pp 311-327, 2011.

834 Jamet C., Thiria S., Moullin C., Crepon M. : Use of a neural inversion for retrieving Oceanic and
835   Atmospheric constituents for Ocean Color imagery : a feasability study.
836   doi:10.1175/JTECH1688.1, J. Atmos. Ocean. Techno. :/ Vol. 22, No. 4, pp. 460–475, 2005

837 Jeffreys S.W. and Vesk M. : Phytoplankton Pigment in Oceanography : Guidelines to Modern
838   Methods, UNESCO, Paris, ed S. W. Jeffery, R.F.C. Mantoura and S. W. Wright, Introduction to
839   marine phytoplankton and their pigment signatures, pp 33-84, 1997.

840 Jouini M., Lévy M. , Crépon M. and Thiria S. : Reconstruction of ocean color images under clouds
841   using a neuronal classification method. Remote Sens. Environ. vol **131**, pp 232-246, 2013

842 Kohavi R. : A study of cross-validation and bootstrap for accuracy estimation and model selection.
843  Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo,
844  CA: Morgan Kaufmann ed.. **2** (12): pp 1137–1143, 1995.

845 Kohonen T : Self-organizing maps (3rd ed.). Springer, Berlin Heidelberg New York. 2001

846 Kruizinga S. and Murphy A : Use of an analogue procedure to formulate objective probabilistic
847  temperature forecasts in the Netherlands. Mon. Wea. Rev., vol **111,** pp 2244–2254, 1983.

848 Le Quéré et al, (2018) : Global Carbon Budget 2018, Earth Syst. Sci. Data, 10, 2141–2194, 2018 ;
849  https://doi.org/10.5194/essd-10-2141-2018

850 Lévy M., D. Iovino, L. Resplandy, P. Klein, G. Madec, A.-M. Tréguier, S. Masson, K. Takahashi, Large-scale
851  impacts of submesoscale dynamics on phytoplankton: Local and remote effects, Ocean Modelling, 77–93,
852  2012

853 Levy, M., Mesoscale variability of phytoplankton and of new production: Impact of the large-scale nutrient
854  distribution, J. Geophys. Res., 108(C11), 3358, doi:10.1029/2002JC001577, 2003.

855 Liu Y. and Weisberg R. H. : Patterns of ocean current variability on the West Florida Shelf using the
856  self-organizing map, J. Geophys. Res., **110,** C06003, doi:10.1029/2004JC002786, 2005

857 Liu Y., Weisberg R. H., and He R. : Sea surface temperature patterns on the West Florida Shelf using
858  growing hierarchical self-organizing maps, J. Atmos. Oceanic Technol., vol **23**(2), pp 325– 338, 2006

859 Longhurst A. R., Sathyendranath S., Platt T., Caverhill C. : An estimation of global primary production
860  in the ocean from satellite radiometer data. J. Plank. Res. vol **17**, pp 1245-1271, 1995

861 Lorenz E. N : Atmospheric predictability as revealed by naturally occurring analogs. J. Atmos. Sci.,
862  vol 26, pp 639–646, 1969

863 Morel A. and Gentili G. : Diffuse reflectance of oceanic waters. III. Implication of bidirectionality for
864  the remote-sensing problem. Appl. Opt. vol 35, pp 4850-4862, 1996.

865 Mouw C. B. and Yoder J. A. : Optical determination of phytoplankton size composition from global
866  SeaWiFS imagery. J. Geophys. Res. vol **115**, C12018, doi:10.1029/2010JC006337, 2010.

867 Ndoye S. , Capet X., Estrade P., Sow B., Dagorne D., Lazar A., Gaye A. and Brehmer P. : SST patterns
868  and dynamics of the southern Senegal-Gambia upwelling center. J. Geophys. Res. Oceans, vol 119,
869  pp 8315–8335. 2014

870 Niang, A., Gross, L., Thiria, S., Badran, F., & Moulin, C. Automatic neural classification of ocean
871  colour reflectance spectra at the top of atmosphere with introduction of expert knowledge.
872  Remote Sens. Environ, vol 86, pp 257–271, 2003.

873 Niang A., Badran F., Moulin C., Crépon M. and Thiria S. : Retrieval of aerosol type and optical
874  thickness over the Mediterranean from SeaWiFS images using an automatic neural classification
875  method. Remote Sens. Environ. vol 100, pp 82-94, 2006.

876 O'Reilly, J.E., Maritorena , S., Siegel, D. A., O'Brien, M. C ., Toole, D., Mitchell, B. G., Kahru, M.,

877   Chavez, F. P., Strutton, P., Cota, G. F., Hooker, S. B., McClain, C. R., Carder, K. L., Muller-

878   Karger, F., Harding, L., Magnuson , A., Phinney, D., Moore, G.F., Aiken, J., Arrigo, K. R.,

879   Letelier, R., and Culver, M.   Ocean color chlorophyll a algorithms for SeaWiFS, OC2 and

880   OC4: Version 4. In S. B. Hooker, and E. R. Firestone (Eds), *SeaWiFS postlaunch calibration and*

881   *validation analyses: Part 3. NASA Tech. Memo. 2000-206892, vol. 11*(pp.9-23). Greenbelt, MD:

882   NASA Goddard Space Flight Center. 2001.

883 Palacz A. P., St. John, M. A., Brewin, R. J.W., Hirata, T., and Gregg,W.W. : Distribution of

884   phytoplankton functional types in high-nitrate low-chlorophyll waters in a new diagnostic

885   ecological indicator model. Biogeosciences 10, 7553–7574. doi: 10.5194/bg-10-7553, 2013.

886 Raitsos D. E., Lavender, S. J., Maravelias, C. D., Haralambous, J., Richardson, A. J., and Reid, P.

887   C. : Identifying phytoplankton functional groups from space: an ecological approach. Limnol.

888   Oceanogr. 53, 605–613. doi: 10.4319/lo.2008.53.2.0605, 2008

889 Reusch D. B., Alley, R. B., and Hewitson, B. C : North Atlantic climate variability from a self-

890   organizing map perspective, J. Geophys. Res., vol **112**, D02104, doi:10.1029/2006JD007460, 2007.

891 Sathyendranath S., Watts S., L., Devred E., Platt T., Caverhill C. M., and Maass H. : Discrimination

892   of diatom from other phytoplankton using ocean-colour data, Mar. Ecol. Prog. Ser., vol 272, pp 59–

893   68, 2004.

894 Sirven J., Mignot J., Crépon M. : Generation of Rossby waves off the Cap Verde Peninsula: the role

895   of the coastline . Ocean Sci., 15, 1–24, 2019

896 Sosik, H.M.; Sathyendranath, S.; Uitz, J.; Bouman, H.; Nair, A. In situ methods of measuring

897   phytoplankton functional types. In Phytoplankton Functional Types from Space. IOCCG report, No.

898   15; Sathyendranath, S., Ed.; IOCCG: Dartmouth, NS, Canada, pp. 21–38, 2014.

899 Uitz J., Claustre H., Morel A. and. Hooker S.B : Vertical distribution of phytoplankton communities

900   in open ocean: an assessment based on surface chlorophyll. J. Geophys. Res. **111,** C08005,

901   doi:10:1029/2005JC003207. 2006

902 Uitz J., Claustre H., Gentili B. and Stramski D. : Phytoplankton class-specific primary production in

903   the world's ocean: seasonal and interannual variability from satellite observations. Global

904   Biogeochem. Cycles, vol **24**, GB 3016, doi:10:1029/2009GB003680, 2010

905 Van den Dool H. : Searching for analogs, how long must we wait? Tellus, vol **46A**, pp 314–324, 1994.

906 Varma, S., Simon, R. : Bias in error estimation when using cross-validation for model selection; BMC

907   Bioinformatics. vol **7**. PMC 1397873 . PMID 16504092. doi:10.1186/1471-2105-7-91, 2006

908 Vidussi F., Claustre H., Manca B. B., Luchetta A. and Marty J. C. : Phytoplankton pigment distribution
909   in relation to upper thermocline circulation in the eastern Mediterranean sea during winter. J.
910   Geophys. Res., vol 106, pp 19,939-19,956, 2001.

911 Westberry T., Behrenfeld M.J., Siegel D. A. and Boss E.: Carbon-based productivity modeling with
912   vertically resolved photoacclimatation. Global Biogeochem. Cycles, vol **22**, *GB2024*,
913   DOI:10.1029/2007GB003078, 2008

914 Zorita E. and Von Storch H. : The Analog Method as a Simple Statistical Downscaling Technique:
915   Comparison with More Complicated Methods. Journal of Climate, vol **12,** pp 2474-2489, 1999.

916

917 **ANNEX 1**

918

919   **A1 Cost function of the SOM**

920 Let us recall the following notation:

921   $D = \{z_1, \cdots, z_i, \cdots, z_K\}$ the dataset composed of K vectors $z_i \in \mathbb{R}^N$

922   $W = \{w_1, \cdots, w_c, \cdots, w_C\}$ the set of weights $w_c \in \mathbb{R}^N$ where $C = p \times q$ is the size of the SOM.

923 The w_c of the SOM are estimated by minimizing a cost function of the form

924

925     $J^T_{SOM}(\chi, W) = \sum_{i=1}^{K} \sum_{c=1}^{p \times q} K^T\left(\delta\left(c, \chi(z_i)\right)\right) \|z_i - w_c\|^2,$ $(A.1)$

926 where c indices the neurons of the SOM map, χ is the allocation function that assigns each element z_i

927 of D to its referent vector w_c which is of the form $\chi(z_i) = \arg\min_c \|z_i - w_c\|^2$,

928 $\delta\left(c, \chi(z_i)\right)$ is the discrete distance on the SOM between a neuron if index c and the neuron allocated

929 to observation z_i, and K^T a kernel function parameterized by T that weights the discrete distance on

930 the map and decreases during the minimization process. T acts as a regularization term (*Kohonen,* 2001,

931 *Niang et al,* 2003). In the present case K^T is of the form :

932 $K^T(\delta) = (1/T)K(\delta/T)$, where K is the gaussian function of mean 0 and standard deviation 1.

933 The cost function (A1) takes into account the proper inertia of the partition of the data set D and ensures

934 that its topology is preserved.

935

936   **A2 Definition of the Algorithm 2S-SOM**

937 The 2S-SOM algorithm is an extension of the Self-Organizing maps (SOM, *Kohonen,* 2001) based on

938 the K-mean method (*Ouattara et al.*, 2014**,** https://www.theses.fr/179489704). It automatically

939 structures the variables having some common characters into conceptually meaningful and

940 homogeneous blocks during the learning phase. The 2S-SOM takes advantage of this structuration of

941 D and the variables into B different blocks, which permits an automatic weighting of the influence of

942 each block and consequently of each variable in the classification phase. The 2S-SOM is based on a

943 modification of the cost function of the SOM algorithm. For a neuron of index c, we define the weights

944 α_{cb} of each block b ($b = 1, ..., B$) and the weights β_{cbj} of the variables j ($j = 1, ..., P_b$) in this block,

945 where P_b is the number of variable in the block indexed by b. The vectors of weighs are denoted

946 $\boldsymbol{\alpha} = \{\alpha_{cb}\}_{1 \leq c \leq C, 1 \leq b \leq B}$ and $\boldsymbol{\beta} = \{\beta_{cbj}\}_{1 \leq c \leq C, 1 \leq b \leq B, 1 \leq j \leq P_b}$

947 The new cost function is:

948 $$J^T_{2S-SOM}(\chi, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_c \left(\sum_{b=1}^B \left(\sum_{zi \in D} \alpha_{cb} K^T \left(\delta(c, \chi(z_i)) \right) d_{\beta_{cb}}(i) + J_{cb} \right) + I_c \right),$$ (A.2)

949 with

950 $$d_{\beta_{cb}}(i) = \sum_{j=1}^{P_b} \beta_{cbj} (z_{ib}^j - w_{ib}^j)^2,$$ (A.3)

951 where c indices the neurons of the 2S-SOM map.

952 under the two constraints:

953 $$\sum_{b=1}^B \alpha_{cb} = 1; \alpha_{cb} \in [0,1] \; \forall c, 1 \le c \le C$$ (A.4)

954 and

955 $$\sum_{j=1}^{P_b} \beta_{cbj} = 1; \beta_{cbj} \in [0,1], \forall c, 1 \le c \le C; \forall b, 1 \le b \le B.$$

956 I_c and J_{cb} are used to regularize the weights α and β. They are defined as negative entropies weighted
957 by μ for the blocks and η for the variables of each block

958

959 $$I_c = \mu \sum_{b=1}^{P_b} \alpha_{cb} log(\alpha_{cb})$$ (A.6)

960 and

961 $$J_{cb} = \eta \sum_{j=1}^B \beta_{cbj} log(\beta_{cbj})$$ (A.7)

962 The topological conservation properties of 2S-SOM are influenced by the weights α_{cb} and β_{cbj} in the
963 classification through the hyper-parameters μ, η and the neighborhood parameter T.

964 The weights α_{cb} and β_{cbj} respectively indicate the relative importance of blocks and variables in the
965 neurons. Thus, the greater the weight of a block b or a variable j, the more the block or the variable
966 contributes to the definition of the class (or neuron) in the sense that it makes it possible to reduce the
967 variability of the observations in the cell and in its close neighborhood. For a high value of η and a
968 fixed one for μ, the β_{cbj} in a block are equal to $1/P_b$. In this case, only the blocks are modified according
969 to their capacity to define the neurons. In this context, the 2S-SOM then makes possible to weight the
970 different blocks for each neuron

971   -   For high values of μ, I_c is large. The minimization of J_{cb} forces all its coefficients to become

972   equal. For a fixed value of η, the α_{cb} associated with the blocks are all equal to $1/B$. In this case,

973   only the β_{cbj} of the variables inside the blocks weight the neurons

974   -   When μ and η tend to very large values, the blocks are equiprobable as well as the variables.

975   Thus, the 2S-SOM algorithm is comparable to the SOM.

976

977   **A.3 How the 2S-SOM algorithm works:**

978   For fixed μ and η, the learning of the 2S-SOM algorithm is as follows:

979   -   Step 0: Initialization with iteration of the algorithm SOM, by setting $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to homogeneous

980   values.

981   The optimization of J_{2S-SOM}^{T} is carried out through an iterative process composed of three steps (1, 2,

982   and 3) presented below.

983   -   Step 1: The w_c referents, the weights α and β are known and fixed, the observations are assigned

984   to the neurons by respecting the assignment function:

985
$$c(zi) = \chi(z_i) = \arg\min_{c\in C}\left(\sum_{r\in C} K^T\big(\delta(r,c)\big)\left(\sum_{b=1}^{B} \alpha_{cb}d_{\beta_{cb}}(i)\right)\right) \quad (A.8)$$

986

987   -   Step 2: Updating the neuron centers (the w_c referents) according to the formula of the SOM

988   algorithm.

989

990   -   Step 3: the assignment function and the referents w_c being fixed, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are determined

991   according to the equations (A.9, A.10, A.11, A.12), by minimizing the cost function

992   J_{2S-SOM}^{T} with respect to α and β under the constraints A.4 and A.5.

993
$$\alpha_{cb} = \frac{\exp\left(\frac{-\psi_{cb}}{\mu}\right)}{\sum_{b=1}^{B}\exp\left(\frac{-\psi_{cb}}{\mu}\right)} \qquad (A.9)$$

994   with

995
$$\psi_{cb} = \sum_{zi\in D} K^T\big(\delta(\chi(z_i),c)\big)d_{\beta_{cb}}(i) \qquad (A.10)$$

996   and

997
$$\beta_{cbj} = \frac{\exp\left(\frac{-\Phi_{cbj}}{\eta}\right)}{\sum_{b=1}^{p_b}\exp\left(\frac{-\Phi_{cbj}}{\eta}\right)} \qquad (A..11)$$

998   with

999
$$\Phi_{cbj} = \sum_{zi \in D} \alpha_{cb} K^T(\chi(z_i), c)(z_{ib}^j - w_{cb}^j)^2 \qquad (A.12)$$

1000

1001   This algorithm is repeated by sampling the hyper-parameters μ and η until convergence.

1002   Finally, at the convergence, the 2S-SOM provides on the one hand a topological map allowing to
1003   visualize the data, and on the other hand a weight system for the neurons of the map allowing us to
1004   interpret the role of the different variables and to choose those that are the most significant for the
1005   classification and neutralizing those which are the least one.

1006

Author Contribution

Dr N'Dye Niang and Maurice Ouattara provided the 2S-SOM code, Khalil Yala processed the data and did the computations with the 2S-SOM, Sylvie Thiria, Michel Crepon and Julien Brajard analyzed the results, Carlos Mejia and Roy El Hourany did the statistical tests presented in tables and figure 13. Prof. Sylvie Thiria conceived and supervised the study.

Code/data availability

The satellite data (ocean color and SST) are available at the web site:
 http://poacc.locean-ipsl.upmc.fr/.

The DPIG data base was kindly provided by Dr. S. Alvain (Severine.alvain@univ-littoral.fr)

The UPSEN data are available at : alban.lazar@locean-ipsl.upmc.fr

The 2S-SOM code is available on request at: carlos.mejia@locean-ipsl.upmc.fr

1020 **Figure Caption**

1021

1022

1023

1024 Figure 1: *Mauritania and Senegal coastal topography. The land is in brown and the ocean depth*

1025 *is represented in meters by the color scale on the right side of the figure. The UPSEN stations are*

1026 *shown at the bottom left cartoon of the figure*

1027

1028 Figure 2: *Geographic positions of the 515 in situ and satellite collocated measurements of the DPIG*

1029 *database.*

1030

1031 Figure 3: *Dispersion diagram of DPIG chl-a computed from the SeaWiFS observations using the*

1032 *OC4V4 algorithm versus in situ chl-a. The coefficient of vraisemblance R^2 and the RMSE (Root Mean*

1033 *Square Error) were computed in in mg m^{-3}*

1034

1035 Figure 4: *Flowchart of the method: top panel - Learning phase; bottom panel – operational phase*

1036 *which consists in pigment retrieval and the determination of the α_{cb} block parameters.*

1037

1038 Figure 5: *Flowchart of the cross-validation procedure for 30 partitions of the DPIG database.*

1039

1040 Figure 6: *2S-SOM Map. From left to right and top to bottom, values of the referent vectors for $\rho_w(490)$,*

1041 *Ra(490), SeaWiFS chl-a, and fucoxanthin, peridinin, divinyl Ratios. The number in each neuron indicates the*

1042 *amount of DPIG data captured at the end of the learning phase, the values indicated by the color bars are*

1043 *centered-reduced and non-dimensional values.*

1044

1045 Figure 7: *2S-SOM map. Weights (α_{cb}) of the four block parameters determined at the end of the learning*

1046 *phase; from left to right and top to bottom: ρ_w, Ra, Pigment, SeaWifs chl-a. The color bars show the % of the*

1047 *weight estimated by 2S-SOM, a value of 1 or 0 indicating that the data in the neuron are assembled with respect*

1048 *to that block only.*

1049

1050 Figure 8: *A) chl-a concentration, (B) fucoxanthin ratio, (C) aerosol optical thickness, (D) peridinin*

1051 *for 1 January 2003. Panels (B) and (D) show that a second-order information was retrieved, which is*

1052 *correlated with the chl-a concentration (A) but not equivalent. The aerosol optical thickness (C) does*

1053 *not seem to contaminate the estimated parameters (fucoxanthin and peridinin ratios).*

1054

1055 Figure 9: *SST for 2 January 2003. Note the well-marked upwelling (cold temperature) north of 13°N.*

1056

1057

1058 Figure 10: *(A) chl-a concentration, (B) fucoxanthin ratio, (C) aerosol optical thickness, (D) peridinin for 6*
1059 *January 2003. Panels (B) and (D) show that a second-order information was retrieved, which is correlated with*
1060 *the chl-a concentration (A) but is not equivalent. It is found that the aerosol optical thickness (C) does not*
1061 *contaminate the estimated parameters (fucoxanthin and peridinin ratios).*

1062

1063 Figure 11: *(A) chl-a concentration, (B) fucoxanthin ratio, (C) aerosol optical thickness, (D) Peridinin for*
1064 *28 February 2003. Panels (B) and (D) show that a second order information was retrieved, which is correlated*
1065 *with the chl-a concentration (A) but is not equivalent. It is found that the aerosol optical thickness (C) does not*
1066 *contaminate the estimated parameters (fucoxanthin and peridinin ratios). The position of the NSB and OFB*
1067 *boxes are figured out by black square boxes*

1068

1069 Figure 12: *Reflectance spectra (in blue) captured the 28 February by six neurons whose referent*
1070 *vector spectra are in yellow: top line, for pixels in the NSB region (long. [-20°, -18°], lat. [12°,*
1071 *14°]); bottom line, for pixels in the OFB region (long. [-28°, -26°], lat. [12°, 14°]).*

1072

1073 Figure 13: *Box plot of the weights of the selected neurons during the decoding of the 28 February*
1074 *data. From left to right, weights of blocks B1, B2, B3, B4. Top panel, in the NSB region (long. [-20°,*
1075 *-18°], lat. [12°, 14°]); bottom panel, in the OFB region (long. [-28°, -26°], lat. [12°, 14°]).*

1076

1077 Figure 14: *Monthly fucoxanthin concentration averaged for an 11- years (1998-2009) for December*
1078 *(A), March (B) and May (C).*

1079

1080 Figure 15: *. chl-a (in blue) and fucoxanthin (in green) concentrations for near-shore pixels (in the NSB*
1081 *region).*

1082

1083 Figure 16: *For ship stations 1, 2, 3, 5a and 5b, we show the averaged spectrum of the in situ*
1084 *spectra of the UPSEN station in blue; the spectrum of the referent vector (in red) of the 2S-SOM*
1085 *neuron, which has captured the closest satellite observations to the UPSEN station; among the*
1086 *different spectra constituting the referent spectrum, the spectrum of the learning database*
1087 *(DGIP) that is the closest to the averaged satellite spectra is shown in black. In the rectangular*
1088 *cartoons, we show the position of the UPSEN station, the number of the neuron of the 2S-SOM*
1089 *which has captured the satellite observation, the Rfuco of the referent vector, the Rfuco$_{DGIP}$ of the*
1090 *closest DGIP and the in situ Rfuco$_{UPSEN}$*

1091

1092

Table Caption

1093

1094

1095 Table 1: *Pigments of the DPIG and their statistical characteristics: :STD (Standard Deviation), MIN*

1096 *(minimum value), MAX (maximum value).*

1097

1098 Table 2: *Statistical parameters (R^2 coefficients, RMSE and P-values) of the cross validation between*

1099 *the DPIG in situ pigments and the pigments given by the 2S-SOM averaged for the 30 2S-SOM*

1100 *realizations*

1101

1102 Table 3: *For ship stations 1, 2, 3, 5a and 5b of the UPSEN campaigns, we show the referent captured*

1103 *by the VIIRS observations, the fucoxanthin-ratio associated with this referent (Rfuco-2S-SOM), the*

1104 *fucoxanthin-ratio of the closest DPIG fucoxanthin-ratio captured by the neuron of the referent and the*

1105 *fucoxanthin-ratio measured in situ during the UPSEN campaign*

1106

1107