Biogeosciences Discuss., 6, 3373–3414, 2009
www.biogeosciences-discuss.net/6/3373/2009/

Biogeosciences Discussions

Biogeosciences Discussions is the access reviewed discussion forum of *Biogeosciences*

Estimating the monthly pCO_2 distribution in the North Atlantic using a self-organizing neural network

M. Telszewski[1], A. Chazottes[2], U. Schuster[1], A. J. Watson[1], C. Moulin[2], D. C. E. Bakker[1], M. González-Dávila[3], T. Johannessen[4], A. Körtzinger[5], H. Lüger[6], A. Olsen[4,8,9], A. Omar[4], X. A. Padin[7], A. Ríos[7], T. Steinhoff[5], M. Santana-Casiano[3], D. W. R. Wallace[5], and R. Wanninkhof[6]

[1]School of Environmental Sciences, University of East Anglia, Norwich, UK
[2]L'Institut Pierre-Simon Laplace/Laboratoire des Sciences du Climat et de l'Environnement, Centre National de la Recherche Scientifique - Commissariat à l'Énergie Atomique, Gif-sur-Yvette, France
[3]Department of Marine Chemistry, Universidad de Las Palmas de Gran Canaria, Las Palmas, Gran Canaria, Spain
[4]Geophysical Institute, University of Bergen, Bergen, Norway

[5] Leibniz Institute of Marine Sciences, Kiel, Germany
[6] Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, Florida, USA
[7] Instituto de Investigacions Marinas (Consejo Superior de Investigaciones Científicas), Vigo, Spain
[8] Bjerknes Centre for Climate Research, UNIFOB AS, Bergen, Norway
[9] Marine Chemistry, Departement of Chemistry, University of Göterborg, Göteborg, Sweden

Received: 10 March 2009 – Accepted: 22 March 2009 – Published: 30 March 2009

Correspondence to: M. Telszewski (m.telszewski@uea.ac.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Here we present monthly, basin-wide maps of the partial pressure of carbon dioxide (pCO_2) for the North Atlantic on a 1° latitude by 1° longitude grid for years 2004 through 2006 inclusive, constructed using a neural network technique which reconstructs the non-linear relationships between 3 biogeochemical parameters and marine pCO_2. A self organizing map (SOM) neural network has been trained using the SeaWiFS-MODIS chlorophyll *a* concentration, the NCEP/NCAR reanalysis sea surface temperature, and the FOAM mixed layer depth. 389 000 such triplets were used. The trained SOM was labelled with 137 000 underway pCO_2 measurements collected in situ during 2004, 2005 and 2006 in the North Atlantic, which span the range of 208 and 437 μatm. The root mean square (RMS) deviation of the neural network fits from the data is 11.55 μatm, which equals to just above 3 per cent of an average pCO_2 value in the in situ dataset. The seasonal pCO_2 cycle as well as the interannual variability estimates in the major biogeochemical provinces is presented and spatial and temporal variability of the estimated fields is discussed. High resolution combined with basin-wide cover makes the maps a useful tool for several applications such as monitoring of basin-wide air-sea CO_2 fluxes or improvement of seasonal and interannual marine CO_2 cycles in future model predictions. The method itself is a valuable alternative to traditional statistical modelling techniques used in geosciences.

1 Introduction

Globally, natural oceanic sink has absorbed around 30 per cent of the total anthropogenic carbon dioxide (CO_2) emissions to the atmosphere since the beginning of the industrial era (Sabine et al., 2004). This natural buffer slows the effects of anthropogenic interference with the global carbon cycle. The North Atlantic Ocean being a highly biogeochemically dynamic basin and one of the strongest sinks of carbon in the world's oceans (Takahashi et al., 2002), plays an important role in the world's car-

BGD

6, 3373–3414, 2009

bon cycle. Understanding of the future behaviour of the global natural carbon sinks and sources as well as related climate development can only be obtained given robust understanding of the current distribution of carbon sink/source regions.

The magnitude of the ocean sink can be determined using air-sea flux estimates based on in situ measurements of the sea surface partial pressure of CO_2 (pCO_2). However, while atmospheric pCO_2 shows relative homogeneity, marine pCO_2 varies radically both temporarily and spatially (Sarmiento and Gruber, 2002). Monitoring the marine pCO_2 distribution on monthly to interannual time-scales is thus crucial for further understanding of the global carbon cycle in the context of current climate dynamics. Due to technical as well as financial restrictions, in situ measurements of marine pCO_2 are sparse even in the relatively well sampled North Atlantic Ocean. Over the last decade though, technical improvements and cooperation with the shipping industry have allowed for installing several autonomous underway systems on board of commercial vessels routinely crossing the ocean basin. Those instruments perform quasicontinuous measurements, offering temporal and spatial cover which allows for regional analysis of the highly variable spatial and temporal distribution of pCO_2 (e.g. Cooper et al., 1998; Lefèvre et al., 2004; Lüger et al., 2004, 2006; Corbière et al., 2007; Schuster and Watson, 2007; Olsen et al., 2004, 2008; Schuster et al., 2008). Most of these authors suggest that the strength of the North Atlantic sink has decreased over the last decade, with the decline especially significant (up to 50%) in the northern part of the basin. This change indicates that increasing fraction of the anthropogenic emissions remains in the atmosphere, which is consistent with some of the recent modelling results. For instance, Canadell et al. (2007) suggest that around 10 per cent of the recent (2000–2006) rise in the atmospheric CO_2 concentrations can be attributed to the weakening of the ocean sink.

Despite the huge community effort to increase the network of in situ measurements in the North Atlantic, the cover still remains unevenly distributed in time and space. The regional character of the existing estimates poses difficulties when generalized into the entire basin, therefore a robust and reliable method to spatially and temporarily inter-

polate available measurements of marine pCO_2 has been long sought (e.g. Lefèvre et al., 1999, 2005; Takahashi et al., 2002, 2008; Olsen et al., 2003; Jamet et al., 2007; Chierici et al., 2008).

In the work presented here, we seek to map oceanic pCO_2 in the North Atlantic at a monthly timescale. We use an artificial neural network (NN), a powerful non-linear modelling tool for mapping performance (Dreyfus, 2005). Neural networks where first used extensively by the pattern recognition community 20–30 years ago (Kohonen, 2001). Since then NN have made their way into geosciences and over the last decade there has been a significant increase in their application to environmental problems. They are now commonly used in atmospheric science (Cavazos, 1999; Hewitson et al., 2002; Niang et al. 2006), oceanography (Richardson et al., 2003; Liu et al., 2005 and 2006a; Reusch et al., 2007) and meteorology (e.g. Ali et al., 2007).

The term *artificial neural network* reflects a mechanistic connection to the processes found in the human brain and therefore generates some confusion. At present their data-processing algorithms are well-understood and may be used in parallel with traditional statistical tools. There are numerous NN types among which the Self Organizing Map (SOM) seems to gain the most attention as being well suited to study empirical relationships in geosciences. It appears to be a particularly powerful tool for the extraction and classification of features (clustering), such as trends between input variables or their relative distribution. The SOM is a "black-box" type of model. While its restrictions and limitations need to be considered, it has an essential advantage over more commonly used knowledge-based models, which are based on equations describing the physical, chemical and biological phenomena that control the quantity to be modelled. As opposed to the latter, the SOM technique is based solely on observations. No equation, whether empirical or theoretical, is generated. Instead, the SOM uses an unsupervised (no need for a priori description of the input – output relations) learning algorithm, enabling us to identify relationships among the state variables of the phenomena under analysis, where our understanding of those is insufficient to be fully described using mathematical equations, and where knowledge-based models appli-

cations are therefore limited. The SOM technique has already been successfully used to interpolate marine pCO_2 maps from in situ measurements by Lefèvre et al. (2005). These authors were able to capture a more complex distribution in the northern North Atlantic using SOM than they could using multiple linear regressions, and also the residuals determined through the validation against an independent subset of the data were smaller for SOM.

In the study presented here, we use 137 000 in situ pCO_2 measurements collected in the North Atlantic throughout 2004, 2005 and 2006 as part of the CarboOcean (http://www.carboocean.org/), an EU-funded Integrated Project and parallel US projects, combined with 389 000 satellite, reanalysis and assimilation data, which allows basin-wide, continuous mapping over extended periods of time. Here we present basin-wide, monthly pCO_2 maps for 3 consecutive years with a 1° latitude by 1° longitude resolution. We show the capacity of the method to synthesize coherent, spatial and temporal distribution patterns of marine pCO_2 fields in the North Atlantic, and propose the method to be used in conjunction with in situ data collection during future oceanic pCO_2 monitoring programs.

2 Data and methods

We hypothesise that sea-surface pCO_2 can be estimated through the SOM based multiple non-linear regression with three parameters (Eq. 1): the sea surface temperature (SST), the wind-mixed layer depth (MLD) and the abundance of photosynthesizing organisms in the surface ocean represented by chlorophyll *a* concentrations (CHL).

$$pCO_2 = \mathrm{SOM}(\mathrm{SST}, \mathrm{MLD}, \mathrm{CHL}) \quad (1)$$

During our SOM analysis three steps are taken in order to estimate basin-wide pCO_2 fields: first, no-pCO_2 data is used and an unsupervised training takes place; second, pCO_2 data measured in situ is used to label such preconditioned SOM map; third, the

trained and labelled SOM neurons are used to assign pCO_2 values to the geographical map of the North Atlantic.

2.1 An overview of the SOM setup

The SOM method is based on a statistical model, introduced by Kohonen (2001). It is a competitive learning method in which an algorithm learns to classify the samples by recognizing existing patterns. It is used to extract pertinent information from the statistical structure of the multivariate dataset. It performs a non-linear projection from the highly dimensional input data onto a usually two-dimensional (2-D) grid, as extensively described by Niang et al. (2003). The SOM analysis was carried out using the SOM Toolbox Version 2 (Vesanto, 2000) for Matlab, developed by the Laboratory of Information and Computer Science at the Helsinki University of Technology and freely available from http://www.cis.hut.fi/projects/somtoolbox. Visualizations of the resulting North Atlantic pCO_2 maps were done using additional procedures in Matlab. For general SOM procedures and parameter settings consult Liu et al. (2006b) and Vesanto et al. (2000). The SOM procedure adopted in this study is outlined below.

The SOM map consists of 2220 $\boldsymbol{i}$ units (often referred to as neurons) organized on a regular two dimensional (2-D) grid. Moderately sized maps (in relation to the training data set) are found to be the most efficient. Too many neurons do not reduce the data enough for extracting characteristic patterns. Too few neurons do not provide sufficient representation of patterns underlying the in situ observations. A flat sheet map shape (60×37) with a hexagonal regional lattice structure was chosen. Each neuron is represented by a multidimensional weight vector $\boldsymbol{m_i}$ (also called reference vector), made of as many components as the number of input variables in the training dataset (3 in this study, one for SST, MLD and CHL). Each component is also organized on a regular 2-D grid (called component plane) "underlying" the actual map. In this way the $\boldsymbol{i_{th}}$ element of the map is a 3 dimensional ensemble of the $\boldsymbol{i_{th}}$ elements of component planes, and can be described as a synthetic sample (Solidoro, 2007). All the values are linearly normalized to acquire an even weight distribution between variables. Ad-

ditionally CHL and MLD values are also log10 normalized (to minimize the influence which their spread throughout 4 and 3 orders of magnitude respectively, would otherwise have on the weight distribution). Initial values of the nodes of the component planes are determined prior to the training process. In a linear initialization (performed in this study) their weights are initialized by first calculating the two eigenvectors of the covariance matrix of the training data that have the largest eigenvalues. Then the linear span along the two eigenvectors determines the two dimensional matrix of the weight vectors. The horizontal and vertical directions of the weight vectors' matrix should be proportional to the two largest eigenvalues described above (Kohonen, 2001).

2.2 Training data set (SST, MLD, CHL)

The training data set consists of 3 subsets, one for each parameter. Basin-scale SST data were obtained from the NCEP/NCAR Reanalysis Project (Kalnay et al., 1996; http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis.html) at daily frequency and 2.5° latitude × 2.5° longitude resolution. The SST analyses were done weekly and interpolated linearly to daily values.

Basin-wide MLD estimates were obtained from the Forecasting Ocean Assimilation Model (FOAM, Meteorological Office, Exeter, UK; http://www.nerc-essc.ac.uk/godiva) at daily frequency and 1° latitude × 1° longitude resolution. FOAM model assimilates both in situ and remotely sensed ocean observations which are available in near real-time including temperature and salinity profiles at all depths from sea stations, Argo profiling floats and PIRATA moored arrays, as well as sea surface temperature from Voluntary Observing Ships' (VOS) reports, buoys, and the satellite mounted advanced high-resolution radiometer (AVHRR). The FOAM mixed layer depth used in this study is determined using the density based criterion (the depth where the density increase of 0.05 kg m^{-3} from the surface value occurs).

CHL data were obtained from Aqua-MODIS/SeaWiFS merged Level-3 Standard maps provided by NASA/GFSC/DAAC at weekly frequency and 9 km resolution (http://oceancolor.gsfc.nasa.gov). The use of the merged product was dictated by its 20%

and 24% improvement in cover in relation to the single mission products (SeaWiFS and MODIS respectively), when weekly maps are considered.

All three products (SST, MLD, CHL) offer almost full basin-wide cover for the years 2004 to 2006. All parameters were re-gridded onto weekly frequency and 1° latitude × 1° longitude resolution. The study area stretches latitudinally from 10.5° N to 75.5° N and longitudinally from 9.5° E to 75.5° W and it is hereafter called the North Atlantic.

We have excluded coastal (<500 m depth) and ice covered waters (<−1.8°C) from the training data set, which therefore consists of 389 000 pixels with assigned SST, MLD and CHL values (used in training) as well as additional information such as time and position, bottom-depth and other ancillary information which is used during mapping and analysis of the results.

The division into seasons was decided as follows: winter includes December, January and February, spring includes March, April and May, summer includes June, July and August and fall includes September, October and November.

2.3 The self organizing process – training the SOM

During the self-organizing process, 389 000 elements from the 3-dimensional training data matrix, referred to as input vectors $\boldsymbol{x}_p$, are presented to the SOM (Fig. 1a). The activation of each neuron's 3-dimensional weight vector, $\boldsymbol{y}_p$, for the presented input vector is computed. For a given input, the "winning" neuron (the one with the highest activation) is considered to be the one whose weight vector is the closest to the presented input vector in Euclidean distance, defined as:

$$D(x, y) = ((x_{SST} - y_{SST})^2 + (x_{MLD} - y_{MLD})^2 + (x_{CHL} - y_{CHL})^2)^{0.5} \quad (2)$$

The weight vector of the winning neuron is updated by adjusting it towards the input data vector by a certain fraction of the difference between the two, as indicated by a linear, monotonically time-decreasing learning rate function $\boldsymbol{\alpha}$. Thus the winner's activation will be even higher the next time a similar input vector is presented. In

addition to the winning-neuron, the weight vectors of units in the neighbourhood of the winner are also stretched towards the input sample, according to a neighbourhood function, $\boldsymbol{H}$, decreasing spatially away from the winner:

$$H_{ci}(t) = \alpha(t) \times \exp\left(d_{ci}^2 / 2\sigma(t)^2\right) \quad (3)$$

where $\sigma(t)$ is the neighbourhood radius at time t, and d_{ci} is the distance between map units c and i on the map grid. The neighbourhood radius $\sigma(t)$ decreases as a function of time along with the learning rate $\boldsymbol{\alpha}(\boldsymbol{t})$. The learning rule which incorporates such a neighbourhood function distinguishes the SOM from other vector quantization algorithms. The learning procedure leads to a topologically ordered mapping of the input vectors.

The training is performed in two phases: the rough training which accounts for an approximate ordination of neurons (map space) in the input data space, followed by the fine-tuning when the convergence between the map space and the data space is sought. The length of each phase, which depends on the ratio between the number of neurons and the number of input vectors, has been set to 20 and 15 repetitions, respectively (Fig. 1b). In the fine-tuning phase relatively small $\boldsymbol{\alpha}$ and $\boldsymbol{H}$ are set right from the beginning. The neighbourhood radius and the shape of the neighbourhood function have to be decided before the training starts. In this study the neighbourhood radius decreases from 8 to 2 neurons during the rough training and further to 0 during the fine-tuning phase. The shape of the neighbourhood function dictates the extent to which the neighbours of the winning-neuron are updated, and how it changes with increasing distance from the winning-neuron. A Gaussian shape has been used in this study.

By virtue of the neighbourhood function the winning-neuron is not a mean of the data it accounts for, but rather an expression of the local ordination of patterns extracted from the input data set (Dreyfus et al., 2005). Similar patterns are mapped onto neighbouring regions on the map, while dissimilar patterns are mapped further apart. After the training, each neuron becomes a synthetic sample with an associated 3 di-

mensional reference vector. It expresses the local relationship between its 3 components in the 2-D map space. Every weight vector (neuron) has a different combination of components, therefore the SOM can account for strong non-linearities in the real system. Figure 2a–c show the distribution of neurons within the input data space, visualizing how the SOM accounts for the non-linear relations between the components. The SOM is well equipped for such a complicated setup, e.g. the distribution of the neurons closely follows the data distribution, even in such an extreme case as MLD versus SST (Fig. 2b). Each neuron represents a relationship between two components within its neighbourhood. The SOM estimates are therefore based on 2220 relationships between each pair of components, and as such can resolve highly non-linear relationships.

2.4 Labelling data (pCO_2 with SST, MLD, CHL)

In order to estimate pCO_2 fields in the North Atlantic, the trained SOM neurons need to be labelled with the pCO_2 values. In the labelling set, in situ pCO_2 measurements are used, each coincided with SST, MLD and CHL according to their time and space coordinates. For the purpose of this work, we used a subset of the North Atlantic data set compiled under auspices of the CarboOcean, an ongoing EU-funded Integrated Project (http://www.carboocean.org). 137 000 data points were collected onboard several vessels routinely crossing the North Atlantic between June 2004 and October 2006 (Fig. 3). This provides a reasonable representation of the spatial and temporal variability of the sea surface pCO_2 field. We match each in situ pCO_2 measurement from the labelling set with the coinciding (in terms of time and space) pixel from the training data set. This way each labelling datum contains pCO_2 information and corresponding SST, MLD and CHL values.

2.5 Data distribution

The data available in the labelling set is not evenly distributed in time. Many more measurements are available in spring and summer than in fall and winter (Fig. 4). Difficulties related to sampling in stormy winter waters make the labelling data set skewed towards the calmer mid-year months. For the three years there are less than 3500 measurements between November and January, with November and December only sampled in 2005. Relatively few data are available for 2004. Apart from a major contribution in June, only 2 days of sampling in July and 7 days in October makes this year's input rather imbalanced.

Such an uneven distribution would make this data very difficult to analyse using traditional statistics. Most linear methods would be biased towards summer waters, and the exceptionally high volume of data from June 2004 would increase mapping discrepancies. In contrast, SOM estimates depend solely on the input values, hence ranges of input parameters are more relevant than the temporal and spatial distribution of the in situ measurements. The training data set offers wide ranges for all parameters, providing sufficient information about their variability as summarized in Table 1. SST varies between −1.8°C and 30°C, the depth of the mixed layer ranges from ∼10 m to more than 2.5 kilometres and chlorophyll *a* concentrations vary from 0 to ∼65 mg/m^3. The variability in the labelling data set should not be significantly smaller than that of the training set in order for the SOM to give optimal mapping results (Kohonen, 2001). In this study the labelling data set capture most of the variability in the training data set (Table 1). The temperature ranges are 2.8°C to 8.2°C smaller than those in the training set. Most of this difference is due to the fact that there are very few in situ measurements from ice-melting zones where water temperature drops to below 0°C. Those regions are negligible in terms of the area covered, and also the number of below zero measurements accounts for less than 1% of the training data; hence the lack of the lowest temperature labels in the labelling data set is unlikely to have a significant effect on the basin-wide pCO_2 maps.

The mixed layer depth is well represented in the labelling set. In winter and spring however, the maximum mixed layer depth in the labelling set is substantially lower than that in the training set. This has two causes, firstly commercial vessels avoid storm regions and therefore measurements in deep vertical mixing areas are rare, especially in winter when the ocean is generally under-sampled (Fig. 4); secondly, the highest MLD's in the training data occur in two very specific regions (Labrador Sea and the Greenland-Norwegian Sea) where deep water formation takes place. Those two relatively small basins are not extensively sampled, therefore the maxima measured are smaller. As a result the SOM output is potentially biased towards shallower mixed layer depths in all regions and seasons where the actual depth of the mixed layer is greater than ~850 m. This affects a small fraction (between 0.4% and 1.8%) of the training data as indicated in Table 1. However the exponential character of the relationship between sea surface pCO_2 and MLD found in the subpolar North Atlantic (Olsen et al., 2008) suggests that the MLDs deeper than around 500 m have little influence on changes in the sea surface pCO_2 (their Fig. 9a). Similar relationship was found for the subtropical North Atlantic in our labelling data set (not shown) with the threshold value of around 200 m.

The chlorophyll *a* concentrations in the labelling data capture most of the variability observed between 2004 and 2006. The seasonal maxima between 2 mg m^{-3} in winter and 27 mg m^{-3} in fall suggest that some strong blooms are not resolved by SOM. Nonetheless more than 99% of the training data falls within the range of the labelling data (Table 1), meaning that the SOM is labelled with a sufficient fraction of the observed variability. Cloud cover and the distribution of the ships' tracks, seem to be the main reason for lower ranges in the labelling data set with regards to the training data set. Additionally, winter chlorophyll *a* data are affected by the lack of satellite cover north of ~45° N in December and January.

2.6 Labelling the trained SOM with the pCO_2 data

At this point of the mapping procedure, each in situ pCO_2 measurement associated with a three-dimensional vector consisting of SST, MLD and CHL components is presented to the already trained SOM as input data (Fig. 1c). Instead of updating the winning neuron and its neighbourhood, such input vector labels the winning neuron with its pCO_2 value. Consequently each pCO_2 measurement is assigned to one of the neurons. Most of the neurons are labelled more than once and the ultimate pCO_2 value of the neuron is an average of all the labels it accounts for. Relationships between in situ pCO_2 measurements and each individual component (SST, MLD and CHL) of associated vectors are strongly non-linear from the basin-wide, year long perspective. Figure 2d–f shows how the density distribution of the SOM neurons follows the density distribution of the pCO_2 data. Most neurons concentrate where the data density is highest, giving a highly discriminative representation of the data. Neurons outside the data cloud mean that for a certain value of the property (x-axis) the SOM will estimate a pCO_2 value other than that measured (y-axis). This suggests that parameters other than those considered in this study control the distribution of pCO_2 in the North Atlantic. Adding sea surface salinity (SSS) as an additional variable in the training data matrix is suggested to improve SOM estimates, especially in subtropical and tropical North Atlantic (J. Boutin and N. Lefèvre, personal communication, 2008). SSS could act as a water mass tracer and a proxy for water parcel history, which would enable the SOM to account for the sea surface pCO_2 variability not determined by changes in SST, MLD and CHL.

2.7 Estimating basin-wide pCO_2 fields

In order to estimate the geographical distribution of pCO_2 for a certain time period, the training input data are used. Each of the 389 000 input vectors has a time coordinate and two space coordinates. June 2005 has been chosen to visualize this step, and is shown in Fig. 1d. All the input vectors with a certain time stamp (month and year in

this study) are presented to the preconditioned (trained and labelled) SOM. The input vector is labelled with the winning-neuron's pCO_2 label. Using the space coordinates of the input vector, this pCO_2 value is then associated with the appropriate pixel on the geographical map. As a final result of the SOM analysis, each pixel used as SOM input data has an estimated pCO_2 value assigned to it. In this study we produce 36 monthly, basin-wide pCO_2 maps between January 2004 and December 2006.

3 Results and discussion

3.1 Monthly pCO_2 maps

For each in situ pCO_2 measurement available, the corresponding SOM estimate was found based on spatial (1° longitude × 1° latitude grid box) and temporal (8 daily intervals between 1 January 2004 and 31 December 2006) coordinates. The residual r value was calculated as a difference between the two. The arithmetic mean of the absolute values of residuals for the entire dataset $\bar{r}$=0.0125, indicating random disparity between measurements and SOM estimates. Root mean-square (RMS) of residuals calculated as:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n}} \tag{4}$$

for the whole dataset (n), provides an estimate of the uncertainty of the method in reproducing the available in situ measurements, and equals to 11.55 μatm, which translates to 3.2% of an average pCO_2 value in the in situ dataset.

Scatter plots (one for each year) of estimated values, versus pCO_2 measured in situ are shown in Fig. 5. The correlation coefficient for the whole data set R=0.93 and varies from 0.96 in 2004, through 0.89 in 2005 to 0.93 in 2006. Presented distribution

in all 3 years indicates that no systematic bias exists in the method. The values are scattered well around the identity line (for which the correlation coefficient=1).

Out of 36 monthly pCO_2 maps generated, one representing each season for the 3 years is shown in Fig. 6. In the columns are 3 full seasonal cycles and in the rows maps for 2004, 2005 and 2006 presenting SOM estimates of the interannual variability. Signatures of physical and biological medium-to-large scale processes can be identified in the basin-wide context. In the subtropical gyre (~20–40° N), high pCO_2 values are found during spring and summer (Fig. 6d–i), with around 20 μatm lower values during fall and winter (Fig. 6a–c and j–l) which confirms the mainly temperature driven variability of pCO_2 in this region (Takahashi et al., 2002; Santana-Casiano et al., 2007). In the subpolar gyre (~40–60° N), massive biological CO_2 drawdown (Takahashi et al., 2002) is reflected in low pCO_2 levels during spring and summer (Fig. 6d–i). Mixing in the fall counteracts the effect of biological carbon uptake on pCO_2, which is visible as strong local maxima and minima in the subpolar waters with values of about 60 μatm apart (Fig. 6j–l). Relatively high pCO_2 values in the northern part of the basin in winter (Fig. 6a–c) are attributed to wind driven deepening of the mixed layer during storms in fall and winter, which brings CO_2 rich waters to the surface (Corbière et al., 2007). The lack of satellite measurements of chlorophyll *a* during late fall and winter in the northern (>60° N) North Atlantic (Kaufman, 1989; Moulin et al., 2001) makes it impossible to estimate the pCO_2 distribution in those regions using the presented SOM set up (Fig. 6a–c and j–l). As discussed in the following section, CHL introduces an important seasonal variability in the pCO_2 field in the northern North Atlantic, however the phytoplankton activity during late fall and winter in that part of the basin is low. In order to cover the region with the "missing" pCO_2 estimates, an additional SOM can be performed where only SST and MLD are used as the training data variables. Such a SOM fails to reproduce the seasonal cycle in most of the basin's sub-regions (not shown) but its winter estimates agree well with those obtained when biology is accounted for.

The influence of seasonally changing oceanographic features on the pCO_2 variability in the North Atlantic can also be distinguished from the maps. Intense upwelling of cold

waters along the coast off Northwest Africa serves as an example. The main upwelling centre follows the seasonal cycle of the belt of northeast trade winds (Hagen, 2001), reaching its northern-most position in summer and its southern-most position in winter. The increased pCO_2 values in this region induced by this upwelling (Pelegri et al., 2005) are recognized by high pCO_2 estimates at around 20° N to 25° N in summer (Fig. 6g–i), and at around 10° N to 15° N in winter (Fig. 6a–c).

3.2 Seasonal cycles in the main biogeochemical provinces

In such a heterogeneous basin as the North Atlantic, a coherent interpolation method ought to accurately extract the seasonal cycles of its most prominent sub regions. We first compare SOM estimates in 2 boxes within 2 biogeochemical provinces to an independent in situ data set. We then compare the SOM results to climatological results for 5 biogeochemical provinces.

Five months of data collected between the UK and Caribbean on board of *MV Santa Maria* during 2006 were not included in the labelling data set. Monthly means of this independent data for the boxes are shown in Fig. 7. The data used for labelling the SOM for these boxes in 2006 are also plotted. SOM estimates are area weighted means for the boxes. In both cases SOM reproduces the labelling data set well.

In the tropical North Atlantic (15° N to 25° N and 50° W to 60° W, Fig. 7a), the absolute value of the mean monthly residual between SOM estimates and independent data ranges from 0.3 to 13.2 μatm (mean 4.3 μatm). The independent data for one (April) month, out of five, falls outside the 1-σ standard deviation of the SOM-predicted values within the box. Given that, based on Gaussian probability distribution, 68% of independent values should fall within such defined error bars, SOM performs well in this direct validation exercise. Interestingly labelling data averaged for April is 25 μatm lower than the independent data. Both voyages used for calculating averages took place between 18 and 30 April 2006, within the same 10° longitude by 10° latitude box. Tracks also crossed (4 days apart) and measured values were around 20 μatm different at the crossing. Such high spatial and temporal sub-pixel variability complicates com-

paring relatively few in situ data from specific sampling region and period, to results obtained with interpolation methods designed for a much larger area.

In the subtropical North Atlantic (26° N to 38° N and 35° W to 60° W, Fig. 7b), the absolute value of the mean monthly residual between SOM estimates and independent data ranges from 0.2 to 31.5 μatm (mean 15.5 μatm). The SOM tends to underestimate the pCO_2 values with regards to the independent data set in late summer and fall and also fails to reproduce exceptionally high April values for the region. Sparse in situ data distribution (not shown) within the box may introduce sampling bias and may explain some of these differences. In addition, we need to suggest using SOM maps with caution when fine scale features and processes are analyzed.

The robustness of SOM estimates is further analyzed for five biogeochemical regions similar to those proposed by Longhurst (2007), shown in Fig. 8. The subpolar North Atlantic is represented by two provinces. The first combines the western part of the sub-Arctic (SARC) and the eastern part of the Arctic (ARCT) and stretches from 58° N to 66° N and from 10° W to 40° W. The second, the North Atlantic Drift Region (NADR) stretches between 46° N and 58° N and 10° W and 40° W. The North Atlantic Subtropical Gyre is divided into a western (NAST(W)) part, between 26° N and 38° N and 35° W and 70° W, and an eastern (NAST(E)) part between 26° N and 42° N and 10° W and 35° W part. The North Atlantic Tropical Gyre (NATR) stretches from 10° N to 26° N and from 20° W to 75° W (with a different latitudinal extent of the province of 20° N to 26° N, between 60° W and 75° W). For each province we show the number of data points available for training and labelling of the SOM. The mid-latitude North Atlantic has the smallest number of in situ measurements, whereas the northern provinces were by far the most sampled. The number of points in the training data depends mainly on the size of the province. The chosen data sources offer full year- around cover except for the occasional lack of the chlorophyll measurements in the SARC/ARCT region.

In Fig. 9, we compare SOM estimates for a reference year 2005 (mean of the monthly SOM estimates for 2004 to 2006) in these provinces, to a climatological distribution of the sea surface pCO_2 constructed for a reference year 2000 based on in situ pCO_2

measurements obtained from 1970 to 2006 (Takahashi et al., 2008). For comparison purposes we show Takahashi's climatological mean distribution (constructed for a reference year 2000) adjusted to a reference year 2005 assuming an annual rate of increase of 1.8 μatm as proposed in Takahashi et al. (2008).

In the SARC/ARCT region (Fig. 9a), SOM estimated pCO_2 values of 330–340 μatm during late spring and summer and of around 370 μatm during fall and winter agree with earlier findings showing that disequilibrium with the atmospheric pCO_2 (not shown here) exists throughout most of the year in this region, with the CO_2 air-sea flux directed into the ocean (Omar and Olsen, 2006; Olsen et al., 2008). Low summer pCO_2 due to strong biological carbon uptake (Takahashi et al., 2002), together with higher winter values (caused by deepening of the mixed layer supplying CO_2 rich waters to the surface) dominate the seasonal cycle. SOM estimates resolve such a pattern for the region. SOM values for the summer months are around 20 μatm higher than the long term climatological mean. However, according to Corbière et al. (2007), who analyzed data from 1993 to 2003 in the Western SARC, the seasonal amplitude can be as low as 20 μatm and as high as around 60 μatm, depending on the year. A variable intensity of the phytoplankton bloom, generally occurring in June, is given as an explanation by Corbière and co-workers. They also show that, at least for the mid-nineties, the climatological distribution proposed by Takahashi et al. (2002), may overestimate the strength of the summer biological carbon uptake and thus underestimate the pCO_2 values in the region. SOM estimates for the region are moreover consistent with those of Chierici et al. (2008) (Fig. 10). These authors estimated sea surface pCO_2 for 2005 in a region with a slightly smaller latitudinal extent than our SARC/ARCT (Fig. 10 shows SOM estimates for a region matching the one used in Chierici et al., 2008). Using measurements obtained onboard *MV Nuka Arctica*, together with remotely sensed data, they applied algorithms based on multiple regression. The in situ measurements used by Chierici and co-workers represent a fraction of the dataset used for the same region of the North Atlantic in this study. The resulting seasonal cycle for 2005 agrees well with SOM estimates for 2005 (Fig. 10). Neither method shows pCO_2 values below

325 μatm during the 2005 bloom. Both methods also produce an annual amplitude in pCO_2 of around 50 μatm. Obtaining similar results to a method designed for regional, high-resolution estimates, increases our confidence in the basin-wide SOM estimates, despite the fact that the pCO_2 data used have a large overlap.

In the NADR (Fig. 9b), the SOM estimates a relatively weak seasonal pCO_2 cycle, with an amplitude of 26 μatm. This is in line with results from Schuster and Watson (2007). These authors report an average annual amplitude of around 20 μatm in the eastern temperate region (35° N to 50° N and 5° W to 30° W) for years 2002 to 2005, 50% smaller than the amplitude found for years 1994 to 1995 (Fig. 3b in Schuster and Watson, 2007). This strong decrease in the amplitude over the last decade might also explain the difference in amplitude between the SOM estimates and the climatology in our region, which is shifted slightly to the North-West.

The seasonal cycle in the subtropical North Atlantic, represented here by two provinces (Fig. 9c–d), has an opposite shape to that further north. SOM pCO_2 estimates in the NAST(W) are characterized by one strong summer maximum in August, which agrees with the peak of the seasonal temperature cycle in the region (Takahashi et al., 2002; Phillips and Joyce, 2007). The annual amplitude of 41 μatm results from generally low primary production (Bates et al., 2002), having a small counteracting effect on the thermo-dynamically driven surface water pCO_2 variability (Bates, 1998, 2001). SOM estimates for the NAST(E) have a relatively low annual amplitude mainly due to a likely underestimation of the summer maximum in August by 10–15 μatm. Santana-Casiano et al. (2007) report summer maxima of 380 to 400 μatm at the ESTOC station (29°10′ N, 15°30′ W) for years 1995 to 2004. Also Schuster et al. (2008) report summer maxima of 400 and 390 μatm (for years 2005 and 2006 respectively) in a 5° latitude × 5° longitude grid box centred at 27.5° N, 17.5° W (Fig. 2 in Schuster et al., 2008). This is in line with the Takahashi climatology adjusted to 2005, which estimates a summer maximum of 386 μatm (Fig. 9d). The SOM, however has a summer maximum of 371 μatm. The SOM's inability to resolve the full annual amplitude of the pCO_2 cycle in NAST(E) requires further investigation. An increase in the spatial and

BGD
6, 3373–3414, 2009

temporal resolution of the training data to reduce the effect that averaging might have on the SOM's estimates as well as better pCO_2 in situ cover has been proposed to address this local "smoothing" effect of the SOM.

In the NATR (Fig. 9e) the SOM estimates a fairly flat seasonal cycle coupled to the temperature variability, which agrees well with the climatology. These warm, homogeneous waters do not support much primary production, and lack of strong winds excludes mixed layer deepening as a control for pCO_2 variability. Values are relatively high throughout the year and vary between 355 and 380 μatm. West African upwelling brings CO_2 rich waters to the surface, thus increasing sea surface pCO_2 values, specially during summer. Overall the SOM proves a robust method for reconstructing seasonal pCO_2 cycles in a diversified suite of biogeochemical provinces of the North Atlantic.

3.3 SOM estimates and the interannual variability of the training data

Medium-to-large scale processes and features of the seasonal pCO_2 cycle are modified in terms of size, strength and to some extent location, by interannual variability. The SOM's basin-wide estimates of such variability for each season are presented in Fig. 6. Visual inspection of this relatively short period (3 years) reveals apparent year-to-year changes. In the western subtropics summer (represented by August) values (Fig. 6g–i), are highest for 2005. Also the region of high (~400 μatm) pCO_2 values covers larger area than in either 2004 or 2006. Similarly in the North Atlantic Drift Region, fall (represented by October) pCO_2 values (Fig. 6j–l) in 2006 are higher and wider spread than those in the two previous years.

According to Eq. (1), SOM predictions are entirely data-based, and therefore the interannual variability in the SOM estimates can only be forced by the interannual variability of the three controlling parameters: SST, CHL and MLD. Their variability affects the pCO_2 distribution in a non-uniform manner, varying with each region's spatio-temporal dependence on the given parameter. Figures 11 and 12 show these relationships for NADR and NAST(W), respectively. Monthly mean values of estimated pCO_2 in each

province for 2004, 2005 and 2006 are represented by red, green and blue curves respectively. Error bars represent standard error of the mean calculated as:

$$\mathrm{SEM} = \frac{s}{\sqrt{n}} \tag{5}$$

where s represents the standard deviation of the sample, and n is the sample size. In the NADR (Fig. 11), during months with high mean CHL concentrations, the interannual variability of this parameter controls the variability in the estimated pCO_2 field. High mean CHL values (0.55–0.75 mg/m^3) during June–August 2004 correspond to the pCO_2 values around 8 μatm lower than during these months in later years; also a comparatively high chlorophyll concentration in March 2005 has decreased the pCO_2 values by ~15 μatm in comparison to other years. The interannual pCO_2 variability in the pre- and post-bloom periods appears to be controlled by variations in the MLD. However, significantly higher MLD values during January–May 2006 are not translated to similarly substantial increase in the pCO_2. This is explained by the non-linear character of the relationship between sea surface pCO_2 and MLD in the subpolar North Atlantic proposed by Olsen et al. (2008) and confirmed for our data (not shown). These authors found that MLDs deeper than around 300 m have no or little influence on the increase in the sea surface pCO_2 (their Fig. 9a), whereas an MLD increase between 0 and 200 m corresponds to the pCO_2 increase of up to 100 μatm. This relationship is to some extent influenced by the variability in the remaining two parameters, but SOM estimates appear coherent with their findings. It is also worth noting that the significant increase in the SST between May and August (almost uniform for 2005 and 2006 and 1.5°C lower for 2004) does not translate into a corresponding increase in estimated sea surface pCO_2. The biological CO_2 drawdown dominates the thermodynamical effect during the bloom period (Olsen et al., 2008), and SOM estimates represent this relationship correctly.

In the NAST(W) (Fig. 12) the estimated interannual pCO_2 variability for the years 2004 to 2006 appears to be controlled by the variations in the MLD field. The seasonal pCO_2 cycle in those regions is controlled by the combination of SST and MLD as

mentioned earlier, but the SST interannual variability for the years under investigation is very weak, hence the MLD dominance in controlling the interannual variability. The strongly non-linear relationship between the sea surface pCO_2 and MLD found in the subpolar North Atlantic is also apparent in the subtropical part of the basin. An opposite sign of this relationship (pCO_2 decreases with increasing MLD) found in the subtropics (not shown) is due to the opposite shape of the, temperature-controlled, seasonal cycle. The deepening of the MLD takes place during the cooling-related pCO_2 decrease, and MLD shallowing occurs during seasonal, warming-related pCO_2 increase. SOM estimates resolve this relationship, which can be observed especially during late winter and early spring, when SST and CHL vary very little.

During January–March period of 2004 the MLD was variable and shallow in the subtropics (55 m decrease, from 96 m in January to 41 m in March), and the SOM predicts variable pCO_2 as a result (9.3 μatm increase). During February–April period of 2005 the MLD was similarly shallow and variable (62 m decrease), and SOM predicts 12.3 μatm increase in the pCO_2. Contrary to that, greater variability (53 m increase followed by 109 m decrease) of the deep MLD during February-April period of 2006 translates to less than 1.5 μatm variability in the predicted pCO_2. The SOM reproduces the pattern previously noted by Olsen et al. (2008) for the northern North Atlantic, and confirmed in our data for the subtropics.

4 Summary and conclusions

A self organizing neural network has been applied to construct 36 basin-wide, monthly pCO_2 maps over the North Atlantic for 2004 to 2006. Estimates of three full seasonal cycles and interannual variability between 2004 and 2006 show that the method can account for medium-to-large scale biological and physical processes. The choice of training parameters has resulted in a powerful mapping performance. The estimated seasonal pCO_2 cycles in five major biogeochemical provinces mostly agrees with other data analyses. The distribution of monthly sea surface pCO_2 for a reference

year 2005 in the northern provinces of the North Atlantic suggests that current pCO_2 values are 20 to 30 μatm higher than the 35-year climatology (Takahashi et al., 2008) indicates. The difference is especially profound in the phytoplankton bloom period (June–September). The lack of estimates in the northern part of the basin in the winter months is a disadvantage of the current SOM set-up for several applications. However, this important issue can be resolved by combining two SOM runs and covering the missing regions with no-biology predictions.

Discrepancies identified in the eastern subtropics reveal the method's tendency to "smooth" highest and lowest values. This behaviour is to some extent expected from the method which is supposed to robustly estimate basin-wide values. An introduction of basin-wide sea surface salinity field as an additional training parameter is suggested to improve SOM estimates. This will be possible following the launch of ESA's Soil Moisture and Ocean Salinity (SMOS) sensor, planned between July and October 2009. However the influence of the smoothing effect on the overall performance seems to be minimal and mainly related to the analyses of features of sub-pixel to a few pixels in size. Very high spatial and temporal natural pCO_2 variability makes the SOM estimates too coarse for such a small-scale analysis and they should be considered as designed for analyses over larger regions.

The estimated interannual pCO_2 variability provides confirmation of the SOM's pattern extraction capabilities. There is no need for implementing a mathematical description of governing relationships a priori, as long as sufficient data are available. This validates the possibility of using the method to examine the interannual variability in the North Atlantic over the last decade or so, during which the region seem to have weakened as a net CO_2 sink (Schuster and Watson, 2007). The estimates of the interannual variability could also add significant value to future model predictions. Current models either lack the interannual variability or disagree with in situ measurements. The sparse nature of in situ observations is often given as an explanation; therefore basin-wide maps should serve as a better input.

As a whole, this is a major improvement over historical efforts to map the pCO_2 in

the entire basin, eliminating the need to divide the basin into several regions in order to derive biogeochemical relationships for each one separately. The SOM's ability to extract numerous existing relationships simultaneously provides an equally good fit to the data and allows for basin-wide analysis over several years.

The continuation of large scale in situ marine pCO_2 measurements will improve our understanding of the actual spatial and temporal variability in the real ocean, and allow us to interpret estimated values with more confidence. It is our current intention that SOMs be used in conjunction with these measurements during the future oceanic pCO_2 monitoring programs.

Acknowledgements. We thank the captains, officers and crews of the entire commercial and research fleet used during data collection for continuous technical assistance and support on board the ships. This work was supported by the European Commission under the CARBOOCEAN project GOCE 511176-2; the EU Marie Curie RTN Greencycles project (MRTN-CT-2004-512464); Spanish project ICCABA CTM2005-03893/MAR; the Norwegian Research Council through A-CARB (178167) and CARBON-HEAT (185093); the Swedish National Space Board through RESCUE – II (d.nr 62/07:1). Special thanks are for Benjamin Pfeil (Bjerknes Centre for Climate Research, Norway) for data handling and for Laurent Bopp (Laboratoire des Sciences du Climat et de l'Environnement, France) and Corinne Le Quèrè (University of East Anglia, UK) for helpful comments and discussions.

References

Ali, M. M., Kishtawal, C. M., and Jain, S.: Predicting cyclone tracks in the North Indian Ocean: An artificial neural network approach, Geophys. Res. Lett., 34, L04603, doi:10.1029/2006GL028353, 2007.

Bates, N. R., Takahashi, T., Chipman, D. W., and Knap, H.: Variability of pCO_2 on diel to seasonal timescales in the Sargasso Sea, J. Geophys. Res., 103, 15567–15585, 1998.

Bates, N. R.: Interannual variability of oceanic CO_2 and biogeochemical properties in the western North Atlantic subtropical gyre, Deep-Sea Res. II, 48, 1507–1528, 2002.

Bates, N. R., Pequignet, A. C., Johnson, R. J., and Gruber, N.: A short term sink for atmospheric CO_2 in subtropical mode water of the North Atlantic Ocean, Nature, 420, 489–493, 2002.

Canadell, J. G. Le Quéré, C., Raupach, M. R., Field, C. B., Buitenhuis, E. T., Ciais, P., Conway, T. J., Gillett, N. P., Houghton, R. A., and Marland, G.: Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks, P. Natl. Acad. Sci. USA, 104(24), 10288–10293, 2007.

Cavazos, T.: Using Self-Organizing Maps to Investigate Extreme Climate Events: An Application to the Wintertime Precipitation in the Balkans, J. Climate, 13, 1718–1732, 1999.

Chierici, M., Olsen, A., Johannessen, T., Trinanes, J., Wanninkhof, R.: Algorithms to estimate the carbon dioxide uptake in the northern North Atlantic using ship-observations, satellite and ocean analysis data, Deep-Sea Res. II, in press, 2009.

Cooper, D. J., Watson, A. J., and Ling, R. D.: Variation of pCO_2 along a North Atlantic shipping route (U.K. to the Caribbean): A year of automated observations, Mar. Chem., 60, 147–164, 1998.

Corbière, A., Metzl, N., Reverdin, G., Brunet, C., and Takahashi, T.: Interannual and decadal variability of the oceanic carbon sink in the North Atlantic subpolar gyre, Tellus, 59B(2), 168–178, 2007.

Dreyfus, G.: Neural Networks: Methodology and Applications, Springer-Verlag, Berlin Heidelberg New York, Germany, 2005.

Hagen, E.: Northwest African upwelling scenario, Oceanol. Acta, 24, S113–S128, 2001.

Hewitson, B. C. and Crane, R. G.: Self-organizing maps: applications to synoptic climatology, Climate Res., 22, 13–26, 2002.

Jamet, C., Moulin, C., and Lefèvre, N.: Estimation of the oceanic pCO_2 in the North Atlantic from VOS lines in situ measurements: Parameters needed to generate seasonally mean maps, Ann. Geophys., 25, 1–11, 2007, http://www.ann-geophys.net/25/1/2007/.

Kalnay, E. Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Redell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Leetmaa, A., Reynolds, R., and Jenne, R.: The NCEP/NCAR Reanalysis Project, B. Am. Meteor. Soc., 77, 437–471, 1996.

Kaufman, Y. J.: The atmospheric effect on remote sensing and its correction, pages 336–428, in: Theory and applications of optical remote sensing, edited by: Asrar, G., Wiley-Interscience, 752 pp., 1989.

Kohonen, T.: Self-Organizing Maps, Third ed., Springer-Verlag, Berlin Heidelberg New York, 501 pp., 2001.

Lefèvre, N., Watson, A. J., Cooper, D. J., Weiss, R. F., Takahashi, T., and Sutherland, S. C.: Assessing the seasonality of the oceanic sink for CO_2 in the northern hemisphere, Global Biogeochem. Cy., 13(2), 273–286, 1999.

Lefèvre, N. Watson, A. J., Olsen, A., Ríos, A. F., Pérez, F. F., and Johannessen, T.: A decrease in the sink for atmospheric CO_2 in the North Atlantic, Geophys. Res. Lett., 31, L07306, doi:10.1029/2003GL018957, 2004.

Lefèvre, N., Watson, A. J., and Watson, A. R.: A comparison of multiple regression and neural network techniques for mapping in situ pCO_2 data, Tellus, 57, 375–384, 2005.

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

Liu, Y. and Weisberg, R.: Sea Surface Temperature Patterns on the West Florida Shelf Using Growing Hierarchical Self-Organizing Maps, J. Atmos. Ocean. Tech., 23, 325–338, 2006a.

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

Longhurst, A. R.: Ecological Geography of the Sea, Second Ed., Academic, Boston, Mass., 542 pp., 2007.

Lüger, H., Wallace, D. W. R., Körtzinger, A., and Nojiri, Y.: The pCO_2variability in the midlatitude North Atlantic Ocean during a full annual cycle, Global Biogeochem. Cy, 18, GB3023, doi:10.1029/2003GB002200, 2004.

Lüger, H., Wanninkhof, R., Wallace, D. W. R., and Körtzinger, A.: CO_2 fluxes in the subtropical and subarctic North Atlantic based on measurements from a volunteer observing ship, J. Geophys. Res, 111, C06024, doi:10.1029/2005JC003101, 2006.

Moulin, C., Gordon, H. R., Chomko, R. M., Banzon, V. F., and Evans, R. H.: Atmosphereic correction of ocean color imagery through thick layers of Saharan dust, Geophys. Res. Lett., 28, 5–8, 2001.

Niang, A., Thiria, S., Badran, F., Moulin, C.: Automatic neural classification of ocean colour reflectance spectra at the top of the atmosphere with introduction of expert knowledge, Remote Sens. Environ., 86, 257–271, 2003.

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

Olsen, A., Bellerby, R. G. J., Johannessen, T., Omar, A. M., and Skjelvan, I.: Interannual variability of the wintertime air-sea flux of carbon dioxide in the northern North Atlantic, 1981–

2001, Deep-Sea Res. I, 50, 1323–1338, 2003.
Olsen, A., Triñanes, J. A., and Wanninkhof, R.: Sea-air flux of CO_2 in the Caribbean Sea estimated using in situ and remote sensing data, Remote Sens. Environ, 89, 309–325, 2004.
Olsen, A., Brown, K. R., Chierici, M., Johannessen, T., and Neill, C.: Sea-surface CO_2 fugacity in the subpolar North Atlantic, Biogeosciences, 5, 535–547, 2008, http://www.biogeosciences.net/5/535/2008/.
Omar, A. M. and Olsen, A.: Reconstructing the time history of the air-sea CO_2 disequilibrium and its rate of change in the eastern subpolar North Atlantic, 1972–1989, Geophys. Res. Lett., 33, L04602, doi:10.1029/2005GL025425, 2006.
Pelegri, J. L., Aristegui, J., Cana, L., Gonzalez-Davila, M., Hernandez-Guerra, A., Hernandez-Leon, S., Marrero-Diaz, A., Montero, M. F., Sangra, P., and Santana-Casiano, J. M.: Coupling between open ocean and the coastal upwelling region off northwet Africa:water circulation and offshore pumping of organic matter, J. Marine Syst., 54, 3–37, 2005.
Phillips, H. E. and Joyce, T. M.: Bermuda's tale of two time series: Hydrostation S and BATS, J. Phys. Oceanogr., 37, 554–571, 2007.
Reusch, D. B., Alley, R. B., and Hewitson, B. C.: North Atlantic climate variability from a self-organizing map perspective, J. Geophys. Res, 112, D02104, doi:10.1029/2006JD007460, 2007.
Richardson, A. J., Risien, C., and Shillington, F. A.: Using self-organizing maps to identify patterns in satellite imagery, Prog. Oceanogr., 59, 223–239, 2003.
Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., Wanninkhof, R., Wong, C. S., Wallace, D. W. R., Tilbrook, B., Millero, F. J., Peng, T. H., Kozyr, A., Ono, T., and Rios, A. F.: The oceanic sink for anthropogenic CO_2, Science, 305, 5682, 367–371, 2004.
Santana-Casiano, J. M., Gonzalez-Davila, M., Rueda, M. J., Llinas, O., and Gonzalez-Davila, E. F.: The interannual variability of oceanic CO_2 parameters in the northeast Atlantic subtropical gyre at the ESTOC site, Global Biogeochem. Cy, 21, GB1015, doi:10.1029/2006GB002788, 2007.
Sarmiento, J. L. and Gruber, N.: Sinks for anthropogenic carbon, Phys. Today, 55, 30–36, 2002.
Schuster, U. and Watson, A. J. W.: A variable and decreasing sink for atmospheric CO_2 in the North Atlantic, J. Geophys. Res., 112, C11006, doi:10.1029/2006JC003941, 2007.
Schuster, U., Watson, A. J., Bates, N., Corbière, A., Gonzalez-Davila, M., Metzl, N., Pierrot, D., and Santana-Casiano, M.: Trends in North Atlantic sea surface fCO_2 from 1990 to 2006,

Deep-Sea Res. II, in press, 2009.
Solidoro, C., Bandelj, V., Barbieri, P., Cossarini, G., and Umani, S. F.: Understanding dynamic of biogeochemical properties in the northern Adriatic Sea by using self-organizing maps and k-means clustering, J. Geophys. Res, 112, C07S90, doi:10.1029/2006JC003553, 2007.
Takahashi, T. Feely, R. A., Weiss, R. F., Wanninkhof, R. H., Chipman, D. W., Sutherland, S. C., and Takahashi, T. T.: Global sea-air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects, Deep-Sea Res. II, 49(9–10), 1601–1622, 2002.
Takahashi, T., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A. J., Bakker, D. C., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T. S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R., Wong, C. S., Delille, B., Bates, N. R., and de Baar, H. J. W.: Climatological mean and decadal change in surface ocean pCO_2, and net sea-air CO_2 flux over the global oceans, Deep-Sea Res. II, in press, 2009.
Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankagas, J.: SOM Toolbox for Matlab 5, Libella Oy, Espoo, 59 pp., 2000.

Table 1. Ranges of sea surface temperature, mixed layer depth and chlorophyll *a* in the training (T) and labelling (L) data sets by season.

Season	Data	Temperature (°C)			Mixed Layer Depth (m)			Chlorophyll *a* (mg/m^3)		
		Min	Max	L cover(%)[a]	Min	Max	L cover(%)[a]	Min	Max	L cover(%)[a]
WINTER	T	−1.80	29.2	99.7	10.0	>1000[b]	98.2	0.04	35.8	98.5
(Dec–Feb)	L	0.45	28.5	–	17.9	571.9	–	0.05	2.0	–
SPRING	T	−1.80	29.7	97.8	10.0	>1000[c]	99.2	0.02	64.6	99.8
(Mar–May)	L	0.17	28.9	–	10.0	834.5	–	0.03	9.6	–
SUMMER	T	−1.80	30.3	95.7	8.4	387.5	99.5	0.02	57.9	99.6
(Jun–Aug)	L	1.92	29.1	–	10.0	337.7	–	0.03	12.7	–
FALL	T	−1.80	30.7	97.9	9.0	484.9	99.6	0.02	32.4	99.1
(Sep–Nov)	L	5.85	30.1	–	12.0	360.4	–	0.04	26.8	–

[a] Percentage of the training data within the range of the labelling data set.
[b] 0.2% of data are above 1000 m.
[c] 0.4% of data are above 1000 m.

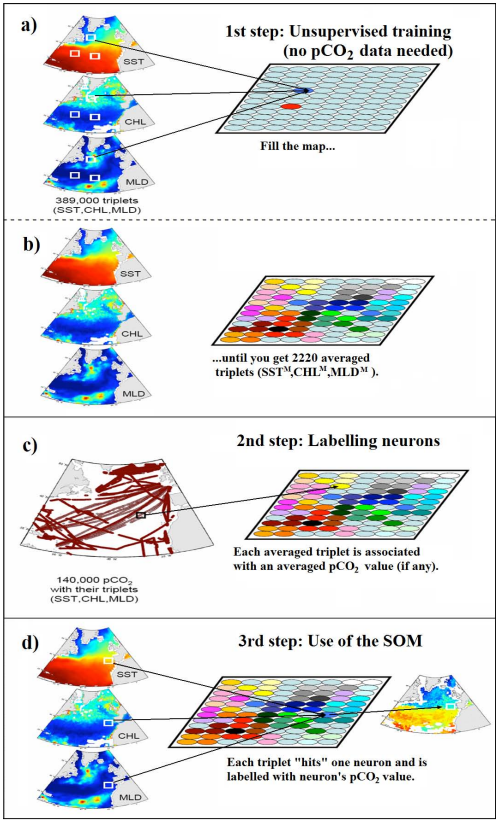

Fig. 1. Visualization of the procedures for the self organizing map (SOM). Three main steps are necessary: first (**a** and **b**), an unsupervised training takes place, and no pCO_2 data is used; second **(c)**, preconditioned neurons are labelled with pCO_2 data measured in situ; third **(d)**, the trained and labelled SOM is used to assign pCO_2 values to the geographical map for the whole basin.

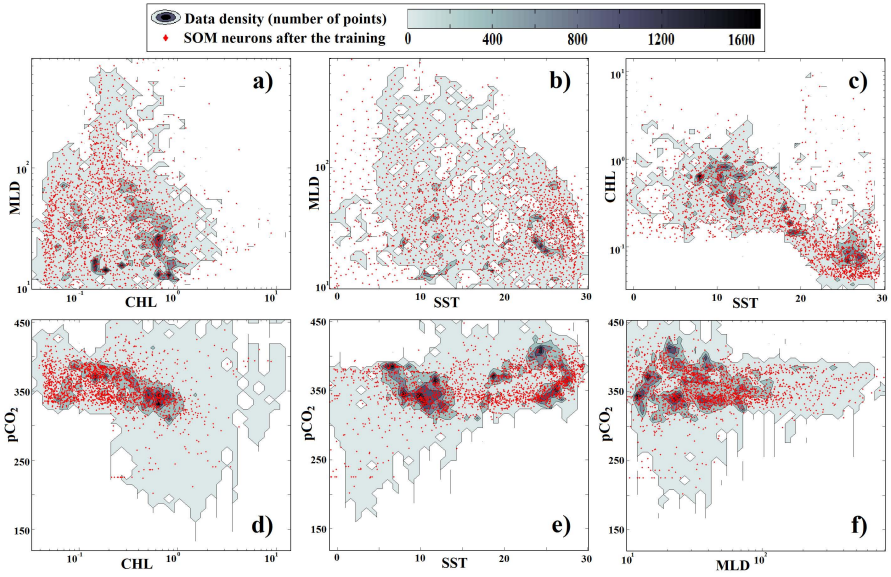

Fig. 2. Property – property plots for pCO_2, MLD, CHL and SST. The distribution of the density of the data of all the 389 000 training data points within each 2-dimensional data space are shown in grey. Overlaid in red is the distribution of 2220 SOM neurons after the training. To account for the non-linearity in the system the distribution of the neurons matches the distribution of the data.

BGD
6, 3373–3414, 2009

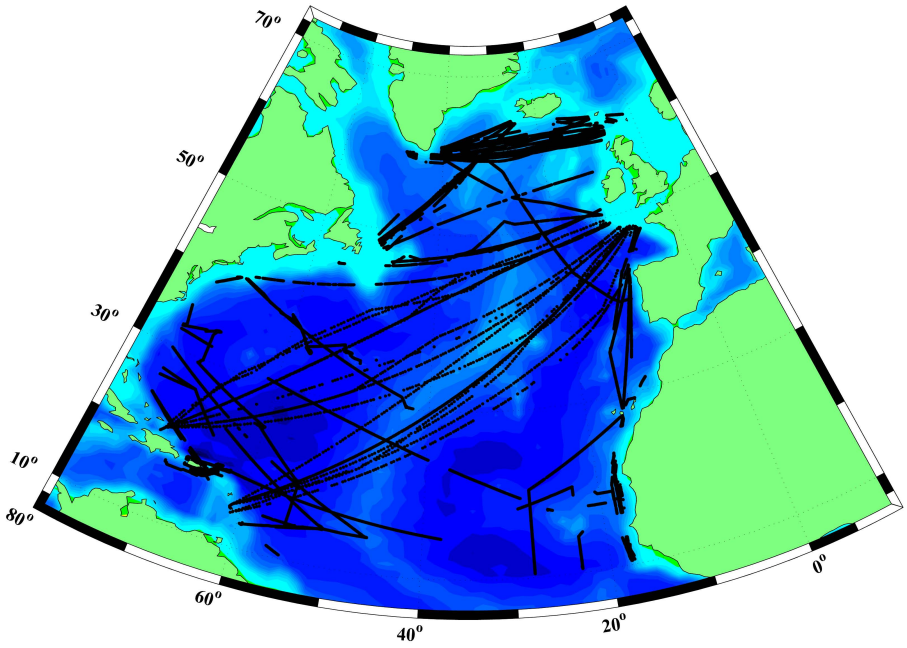

Fig. 3. The spatial distribution of the pCO_2 measurements used in this study. Data constitute the subset of the CarboOcean dataset for 2004 through 2006.

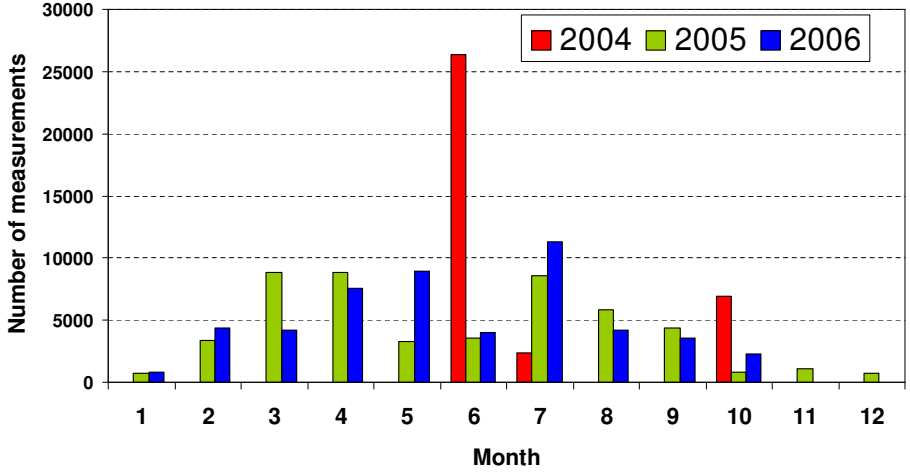

Fig. 4. Number of in situ pCO_2 measurements in the North Atlantic used for labelling the preconditioned SOM, versus month, for 2004, 2005 and 2006.

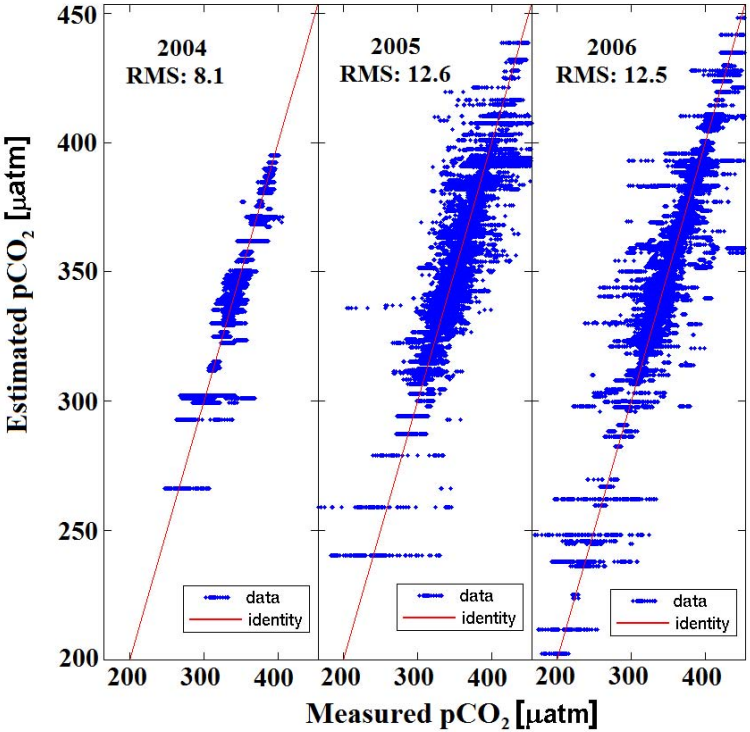

Fig. 5. Scatterplots of the pCO_2 estimated by the SOM, versus the measured pCO_2 (binned into 1°×1°×week) for 2004, 2005 and 2006. Root mean-square of residuals (RMS) for each year provides the accuracy of the method.

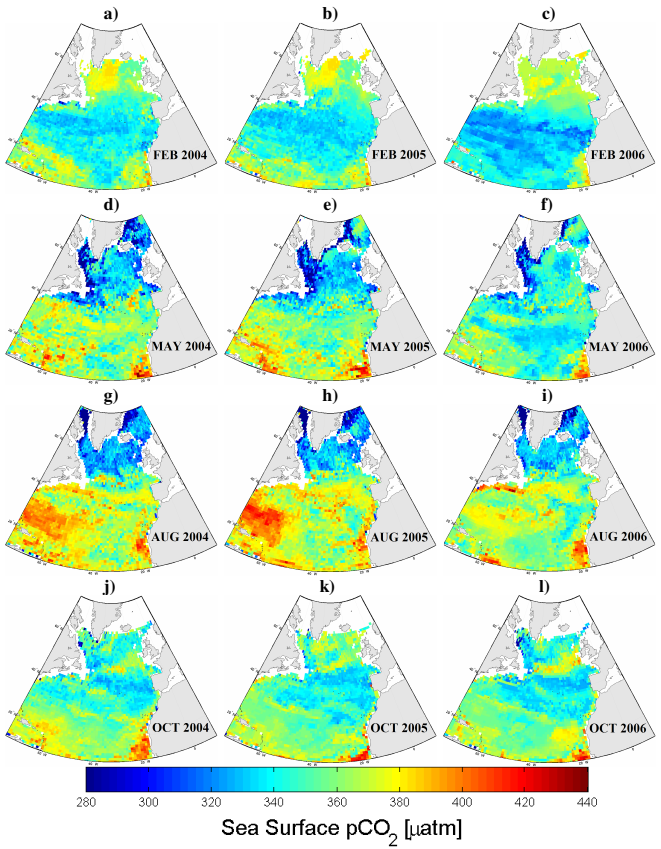

Fig. 6. Seasonal (in columns) and interannual (in rows) variability of the sea surface pCO_2 in the North Atlantic for years 2004 to 2006.

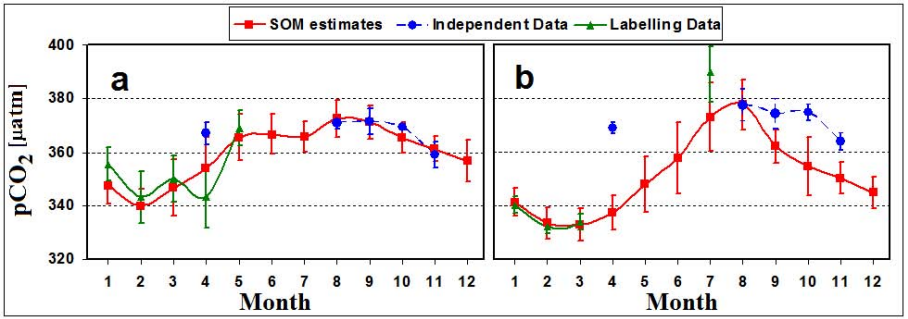

Fig. 7. Monthly mean pCO_2 versus month in 2006 estimated by SOM for 2 regions where independent (not used in SOM training or labelling) data from *MV Santa Maria* are available. The regions are well within (away from provinces' borders) the tropical North Atlantic (**a**, 15° N to 25° N and 50° W to 60° W) and the western subtropical North Atlantic (**b**, 26° N to 38° N and 35° W to 60° W). For comparison, monthly pCO_2 means from the labelling data of 2006 are also shown. The vertical bars extend from -1σ to $+1\sigma$ of the area weighted distribution for the given region.

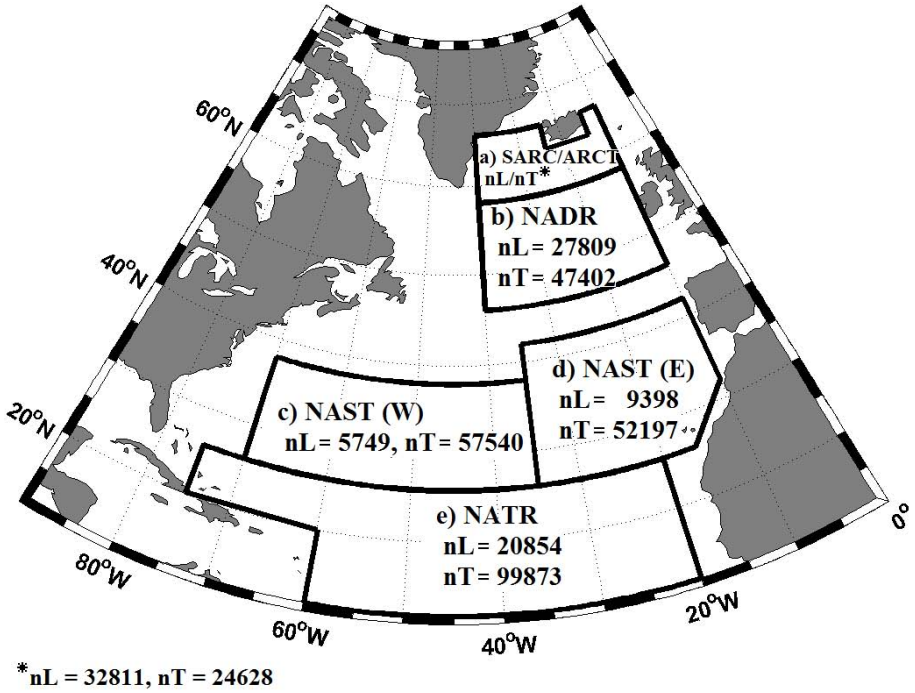

Fig. 8. The biogeochemical provinces of the North Atlantic proposed by Longhurst (2007) as used here for an analysis of SOM estimates. nT and nL represent the number of data points available for training and labelling of the SOM, respectively.

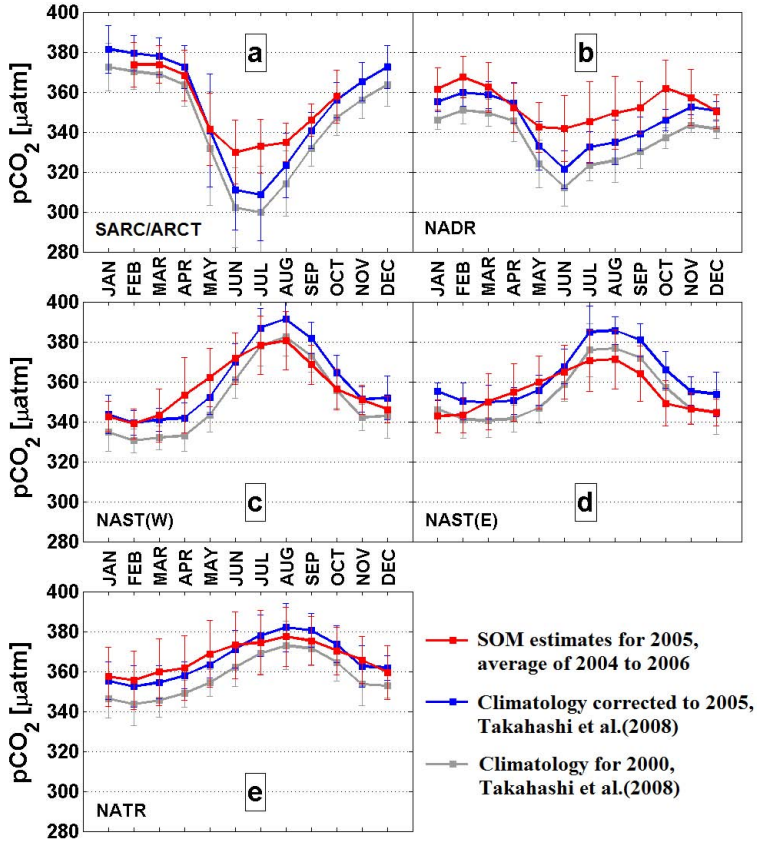

Fig. 9. Seasonal cycle of the sea surface pCO_2 in five biogeochemical provinces of the North Atlantic. The vertical bars extend from -1σ to $+1\sigma$ of the area weighted distribution for the given region.

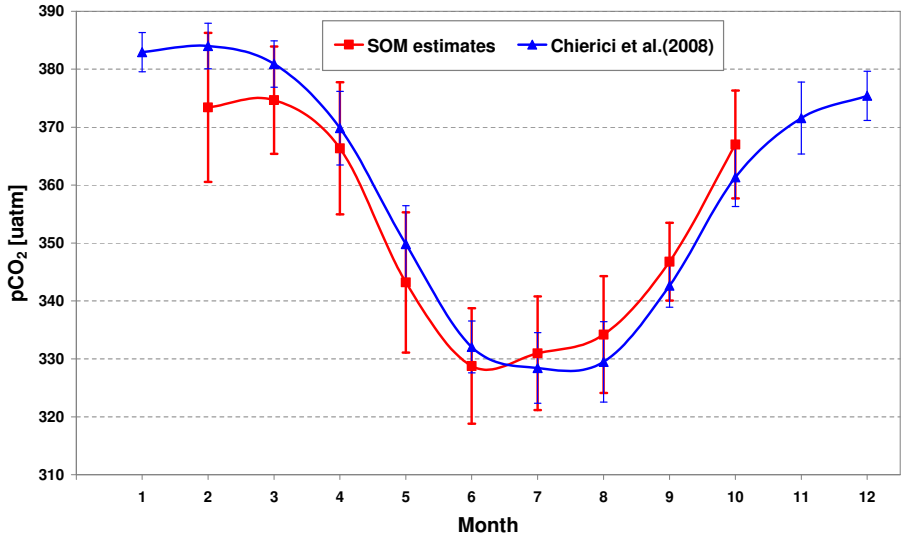

Fig. 10. Monthly, area weighted pCO_2 fields during 2005 in the subpolar North Atlantic (58° N to 63° N, 10° W to 40° W) estimated by SOM, compared to the multiple regression estimates for 2005 in the same region performed by Chierici et al. (2008).

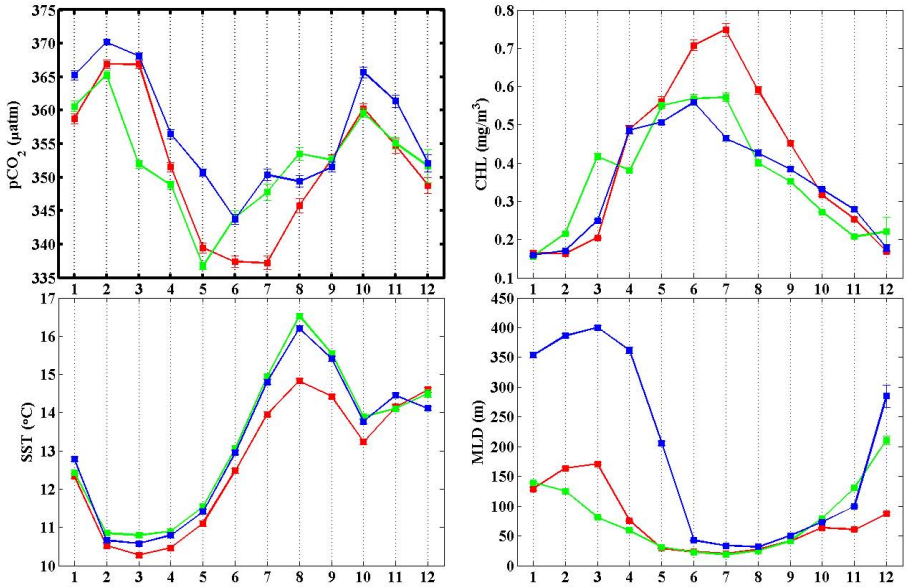

Fig. 11. SOM estimates of the pCO_2 interannual variability during 2004 (red), 2005 (green) and 2006 (blue) in the *NADR* province (46° N to 58° N and 10° W to 40° W) compared to the variability of SST, CHL and MLD.

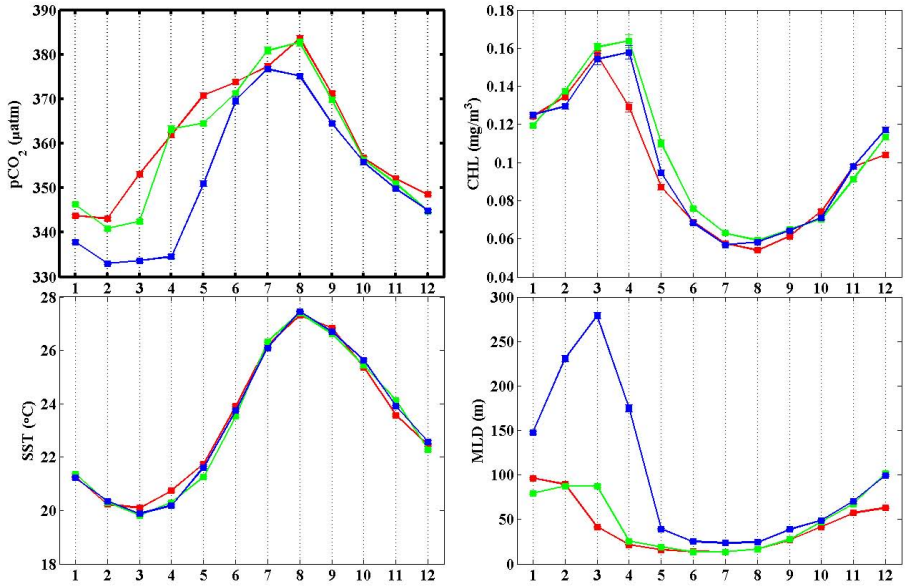

Fig. 12. SOM estimates of the pCO_2 interannual variability during 2004 (red), 2005 (green) and 2006 (blue) in the *NAST(W)* province (26° N to 38° N and 35° W to 70° W) compared to the variability of SST, CHL and MLD.