Supplement of Ocean Sci., 19, 973–990, 2023
https://doi.org/10.5194/os-19-973-2023-supplement

Ocean Science Open Access EGU

Supplement of

Global variability of high-nutrient low-chlorophyll regions using neural networks and wavelet coherence analysis

Gotzon Basterretxea et al.

Correspondence to: Gotzon Basterretxea (gotzon@imedea.uib-csic.es)

Supplementary material

SOM analysis

SOM is a subtype of artificial neural network that uses an unsupervised machine learning algorithm to process and extract hidden structures in large amount of data. The learning process seeks to identify low-dimensional features in high-dimensional data according to a similarity measure while preserving the topological properties of the input space. This technique provides some advantages as compared to other widely used statistical approaches such as proper orthogonal decomposition, principal component analysis or *k*-means, for several reasons: i) it allows an appropriate exploratory data analysis of high dimensional data, such as the long time series of global Chl and nutrients maps; ii) identifies and extracts underlying nonlinear regressions in the data, for example using step-like or Gaussian functions for the neighborhood relationship; iii) it is an efficient method for feature extraction and afterwards classification, providing the patterns that explain the overall behavior of the system; and iv) because of the preservation of the topology, SOM has an additional benefit of ordering the patterns in the neural network according to their similitude, enabling an easy visualization and interpretation of large number of patterns (Brunton et al., 2020; Liu et al., 2006). The SOM is a powerful machine learning technique that has been widely used in oceanographic research (Richardson et al., 2003; Liu and Weisberg, 2005; Liu et al., 2006).

The SOM algorithm is basically composed of two main steps: initialization and training. In the initialization the architecture of the neural network used in this study is set in a hexagonal map lattice of neurons, or units, in order to have equidistant neurons (to avoid anisotropy artifacts). Each unit is represented by a weight vector with a number of components equal to the dimension of the input data, i. e. number of rows or number of columns in the Chl and NO3 matrices, depending on whether the analysis is performed in the temporal or in the spatial domain. We use an initial network composed of units of random values. In the training process, the initial neural network is transformed by iteratively presenting the input data. In each successive iteration the neuron, or unit, with the greatest similarity (excited neuron), called Best Matching Unit (BMU) is updated by replacing their values with the Chl and NO3 values of the input sample data. The similarity is estimated by computing the Euclidean distance between components of the input sample and components of the weight vector of the unit. The unit most similar to the input sample is the one with the minimum distance. In the learning process, Chl and NO3 values of the neighboring neurons of the excited neuron are also updated replacing their values with values determined by a neighborhood function. In this way, the topological neighbors of the BMU are also updated through the neighborhood function. In this study, we use the imputation batch training algorithm (Vatanen et al., 2015) and a Gaussian neighboring function. After repeating the training process a number of times until a stable convergence of the map is achieved, we obtain a neural network with the final NO3:Chl patterns. Therefore, the resulting patterns will exhibit some similarity because the SOM process assumes that a single sample of data (input vector) contributes

to the creation of more than one pattern, as the whole neighborhood around the best-matching pattern is also updated in each step of training. It also results in a more detailed assimilation of particular features appearing on neighboring patterns, if the information from the original data enables to do so.

The size of the neural network (number of neurons) depends on the number of samples and on the complexity of the patterns and an optimal choice is important to maximize the quality of the SOM. In the present study, the map size is set to be [4 x 3] with 12 neurons for the time domain analysis, and a [3 x 3] neural network used in the spatial domain. Using larger map sizes, the patterns are slightly more detailed and more regions of a particular variability emerge, but the occurrence of probability of the patterns decreases, without affecting the results noticeably (Basterretxea et al., 2018; Hernandez-Carrasco and Orfila, 2018). If a reduced neural map, such as [2 x 2] is used, patterns are concentrated together with the occurrence probability in few rough patterns but increasing, in this case, the topological error.

Basterretxea G., Font-Muñoz J.S., Salgado-Hernanz P.M., Arrieta J. et al. (2018). Patterns of chlorophyll interannual variability in Mediterranean biogeographical regions. Remote Sens. of Environ. 215, 7-17. DOI: 10.1016/j.rse.2018.05.027.

Brunton S.L., Noack B.R., Koumoutsakos P. (2020). Machine Learning for fluid mechanics. Annu. Rev. Fluid Mech. 52:1. DOI: 10.1146/annurev-fluid-010719-060214

Hernández-Carrasco I., Orfila A. (2018). The role of an intense front on the connectivity of the Western Mediterranean Sea: The Cartagena-Tenes front. J. Geophys. Res. 123, 4398-4422. DOI: 10.1029/2017JC013613.

Liu, Y., and R.H. Weisberg (2005), Patterns of ocean current variability on the West Florida Shelf using the self-organizing map, *Journal of Geophysical Research,* 110, C06003, DOI: 10.1029/2004JC002786.

Liu Y., Weisberg R.H., Mooers C.N.K. (2006). Performance evaluation of the self-organizing map for feature extraction. J. Geophys. Res. 111, C05018, DOI: 10.1029/2005JC003117.

Richardson, A. J., Risien, C., Shillington, F. A. (2003). Using self-organizing maps to identify patterns in satellite imagery, Progress in Oceanography, 59 (2–3). 223-239, DOI: 10.1016/j.pocean.2003.07.006.

Vatanen T., Osmala M., Raiko T., Lagus K. et al. (2015). Self-organization and missing values in SOM and GTM. Neurocomputing. 147, 60–70. DOI: 10.1016/j.neucom.2014.02.061

Wavelet coherence analysis

Wavelet transform of a time series x_n (W^X) performs a time-frequency domain decomposition of the time series by estimating its spectral characteristics as a function of time (Torrence and Compo, 1998). It is preferred to classical Fourier analysis (Brockwell and Davis, 1987; Priestley, 1992) since it resolves non-stationary signals, therefore being well suited for identifying periodic phenomena with changing spectra. Here we used a Morlet wavelet transform with an adimensional frequency $\omega_0=6$ (i.e. it contains 6 complete cycles of the temporal scale that is being analyzed), as the wavelet base function because it is adequate to be localized in both time and frequency space and therefore to properly assess changes in the wavelet amplitude over time (Torrence and Compo, 1998). To distinguish a signal from noise a threshold above the 95% confidence interval of a red-noise spectrum was used.

Using the cross-wavelet transform (XWT) we determine the cyclic changes at each of the HNLC regions and their relationship with global forcings mentioned above. The XWT of two time-series x_n and y_n indicates common power and relative phase in the frequency-time domain and is given by:

$$W^{XY} = W^X W^{Y*}, \qquad \text{Eq. 1}$$

where * represents the complex conjugate. $|W^{XY}|$ is the cross-wavelet power and arg (W^{XY}) is the relative phase between the two series (shown in the figures as arrows).

Finally the wavelet coherence value at each time point was calculated as the absolute value squared of the smoothed cross-wavelet spectrum, normalized by the product of the smoothed wavelet individual spectra, for each scale (Torrence and Compo, 1998). To quantify the degree of coherence of cross wavelet transform in time-frequency space we use the wavelet coherence coefficient defined as follows (Torrence and Webster, 1999; Grinsted et al., 2004):

$$R_n^2(s) = \frac{\left|S\left(s^{-1}W_n^{XY}(s)\right)\right|^2}{S\left(s^{-1}\left|W_n^X(s)\right|^2\right) * S\left(s^{-1}\left|W_n^Y(s)\right|^2\right)}, \qquad \text{Eq. 2}$$

where R^2 takes values between 0 (no correlation) and 1 (perfect correlation). R^2 can be interpreted as a localized correlation coefficient in the frequency-time domain. S is a smoothing operator in the time-frequency (scale) domain. The smoothing process is necessary to remove the singularities in the wavelet power spectra of the time series. In the time domain, the smoothing was implemented as a weighted moving average function, with weights defined by a Gaussian function, and a width equal to the wavelet size in the time domain. In the frequency domain, a boxcar filter with a width equal to the scale decorrelation length was used; for the Morlet wavelet this is 0.6 (Torrence and Compo, 1998).

Monte Carlo simulations based on two uniform white noise time series are used to determine the significance level for the wavelet coherence. In this study we use the MATLAB software package (Grinsted et al., 2004) for wavelet coherence analysis. It should be noted that cross-wavelet analysis does not establish causative relationships but

only allows identifying possible linkages between variables through the synchrony of their time series.

Brockwell P., Davis R. (1987). Time Series: Theory and Methods. Springer. NY. DOI: 10.1007/978-1-4899-0004-3

Grinsted A., Moore J.C., Jevrejeva S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. Nonlinear Process. Geophys. 11, 561-566. DOI: 10.5194/npg-11-561-2004

Priestley M. (1992). Spectral Analysis and Time Series, Academic Press. ISBN: 9780125649223

Torrence C., Compo G. (1998). A practical guide to wavelet analysis. Bull. Amer. Meteor. Soc. 79, 61–78. DOI: 10.1175/1520-0477

Torrence C., Webster P.J. (1999). Interdecadal changes in the ENSO-monsoon system. Journal of Climate, 12, 2679-2710. DOI: 10.1175/1520-0442

Supplementary figures

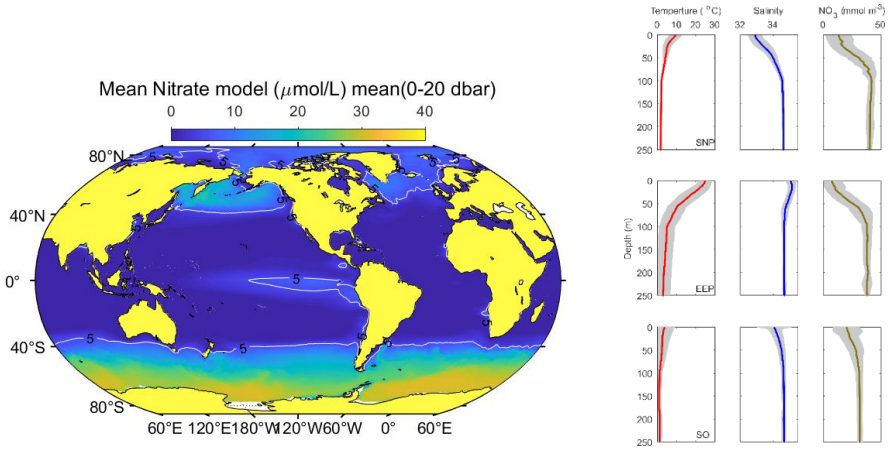

Fig. S1. Mean global surface NO3 concentrations obtained from PISCES model (Aumont, 2015) and mean measured vertical temperature, salinity and NO3 profiles (0-250m) for each HNLC region from WOD18 (Boyer et al., 2018; https://www.nodc.noaa.gov/).

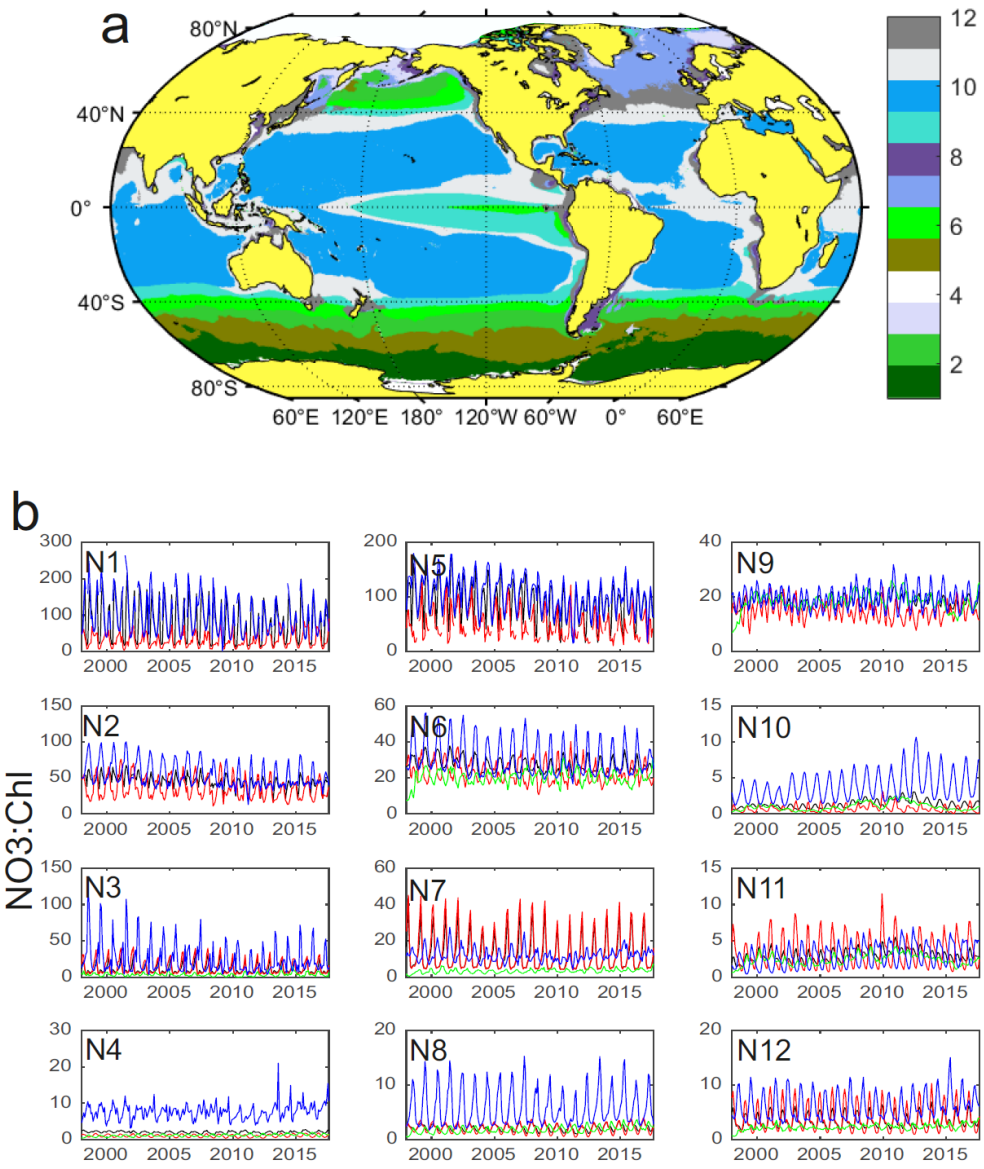

Figure S2. a) Regionalization obtained from coupled [4 x 3] SOM analysis applied to NO3:Chl values (in mmol/mg) and b) associated characteristic time series of NO3: Chl variability. The black line corresponds to the time series obtained from the SOM analysis. The red, green, and blue lines are the averages of the NO3: Chl values over the northern, equatorial, and southern sub-regions. Note that some HNLC regions have been renamed for the simplified map shown in figure 3.

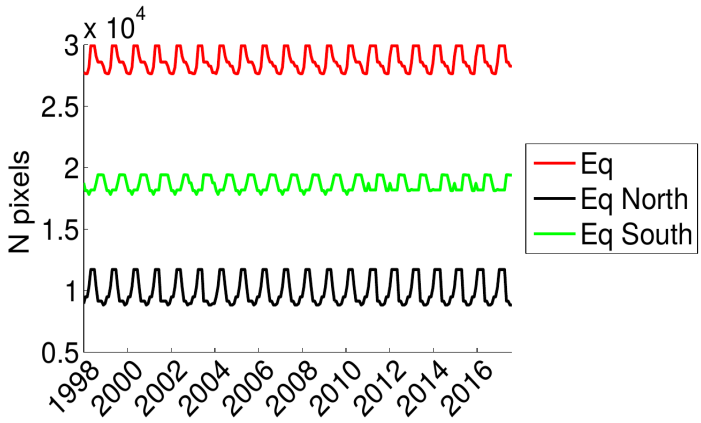

Fig. S3a. Time evolution of the total number of pixels covering each area: Whole equatorial region (in red). Pixels in the equatorial region which is located in the northern hemisphere are shown in black and those in the southern hemisphere in green. Note that while the mean value in the Southern hemisphere is larger, the northern region is more variable.

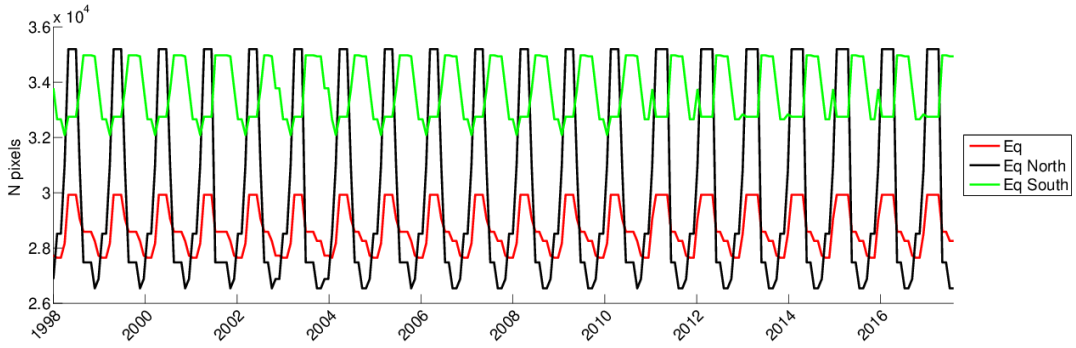

Fig. S3b. Time evolution of the total number of pixels covering each area scaled to the same range of values for better comparison: Whole equatorial (in red). Pixels in the equatorial region which is located in the northern hemisphere are shown in black and those in the southern hemisphere in green.