Data science Case study

Dorra BENNOUR
Khalil BOUGUERRA

ABSTRACT

As we are interested in the study of physical phenomena, and more particularly in phenomena related to the phytoplankton environments, we therefore , aim in this paper to find the locations in the senegal-mauritanian ocean (coast) that are more favourable to produce these kind of organisms. Thus the purpose of this paper consists to identify the concentration of the phytoplankton organisms in the senegal-mauritanian upwelling areas. Our method consists of applying a self organizing map in order to, first, summarize the information held by the data, then proceed with hierachical ascendant clustering HAC.

1 INTRODUCTION

Coastal upwelling systems have received widespread attention for several decades owing to their importance for human society. Although the primary driving mechanism is generic, important differences exist between systems and also between sectors of each given system. The Senegalo-Mauritanian upwelling is a very productive coastal region occurring along the West coast of Africa, Thus it would be interesting to go through its phytoplanktonic system with a deep mindset. In this paper, we aim to define the phytoplanktonic system on the senegal coast by combining the different physical and biogeochemical variables that are related to the system.

2 RELATED WORK

To characterize a rather particular phytoplankton system, several methods have been applied. One of the most widely used methods is Kohonen topological maps known as SOM "Self Organizing Map". These methods have been implemented as part of the characterization of phytoplankton diversity in the Mediterranean Sea (El Hournay et al. 2019) and of a multivariate analysis of the Senegalo-Mauritanian surface (Farikou et al. 2013). SOM maps are indeed a method of vector quantification or reduction of dimensions in order to summarize the information contained in the data. The methodologies followed by the two articles mentioned are not very different, we will then summarize them by the methodology followed by the first article [3]:
In this paper, the objective was to identify the functional types of phytoplankton (PFT), thus, 3 steps were applied:
1. Constitution of the database of pigments, by eliminating empty pixels by constitution of a second database from SOM-Pigments data.
2. Identification of PFTs: Application of the SOM algorithm on the data followed by an ascending hierarchical classification.
3. Validation of the classification results. In another article [1], the authors used an algorithm called the "NASA algorithm", focusing on the spatial and interannual evolution of the biomass of phytoplankton. Thus two different approaches were carried out:
1. synthetic representation of the temporal and spatial variations of the biomass field, i.e. a Hovmoller diagram.

2. A clustered K-means analysis was then performed on the normalized climatology, in order to statistically organize the time series and create clusters representing regions of similarity.

3 EXPLORATORY ANALYSIS

3.1 Data

Daily satellite images are available for this paper between 2004 and 2006 with a resolution of 4 km per pixel (901 * 531 pixels) for 8 oceanographic and biogeochemical variables in coastal waters and off the Senegalese-Mauritanian coasts. These variables are defined as follows: 3 Physical variables (PAR: photosynthetic available radiation (Einstein / m2 / day),ZEU: Euphotic depth (m), SST: Sea Surface Temperature (degree C)) and 5 biogeochemical variables which are phytoplankton pigments(Chla: Chlorophyll-a (mg.m $^{-3}$), Chlb: Chlorophyll-b (mg.m $^{-3}$), Fuco: Fucoxanthin (mg.m $^{-3}$), 19HF: 19-Hexanoyloxy-fucoxanthin (mg.m $^{-3}$),Zea: Zeaxanthine (mg.m $^{-3}$)).

3.2 Univariate Analysis

To fully understand the behavior of the variables presented, as well as to explore the database available, we begin by computing the elementary statistics of each variable resumed by the table 1 below. As our database has a lot of empty pixels where we don't have values due to clouds, we replaced them with the average values. We proceeded then by averaging our database to 8 days. We only work on this database from now on.

	Min	Max	Avg	std
Physical variables				
SST	1.469 10^{1}	39.985	25.308	0.319
PAR	7.999 10^{-3}	63.362	44.953	1.412
ZEU	5.535	171.040	87.793	2.278
biogeochemical variables				
FUCO	-6.938 10^{-8}	0.555	0.014	0.009
19HF	0	0.270	0.0235	0.003
Zea	0	0.121	0.044	0.002
Chla	2.120 10^{-3}	65	0.221	0.2180
Chlb	0	0.154	0.012852	0.002

Table 1: Elementary statistics of all 8 variables

The Table 1 shows that all variables don't have the same magnitude, which is highlighted by the std values that goes from 0.002 for chlb variable to 2.278 for ZEU variable. Thus it is very important to normalize the data.

3.2.1 PAR. The photosynthetic radiation (Figure 1) over the years does not seem to have a great variability. However, by observing more closely, this variable corresponds to high values on the ocean coast (surfaces in yellow).

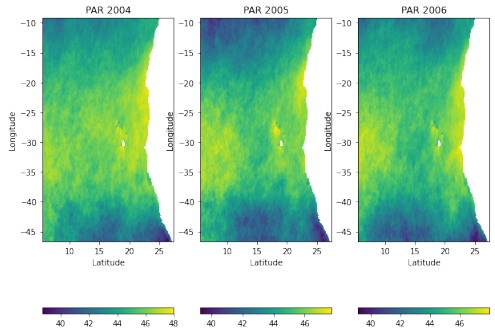

Figure 1: Image of photosynthetic radiation (Einstein / m2 / day) averaged over the year

3.2.2 SST. Among the variables which, plays a very important role during the study of such a system, it is the variable of the temperature of the ocean surface (Figure 2). The variability of the temperature at the sea surface is not perceivable over the years, in fact this is explained by the fact that by averaging the temperatures, we lose the effect of seasonality.Thus it will be more interesting to represent the data as monthly or even weekly data (8 days). The representations of this variable during the different months is represented by the Figure 4 **(à revoir.....)** in the appendix.

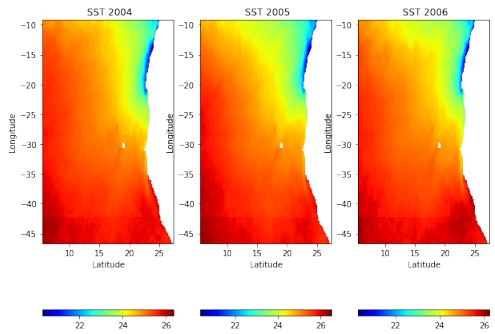

Figure 2: Image of the sea surface temperature SST (° C) averaged over the year

3.2.3 ZEU. Similarly to the other variables, the distribution of the euphotic area is almost the same over the years. Indeed, it is more concentrated in the deep waters of the ocean. Nevertheless, when approaching the coast it seems to be less concentrated. As it is a variable characterizing the marine surface where the residual light intensity makes the photosynthetic activity possible. Figure 3 shows the distribution of this physical variable through the years 2004, 2005 and 2006.

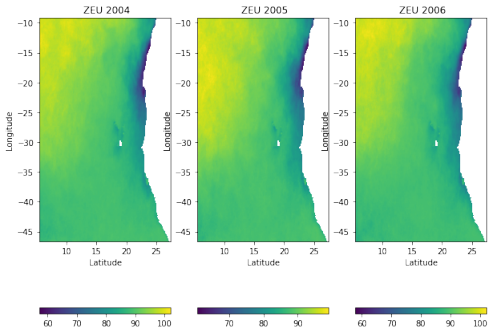

Figure 3: Image of euphotic depth averaged over the year

3.3 Bivariate Analysis

To study the relationship between the different variables, we compute as a first step the correlation matrix, which is demonstrated by figure 4. The Pigments related variables are highly correlated, except for Zea variable which seems to have a small correlation with pigments. Plus, the SST variable is highly correlated to PAR, however it is not much correlated with ZEU. We can also notice that the physical variables are negatively correlated to the biogeochemical variables.

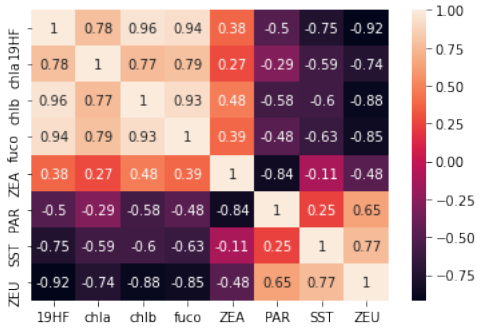

Figure 4: Correlation matrix between variables

We can see from the figure 11 relative to the pairplot that the relationship between the pigment variables seems linear. However, we can't conclude the type of the relationship between the other variables.

4 METHODOLOGY

Our main objective is to find clusters within the provided data to identify zones with high concentration of phytoplankton, with the methods that are inspired from the papers cited above. We choose then to characterize the phytoplankton system by going through these steps:

(1) **8-days averaging:** As the images in the provided data are sparse (non defined values due to cloud) we proceeded by a space (row averaging) and time averaging. As this being said, we would lose the pixels that were nans only on the earth,

thus we've created a mask describing the map, subject to our study. So we now have 137 row for each of the 8 variables.

(2) **Data downsampling:** As our images are of size 531 × 901, which is considered as voluminous, and as we are restrained by the computing power (we work on a 12 GB CPU), we downscale our data by 10, the size of each image become then 54 × 91.

(3) **Normalizing data by chla:** This step is very important in our study, since the phytoplankton depends highly on chla values.
Thus the resulted data is described by the following figure:

	19HF	chla	chlb	fuco	ZEA	PAR	SST	ZEU
2004-01-01	[[0.15280661, 0.13085131, 0.19342797, 0.207742...	[[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,...	[[0.048191722, 0.027165627, 0.045743715, 0.052...	[[0.07255677, 0.06776115, 0.045740616, 0.05178...	[[0.40283254, 0.2954322, 0.25393918, 0.2882731...	[[383.70178, 357.109, 359.70837, 380.32645, 38...	[[285.32538, 276.4334, 290.02307, 291.3764, 29...	[[1173.7472, 1146.6173, 1272.6797, 1266.6656, ...
2004-01-09	[[0.21707366, 0.21484023, 0.25091705, 0.261171...	[[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,...	[[0.044822045, 0.07043291, 0.07150699, 0.08398...	[[0.077381365, 0.09027895, 0.06925899, 0.06518...	[[0.4396278, 0.5272651, 0.3435001, 0.28321567,...	[[431.18286, 465.76132, 480.3882, 509.05432, 5...	[[335.9662, 347.09113, 347.5208, 379.67548, 38...	[[1523.0735, 1531.8356, 1514.7014, 1690.13, 17...
2004-01-17	[[0.23685917, 0.2487284, 0.2769239, 0.26182088...	[[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,...	[[0.07137288, 0.09344571, 0.0793357, 0.0963215...	[[0.070792235, 0.089675814, 0.08024574, 0.0705...	[[0.5513797, 0.6404582, 0.5230715, 0.50188786,...	[[497.2236, 508.39478, 532.7194, 572.8767, 638...	[[402.71667, 424.5876, 421.76343, 429.39407, 4...	[[1825.2471, 1949.5663, 1927.596, 1989.3441, 2...
2004-01-25	[[0.23110248, 0.19293109, 0.1849323, 0.1849323...	[[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,...	[[0.09986393, 0.098094486, 0.092537396, 0.0925...	[[0.0709232, 0.05904688, 0.0822366, 0.0822366,...	[[0.47878838, 0.7776081, 0.40051222, 0.4005122...	[[534.63955, 556.26385, 512.43384, 510.8002, 5...	[[377.45728, 395.96835, 380.23203, 374.76712, ...	[[1765.5181, 1916.2206, 1784.8286, 1720.2325, ...
2004-02-02	[[0.18155847, 0.17532943, 0.19496985, 0.195421...	[[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,...	[[0.054145418, 0.056791425, 0.083795756, 0.056...	[[0.065331735, 0.08083127, 0.08583814, 0.06492...	[[0.337094, 0.51916796, 0.5625676, 0.36588576,...	[[485.49188, 463.3164, 469.01636, 478.11493, 5...	[[317.10767, 306.41827, 303.78357, 308.62192, ...	[[1327.9946, 1264.3059, 1282.7346, 1333.1127, ...

Figure 5: The resulted sampled data

(4) **Applying the Self organizing Map SOM:** The SOM's aim is to summarize the information when reducing the dimensions on the map. Thus the projected neurons on each dimension will allow us to construe the data type. The phytoplankton concentration will then be determined as we cluster the results obtained by the SOM algorithm. Before we apply the algorithm, we proceed first by flattening each image. The data in hand is then a bidimensional array of size (54*91,1096)=(4914, 1096), where the rows are the pixels, and the columns correponds to the value of the pixel in each day of each variable (see figure 6)

	19HF2004-01-01T00:00:00.000000000	19HF2004-01-09T00:00:00.000000000	19HF2004-01-17T00:00:00.000000000	19HF2004-01-25T00:00:00.000000000
0	-1.389885	0.106593	-2.292215	-0.571749
1	-2.492894	0.106593	-3.653084	-0.854275
2	0.650884	0.106593	-2.450657	-2.151569
3	1.370050	0.106593	-2.010465	-1.795525
4	0.292533	0.106593	-1.741147	-0.969463

Figure 6: The SOM input data

The applied self-organizing map parameters were mainly the neurons dimensions, so to choose the neurons bidimensional mapsize , we proceed with a grid search where the number of neurons varies from 10 to 50 with a step of 5. As the optimum mapsize is found (the map that correspond to a compromize between the quantization error [1] ,topographic error [2], and the silhouette criterion [3]) which is a 15 × 30 map. We can infer this result from the figure 7 below:

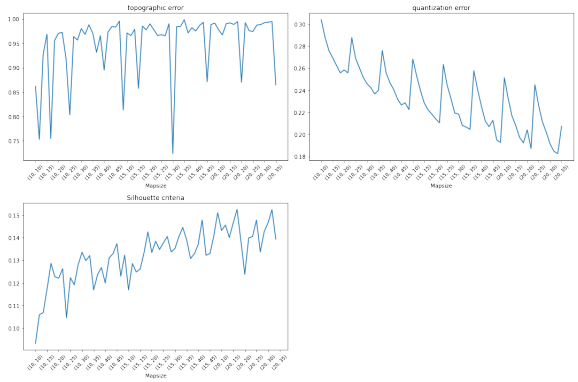

Figure 7: Plots of topographic error, quantization error, and silhouette criterion

We choose to train the topological card on 10 epochs, with a pca initialization.

(5) **Unsupervised clustering using HAC:** A hierarchical ascending clustering is then applied to the 525 neurons. As the figure 8 shows, we obtain 5 significant classes.

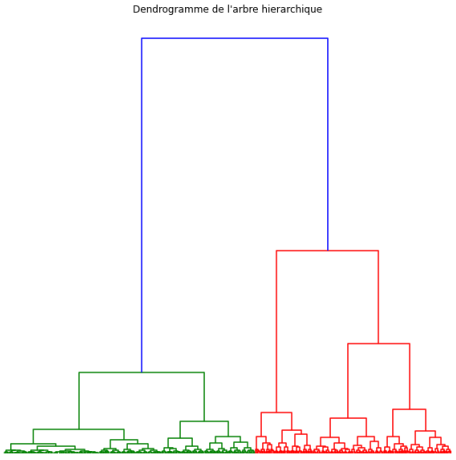

Figure 8: Dendrogram of the applied CAH

[1] Quantization error is the average difference of the input samples compared to its corresponding winning neurons (BMU). It assesses the accuracy of the represented data, therefore, it is better when the value is smaller

[2] Topographical error assesses the topology preservation . It indicates the number of the data samples having the first best matching unit and the second best matching unit being not adjacent.

[3] The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from 1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

5 RESULTS

The results of the "5 classes" CAH clustering of the 525 neurons is shown by the figure below:

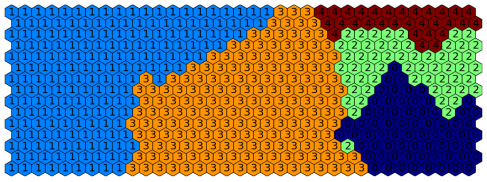

Figure 9: Clustering of the neurons

To study more the meaning of these classes, we carry out a summary of the values of these neurons, which is demonstrated by the below boxplot:

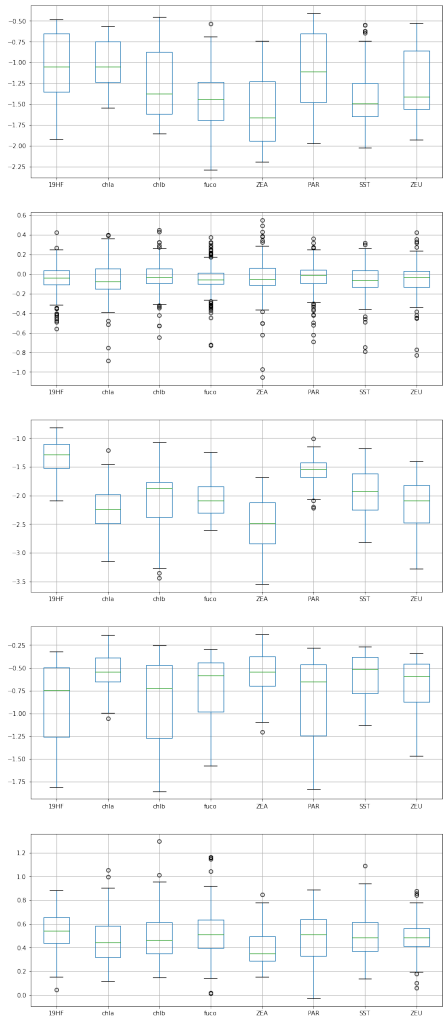

Figure 10: Boxplot of each cluster relatively to each of the eight variables

6 DISCUSSION

The previous results lead us to this part of the paper, where we discuss the clusters. Gathering up all the information that have been said previously we can then describe each class as follows:

- **Cluster 0:** This cluster corresponds to low proportions of pigment variables as well as physical variables relatively to chla, which means that this region correponds to high concentration of chla, slightly lower concentration of chlb and 19HF, and lower concentration of Fuco.
- **Cluster 1:** This cluster, a wide region in the map, correponds to very low proportions of physical values and pigments relatively to chla (median near to zero), which means that it presents very high values of chla or low values of the other variables. This cluster is then charachterized by a second type of region where the phytoplancton organisms develops in a high chla concentrated environment, and low sea surface temperature.
- **Cluster 2:** This region is characterized by warmer temperature and lower proportions of ZEU and ZEA.
- **Cluster 3:** This cluster designate higher range of values of proportions of chlb, 19HF and PAR relatively to chla. This region is then characterized by a high variability of pigments and PAR.
- **Cluster 4:** Corresponds to a wide area on the map, this cluster corresponds to higher proportions of variables, indicating that phytoplancton is barely present in the area.

Due to their geographical location ,ocean characteristics and high concentration of chlorophyll (a) and (b) class-0, class-1 and class-3 corresponds to high probability environment for the phytoplankton organisms. As the presence of phytoplankton is detectable from the chlorophyll a concentration, those clusters seems to coincide with a convenient environment for the production of diverse phytoplankton types.

7 CONCLUSION

A self organizing map, in association with the CAH clustering have provided an effecient way to study the phytoplankton distribution along the ocean. In fact, we could classify, in an unsupervized way, the parts of images provided by the original data set into 5 clusters where the physical variables are slightly variable, on the contrary of the pigment variables relatively to chla. Thus we could come to the conclusion that the higher of values of pigments are, the more likely is to produce phytoplankton in a warm area of the sea.

REFERENCES

[1] F. D'Ortenzio , M. Ribera d'Alcal'. *On the trophic regimes of the Mediterranean Sea: a satellite analysis* Biogeosciences, 6, 139–148, 2009

[2] O. Farikou, S. Sawadogo, A. Niang, J. Brajard, C. Mejia, M. Crépon and S. Thiria *Multivariate Analysis of the Senegalo-Mauritanian Area by Merging Satellite Remote Sensing Ocean Color and SST Observations* Research Journal of Environmental and Earth Sciences 5(12): 756-768, 2013

[3] Roy El Hourany, Marie Abboud-Abi Saab, Ghaleb Faour, Carlos Mejia, Michel Crépon, and Sylvie Thiria *Phytoplankton Diversity in the Mediterranean Sea From Satellite Data Using Self-Organizing Maps* , JGR OCEAN Research article

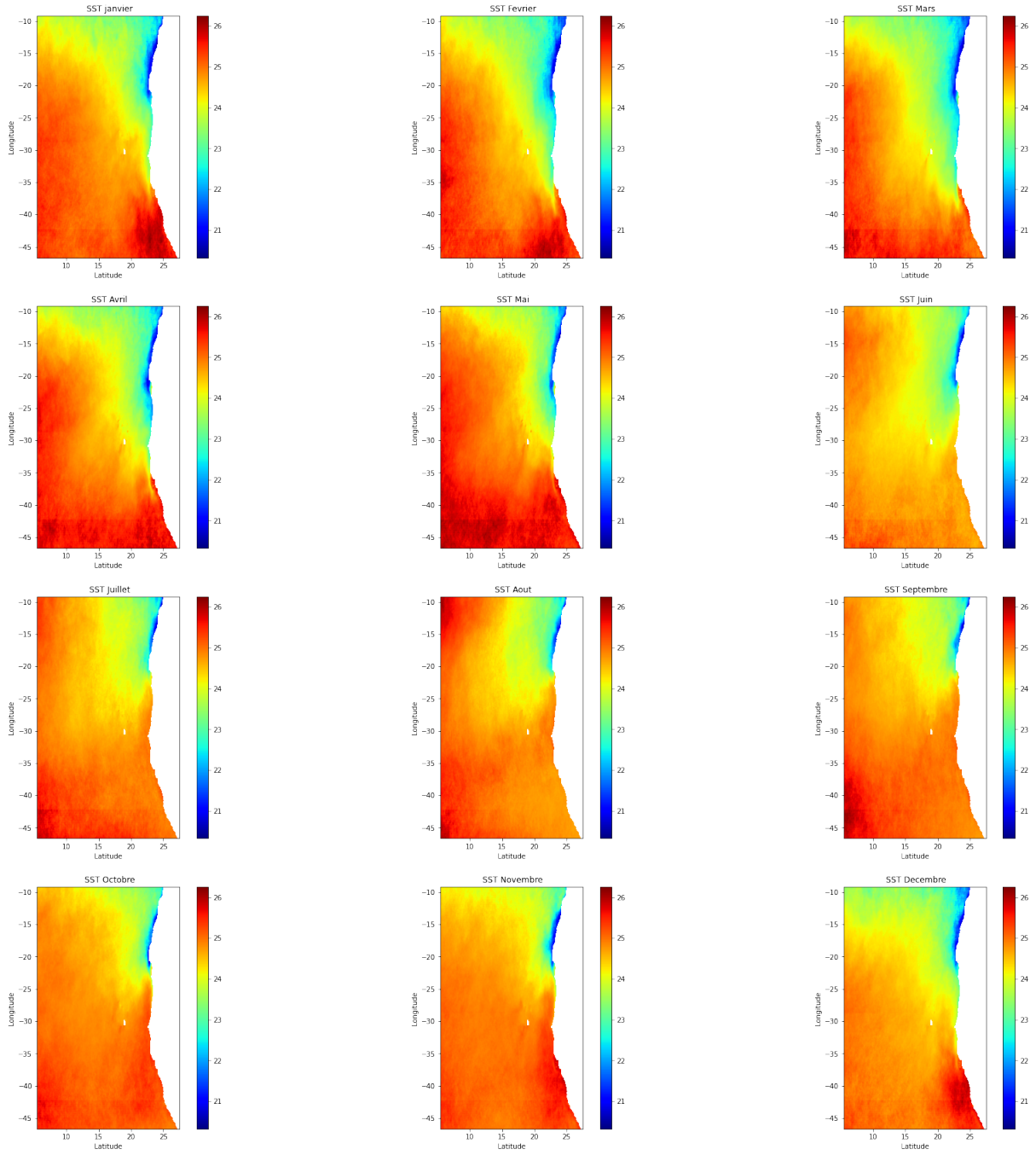

Figure 11: Image of the sea surface temperature SST (° C) averaged over the months

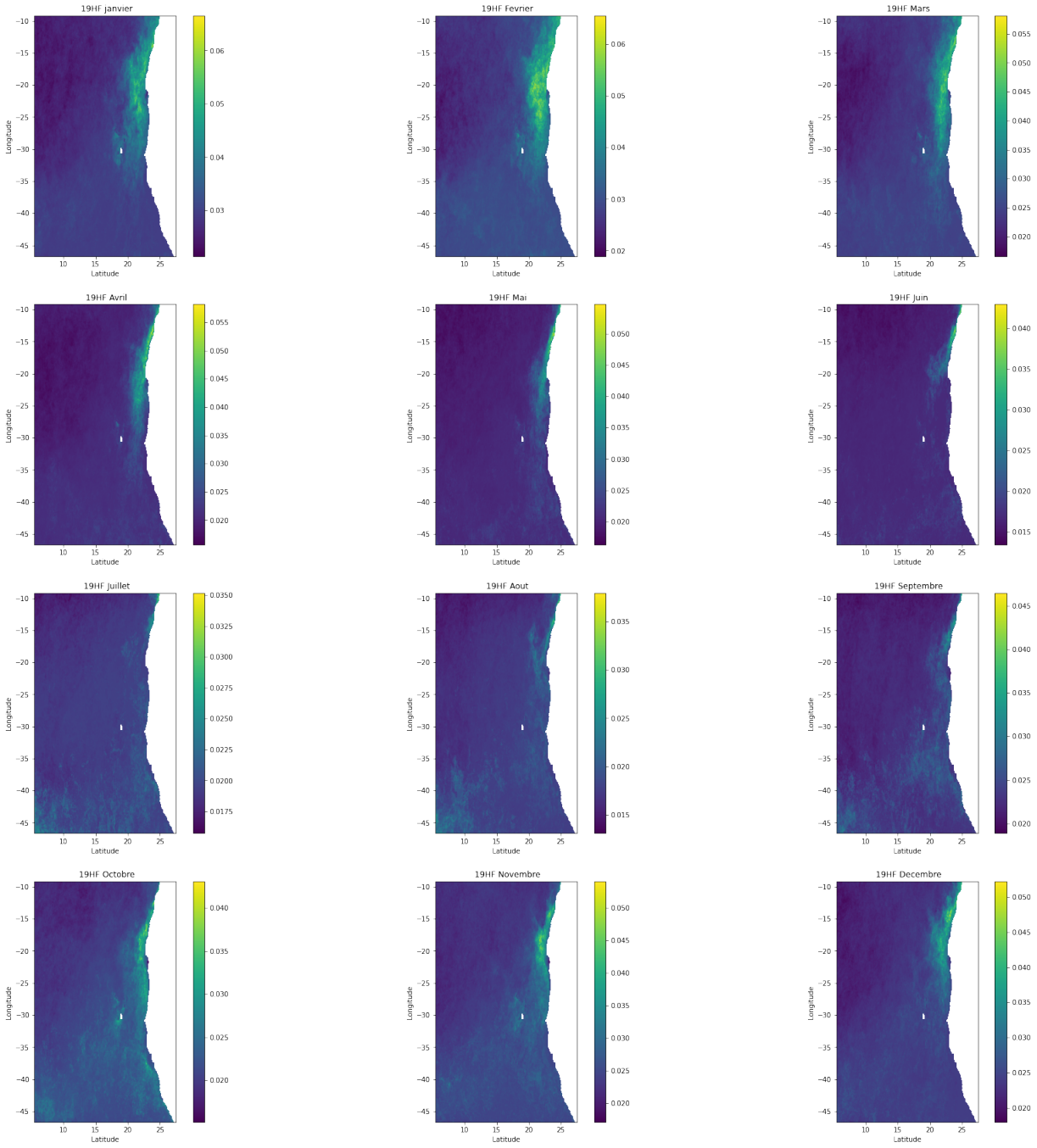

Figure 12: Image of the 19-Hexanoyloxy-fucoxanthin averaged over the months

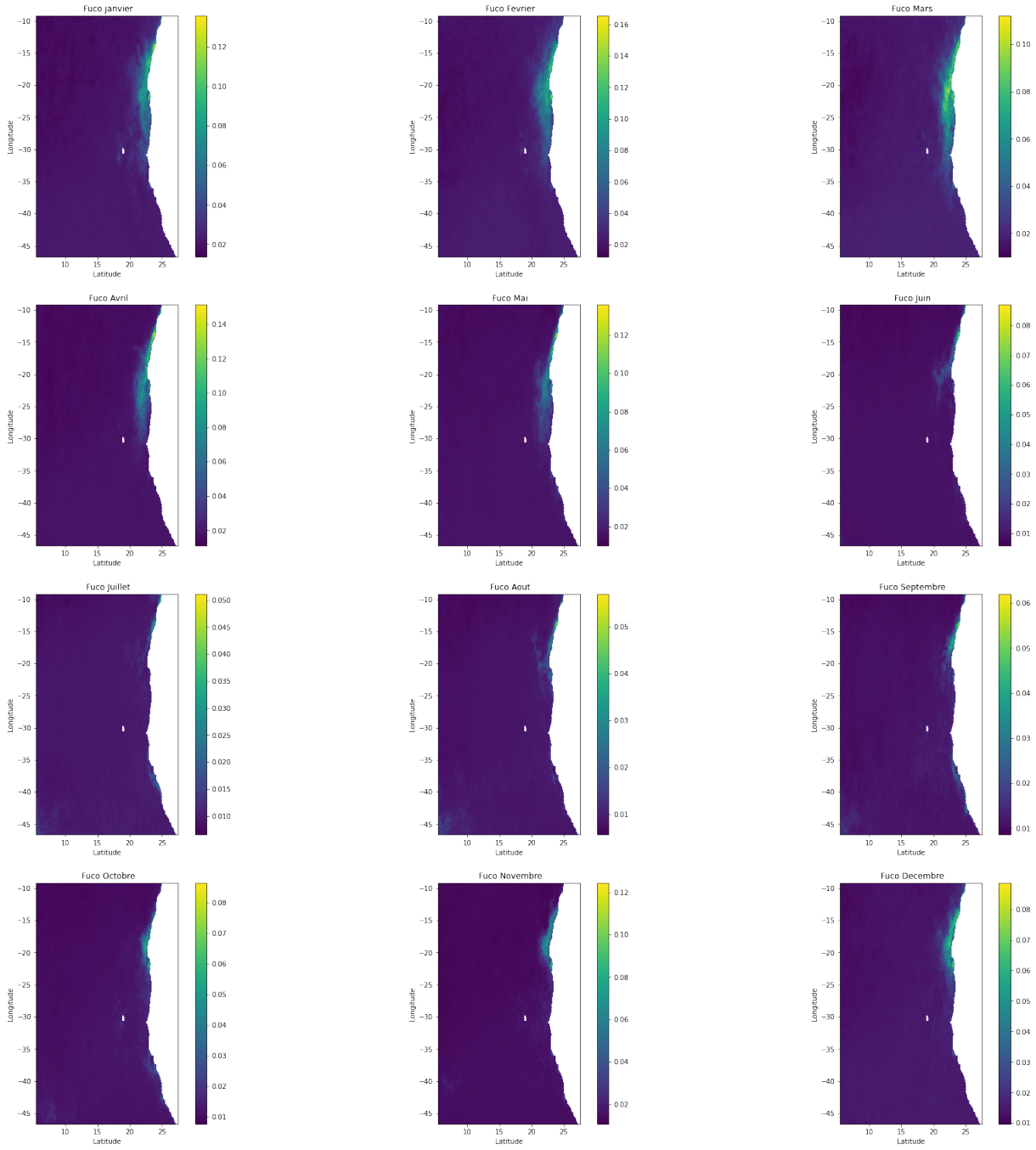

Figure 13: Image of the Fuco averaged over the months

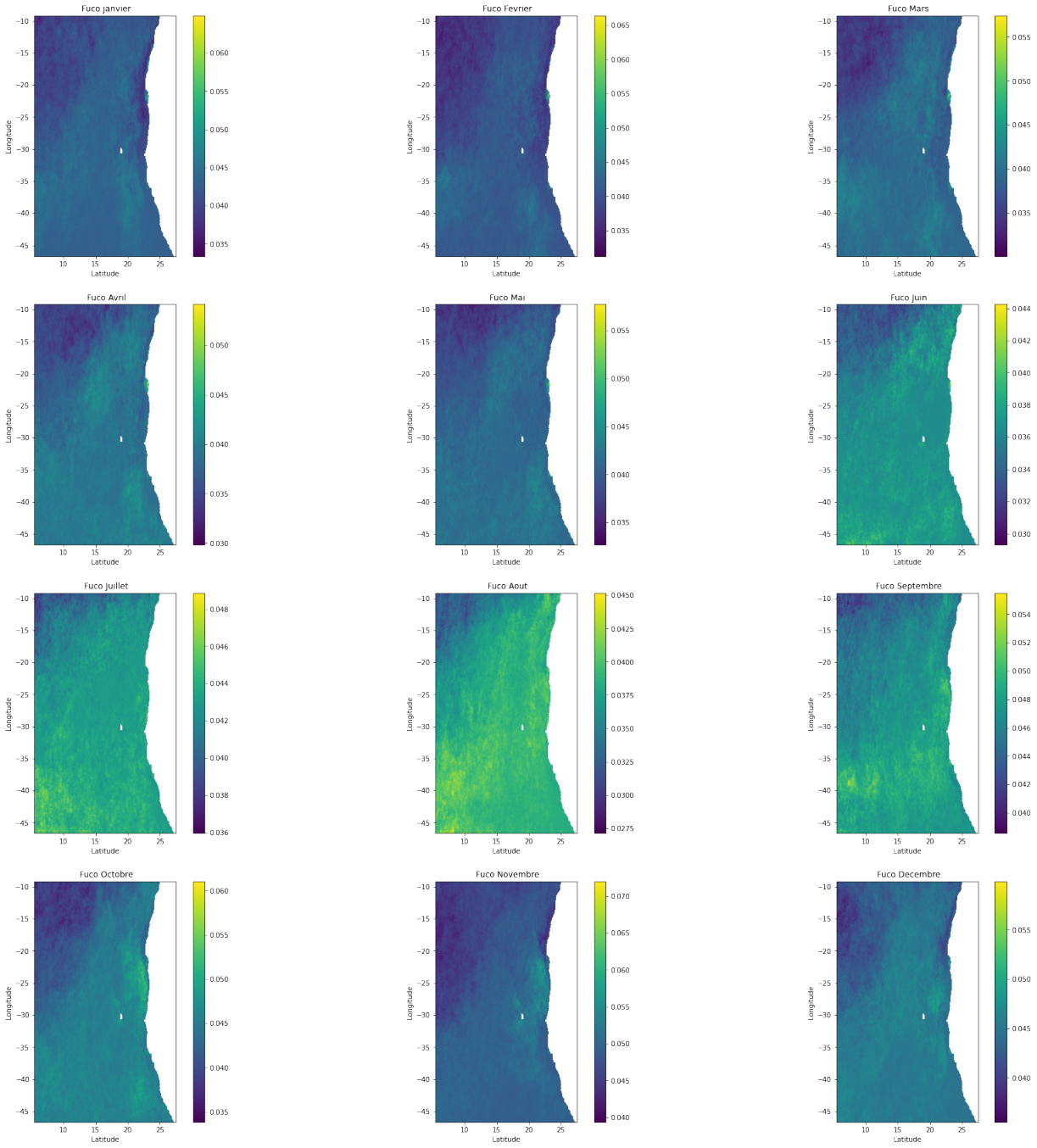

Figure 14: Image of the Zea averaged over the months

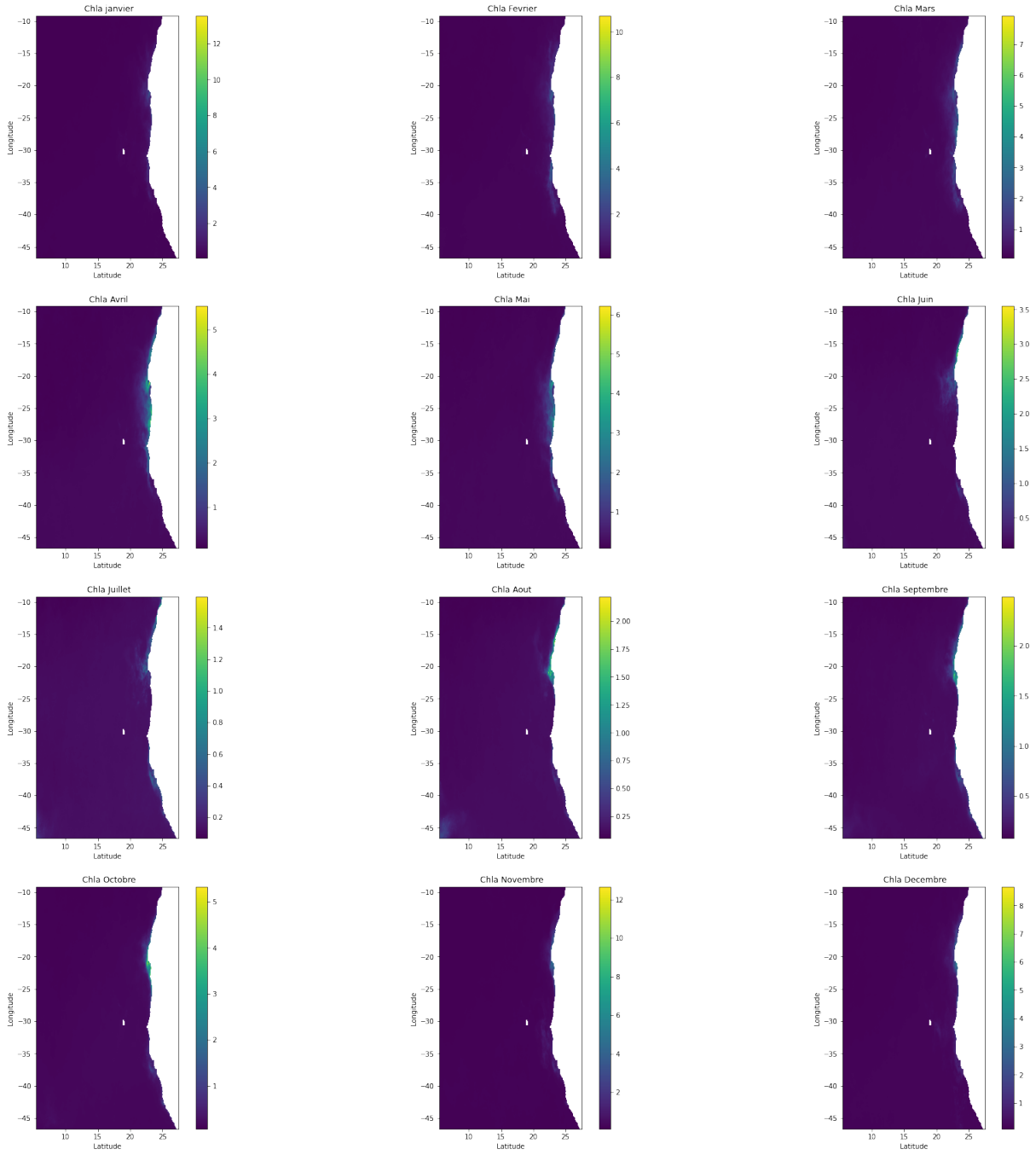

Figure 15: Image of the chla averaged over the months

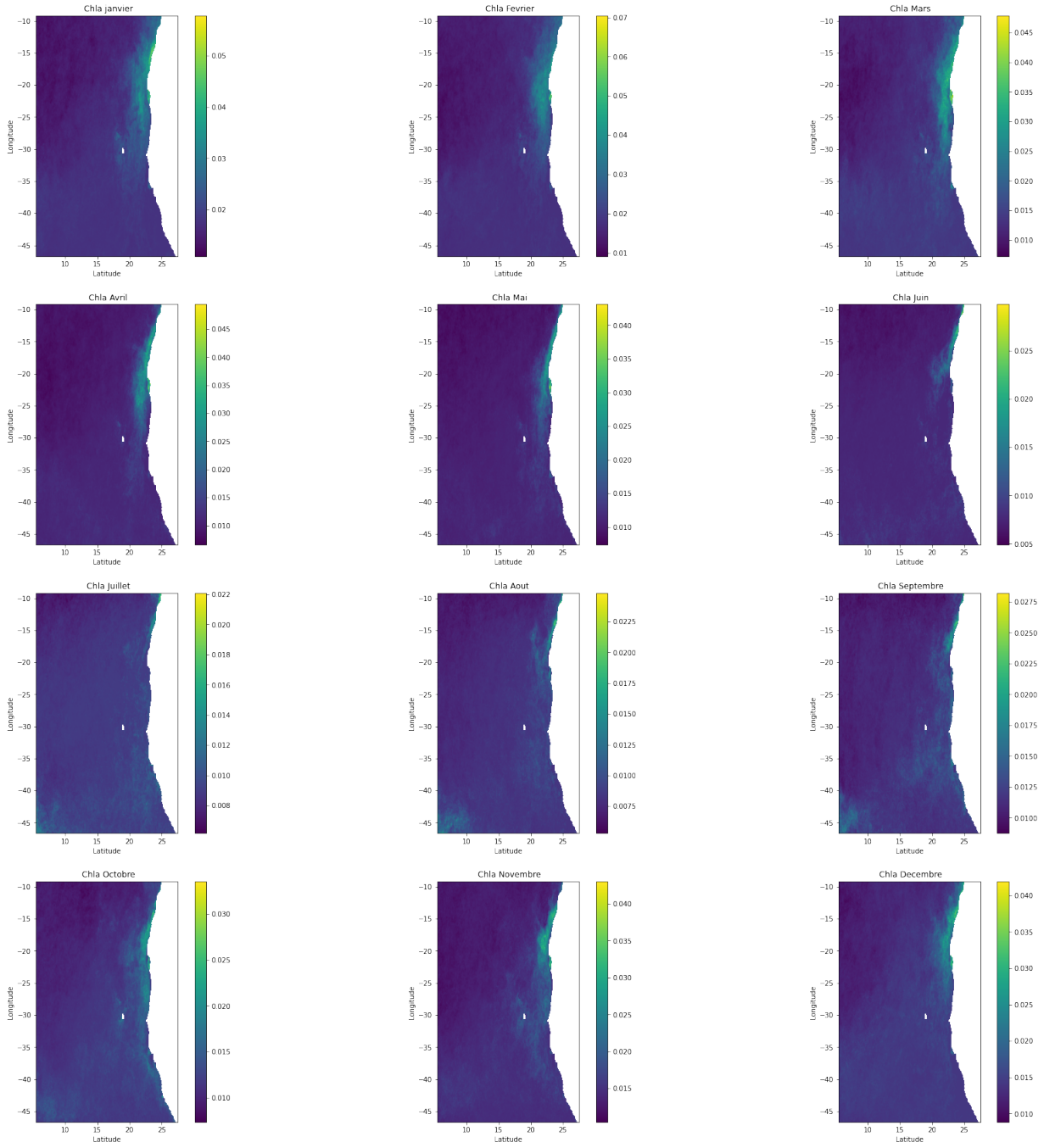

Figure 16: Image of the chlb averaged over the months

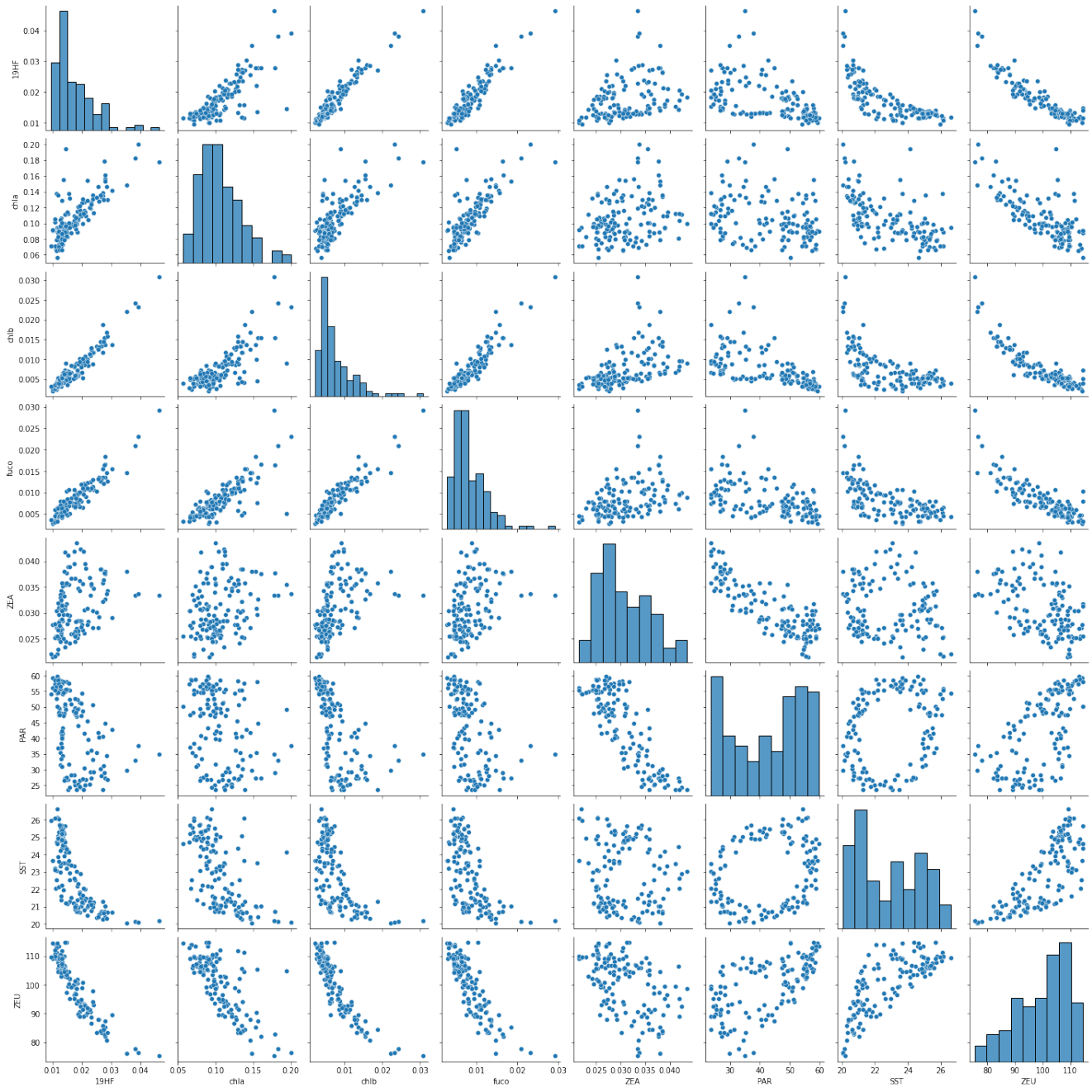

Figure 17: Variables pairplot