Biogeosciences, 6, 1405–1421, 2009
www.biogeosciences.net/6/1405/2009/

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. F. 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 Investigaciones Marinas, Consejo Superior de Investigaciones Científicas (CSIC), 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 – Published in Biogeosciences Discuss.: 30 March 2009
Revised: 21 July 2009 – Accepted: 31 July 2009 – Published: 5 August 2009

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. The maps have been computed using a neural network technique which reconstructs the non-linear relationships between three biogeochemical parameters and marine pCO_2. A self organizing map (SOM) neural network has been trained using 389 000 triplets of the SeaWiFS-MODIS chlorophyll-*a* concentration, the NCEP/NCAR reanalysis sea surface temperature, and the FOAM mixed layer depth. The trained SOM was labelled with 137 000 underway pCO_2 measurements collected in situ during 2004, 2005 and 2006 in the North Atlantic, spanning the range of 208 to 437 μatm. The root mean square error (RMSE) of the neural network fit to the data is 11.6 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 estimates of the interannual variability in the major biogeochemical provinces are presented and discussed. High resolution combined with basin-wide coverage makes the maps a useful tool for several applications such as the 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.

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

1 Introduction

Globally, the oceans have 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 carbon cycle. Understanding the future behaviour of the global carbon sinks and sources, as well as related effects on the planet's climate, can only be obtained given a robust understanding of the current distribution of carbon sink and 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 strongly 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. However, over the last decade, technical improvements and cooperation with the shipping industry have allowed for the installation of several autonomous underway systems on board commercial vessels routinely crossing the ocean basin. Those instruments perform quasicontinuous measurements, offering temporal and spatial coverage 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 and 2006; Corbière et al,. 2007; Schuster and Watson, 2007; Olsen et al., 2004 and 2008; Schuster et al., 2009). 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 an increasing fraction of the anthropogenic emissions remains in the atmosphere, which is consistent with some 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 pCO_2 measurements in the North Atlantic, the coverage still remains unevenly distributed in time and space. The regional character of the existing estimates poses difficulties for extrapolation to the entire basin; therefore a robust and reliable method to spatially and temporarily interpolate available measurements of marine pCO_2 has been long sought (e.g. Lefèvre et al., 1999 and 2005; Takahashi et al., 2002 and 2009; Olsen et al., 2003; Jamet et al., 2007; Chierici et al., 2009).

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 were 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. Among the numerous NN types, the Self Organizing Map (SOM) seems to gain the most attention as being well suited to study empirical relationships in geosciences. It is a particularly powerful tool for the extraction and classification of features, such as trends, in (and between) input variables. 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. The SOM uses an unsupervised (no need for a priori, empirical or theoretical description of the input – output relationships) learning algorithm, enabling us to identify relationships among the state variables of the phenomena under analysis, where our understanding of these is insufficient to be fully described using mathematical equations, and where applications of knowledge-based models are therefore limited.

The SOM technique has been successfully used to synthesise regional marine pCO_2 maps from in situ measurements. Lefèvre et al. (2005) have constructed a monthly climatology for years 1995–1997 using the reanalyzed SST fields as the SOM input. Their estimates cover the North Atlantic subpolar gyre (50° N to 70° N and 10° W to 60° W). Lefèvre et al. were able to capture a more complex distribution in the northern North Atlantic using SOM than they could using multiple linear regressions. Also the residuals determined through the validation against an independent subset of the data were smaller for the SOM.

In this study, we construct basin-wide (10.5° N to 75.5° N and 9.5° E to 75.5° W), monthly pCO_2 maps for three consecutive years with a 1° latitude by 1° longitude resolution. We use 137 000 in situ pCO_2 measurements collected in the North Atlantic throughout 2004, 2005 and 2006 as part of CarboOcean (http://www.carboocean.org/), an EU-funded Integrated Project, and parallel US projects. The pCO_2 data are combined with 389 000 satellite, reanalysis and assimilation data of chlorophyll-*a*, sea surface temperature, and mixed layer depth, which allows basin-wide, continuous mapping over extended periods of time.

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): sea surface temperature (SST),

wind-mixed layer depth (MLD) and the abundance of photosynthesizing organisms in the surface ocean represented by the chlorophyll-*a* concentration (CHL).

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

Lefèvre et al. (2005) and Friedrich and Oschlies (2009) used position and time as additional training parameters for their SOM-based mapping. The different training scheme (compared to that used by Friedrich and Oschlies, 2009) applied in this study allows for improved determination of the statistical structure of the basin-wide input data (discussed in Sect. 3.1). However, the patterns found are more strongly implemented in the resulting maps. The maps obtained using position and time are unrealistic (not shown), as also reported by Jamet et al. (2007) who compared three different combinations of parameters needed to generate pCO_2 maps in the North Atlantic. Using latitude or longitude causes concentration of similar values along east-west or north-south lines, respectively. Using both causes clustering of similar pCO_2 values in patches, with surprisingly equal distances between one another. Finally, using time increases the influence of seasonality on the pCO_2 maps. Thus, whereas using position and time can be sufficient to work with small regions (e.g. Lefèvre et al., 2005), they are definitely not applicable as a basin-wide training parameters.

During our SOM analysis three steps are taken in order to estimate basin-wide pCO_2 fields: first, an unsupervised training takes place without pCO_2 data; second, in situ pCO_2 data is used to label the preconditioned SOM neurons; third, the trained and labelled SOM neurons are used to assign pCO_2 values to the (geographical) grid points of the North Atlantic.

2.1 An overview of the SOM setup

The SOM, introduced by Kohonen (2001), is a competitive learning method in which an algorithm learns to classify the samples by recognizing and extracting patterns 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 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.

Our SOM-map consists of 2220 i units (often referred to as neurons) organized on a regular 2D 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 three-dimensional weight vector $\boldsymbol{y_i}$, with one component for each input variable (SST, MLD and CHL). All the values are linearly normalized to acquire an even weight distribution between the input variables. Additionally CHL and MLD values are $\log_{10}$ normalized to minimize the influence which their spread throughout four and three orders of magnitude, respectively, would otherwise have on the weight distribution. Linear initialization (performed in this study) of the components of weight vectors applied prior to the training process decreases the computing time required for the SOM to converge with the input data (Kohonen, 2001).

2.2 Training data set (SST, MLD, CHL)

The training data set consists of three subsets, one for each parameter. Basin-scale SST data were obtained from the NCEP/NCAR Reanalysis Project (http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis.html) at daily frequency and 2.5° latitude $\times$2.5° longitude resolution. The SST data (used in this study) contain the values as described in Kalnay et al. (1996). As such, over open water the temperature is fixed at its initial weekly value and linearly interpolated to daily frequency in the NCEP data product.

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$\times$1° longitude resolution. The FOAM model assimilates both in situ and remotely sensed ocean observations in near real-time including: vertical temperature and salinity profiles from sea stations and research vessels, Argo profiling floats and PIRATA moored arrays, as well as sea surface temperature from Voluntary Observing Ships (VOS), buoys, and the satellite mounted Advanced Very High Resolution Radiometer (AVHRR). The mixed layer depth used in this study is determined by the FOAM model using the density based criterion as the depth where a density increase of $0.05\,\mathrm{kg\,m^{-3}}$ from the surface value occurs (Chunlei Liu, Environmental Systems Science Centre of the UK National Environmental Research Council, personal communication, 2007).

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 considerable improvement in coverage in relation to the single mission products (20% and 24% for SeaWiFS and MODIS 8-daily product, respectively).

All three products (SST, MLD, CHL) offer almost full basin-wide coverage for the years 2004 to 2006. All

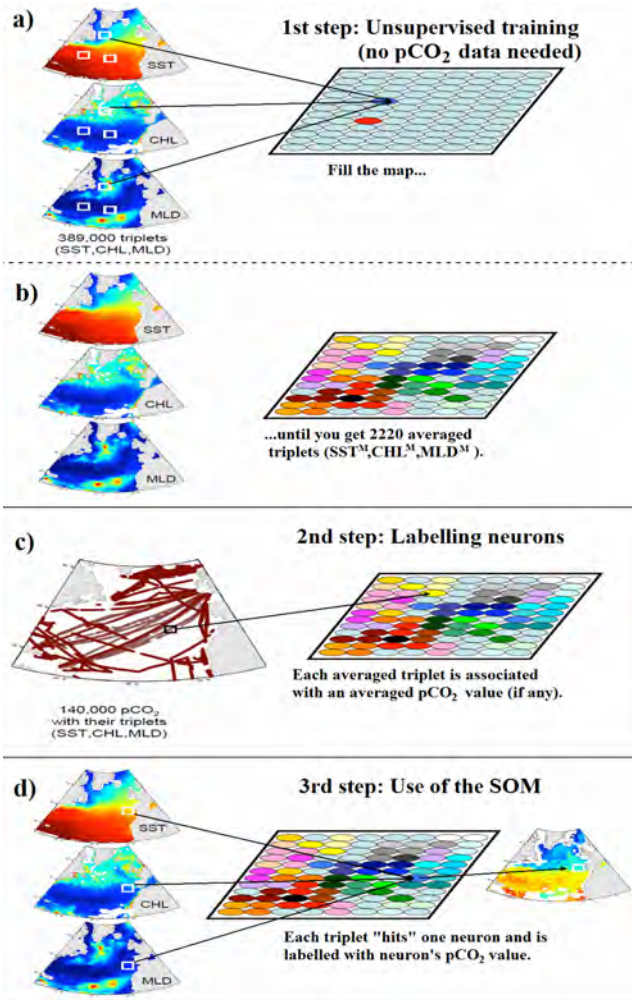

Fig. 1. Visualization of the procedures for the self organizing map (SOM). Three main steps are necessary: first (**a** and **b**), an unsupervised training takes place, and no pCO_2 data is used; second **(c)** preconditioned neurons are labelled with pCO_2 data measured in situ; third **(d)** the trained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

parameters were re-gridded onto a 8-daily frequency and 1° latitude×1° longitude resolution. The study area stretches from 10.5° N to 75.5° N and from 9.5° E to 75.5° W and is hereafter called the North Atlantic.

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

Seasons are 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 training vectors $\boldsymbol{x_i}$, are presented to the SOM (Fig. 1a). The activation of each neuron's weight vector, $\boldsymbol{y_i}$, is computed for the presented training vector. For a given training vector, the "winning" neuron (the one with the highest activation) is the one whose weight vector is the closest to the presented training vector in Euclidean distance $D(x_i, \boldsymbol{y_i})$, defined as:

$$D(x_i, \boldsymbol{y_i}) = \Big[(x_{i\,SST} - y_{\mathrm{iSST}})^2 + (x_{\mathrm{iMLD}} - y_{\mathrm{iMLD}})^2 + (x_{\mathrm{iCHL}} - y_{\mathrm{iCHL}})^2\Big]^{0.5} \quad (2)$$

The weight vector of the winning neuron is updated by adjusting it towards the training vector by a certain fraction of the difference between the two, as indicated by a linear, monotonically time-decreasing learning rate function α. 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 neurons in the neighbourhood of the winner are also stretched towards the input vector, according to a neighbourhood function, H, which decreases with each neuron away from the winner:

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

where $\sigma(t)$ is the neighbourhood radius at time t, and d_{ci} is the distance between map units c (winner) and i on the map grid. The neighbourhood radius $\sigma(t)$ decreases as a function of time along with the learning rate $\alpha(t)$. The neighbourhood radius and also 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. The learning rule which incorporates such a neighbourhood function leads to a topologically ordered mapping of the input vectors and distinguishes the SOM from other vector quantization algorithms (Kohonen, 2001). 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 SOM-map, while dissimilar patterns are mapped further apart.

After the training, each neuron becomes a synthetic sample with an associated weight vector (Fig. 1b). Every weight vector has a different combination of components, therefore the SOM estimates are based on 2220 relationships between the three training parameters. To account for strong nonlinearities in the real system it is important that the frequency

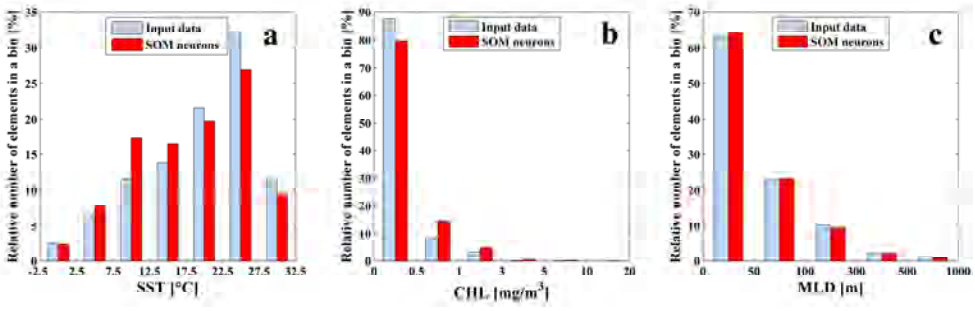

Fig. 2. The frequency distribution of each training parameter in the input data and the SOM neurons. Note that scales in the x-axes in **(b)** and **(c)** are not increasing monotonically.

distribution of each parameter in the input data is well represented by the SOM neurons (Fig. 2). Mean values of parameters (SST_{SOM}=18.1, SST_{INPUT}=19.4; MLD_{SOM}=63.6, MLD_{INPUT}=66.3; CHL_{SOM}=0.37, CHL_{INPUT}=0.27) and their ranges are also similar in the training data and SOM neurons. More importantly, neurons follow highly non-linear relationships between each pair of components in the training data set (Fig. 3a–c) visualizing how well the SOM is equipped for such a complicated setup. The distribution of neurons generally follows the training data distribution, even in such an extreme case as MLD versus SST (Fig. 3b).

2.4 Labelling the trained SOM with the pCO_2 data

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, all accompanied by corresponding SST, MLD and CHL values (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 CarboOcean, an EU-funded Integrated Project (http://www.carboocean.org). A total of 137 000 pCO_2 data points were collected on several vessels routinely crossing the North Atlantic between June 2004 and October 2006.

2.4.1 Distribution of the in situ measurements

The data in the labelling set is not evenly distributed in time and space (Fig. 4). More measurements are available in spring and summer than in fall and winter (Fig. 5) as a result of difficulties related to sampling in stormy winter waters. For the three years there are less than 3500 measurements between November and January. Few data are available for 2004: a major contribution in June, 2 days in July and 7 days in October make that year's input rather imbalanced.

Such an uneven distribution would make this data very difficult to analyse using traditional statistical techniques. Most linear methods would be biased towards summer waters, and the exceptionally high volume of data from June 2004 would create mapping discrepancies. In contrast, the ranges of input parameters captured in the labelling data set (pCO_2, SST, MLD, CHL) are more relevant for SOM estimates than their temporal and spatial distribution. This is one of the advantages that the SOM has over the other interpolation techniques. However, 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).

The training data set (as introduced in Sect. 2.2) offers wide ranges for all parameters, providing sufficient information about their variability as summarized in Table 1. The SST varies between −1.8°C and 30°C, the depth of the mixed layer ranges from ∼10 m to more than 1000 m (in total, 0.15% of data has MLD values greater than 1000 m) and chlorophyll-*a* concentrations vary from 0 to ∼10 mg/m^3 (0.04% of data has CHL values greater than 10 mg/m^3). The labelling data set captures 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 data set. Most of this difference is due to the fact that there are few in situ measurements from ice-melting zones, where water temperature drops below 0°C. Those regions are negligible in terms of the area covered, and the number of below 0°C 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 data set. In winter 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. 5); 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, and the deepest MLDs are not measured. As a result the SOM output is potentially biased towards shallower mixed layer depths in all regions

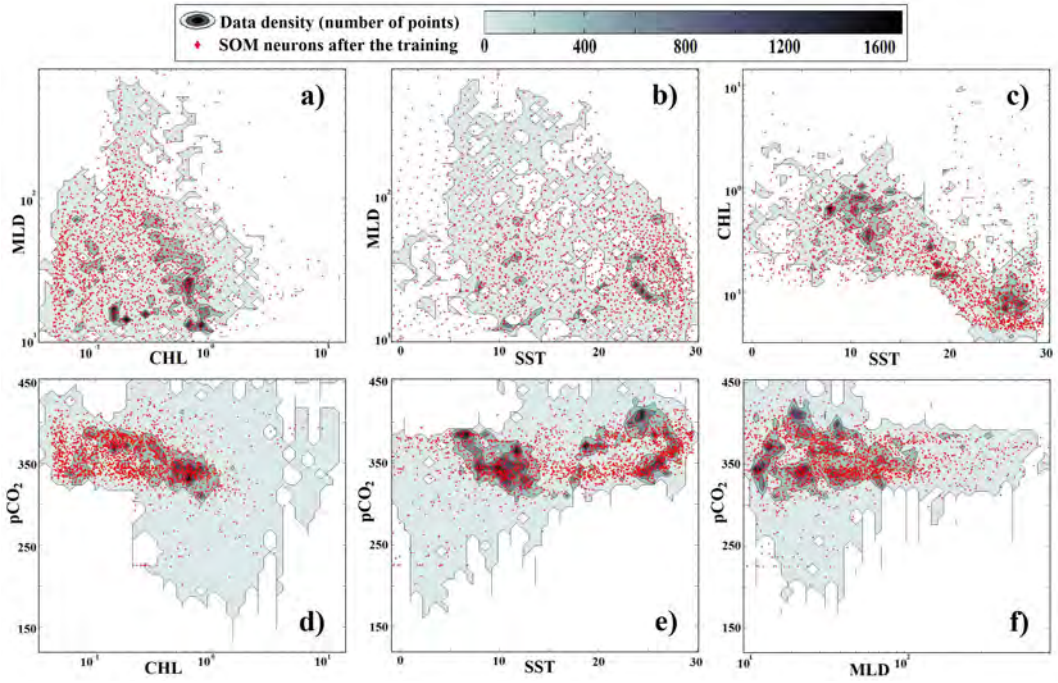

Fig. 3. (a–c) property – property plots for MLD, CHL and SST. The distribution of the density of the 389 000 training data points within each two-dimensional data space is shown in grey. Overlaid in red is the distribution of 2220 SOM neurons after the training. **(d–f)** property – property plots for pCO_2 and CHL, SST and MLD. The distribution of the density of the 137 000 labelling data points within each 2-dimensional data space is shown in grey. Overlaid in red is the distribution of 2220 SOM neurons after the training.

Table 1. Ranges of sea surface temperature, mixed layer depth and chlorophyll-*a* in the training (T) and labelling (L) data sets by season. Percentage of the training data within the range of the labelling data set is given for each parameter (L cover).

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	>10[c]	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[b]	99.2	0.02	>10[c]	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	>10[c]	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	>10[c]	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] 0.15% of the training data has MLD values greater than 1000 m.
[c] 0.04% of the training data has CHL values greater than 10 mg/m^3.

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. The exponential character of the relationship between sea surface pCO_2 and MLD in the subpolar North Atlantic (Olsen et al., 2008) suggests that MLDs deeper than 500 m have little influence on sea surface pCO_2 (their Fig. 9a). A similar relationship was found for the subtropical North Atlantic in our labelling data set (not shown) with a threshold value of 200 m.

The chlorophyll-*a* concentrations in the labelling data capture most of the variability between 2004 and 2006. The seasonal maxima between 2 mg m^{-3} in winter and 27 mg m^{-3} in fall suggest that even the strongest blooms are represented.

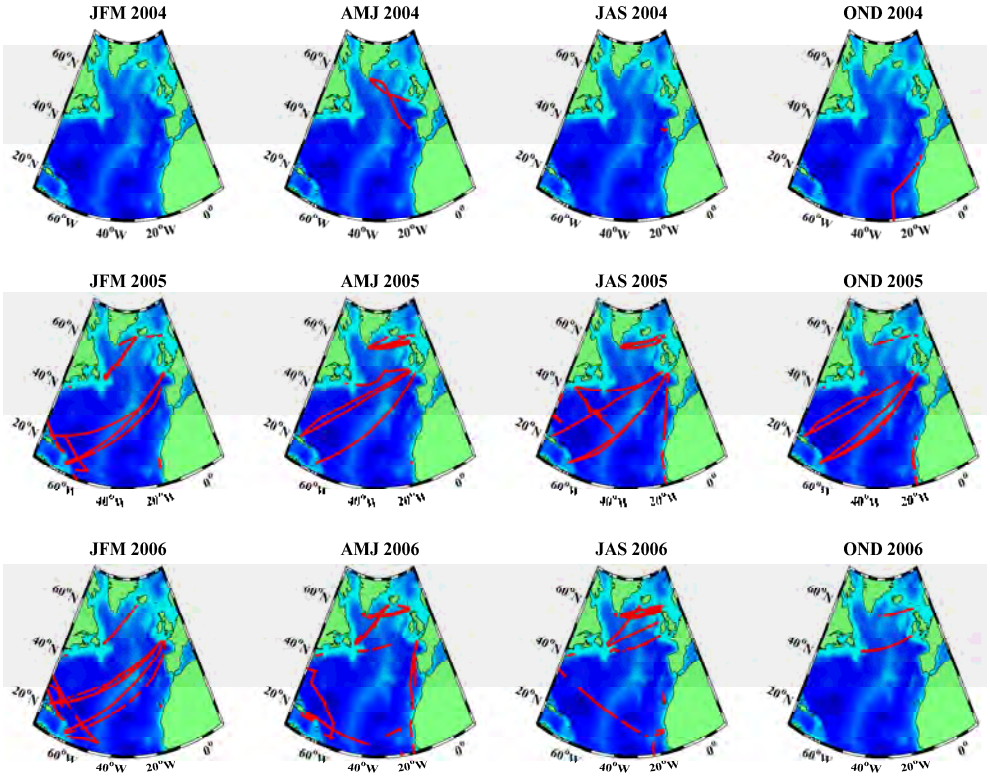

Fig. 4. The spatial distribution of the pCO_2 measurements used in this study. Data constitute the subset of the CarboOcean dataset for 2004 through 2006. Quarterly plots show data for January to March (JFM), April to June (AMJ), July to September (JAS) and October to December (OND).

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. Additionally, chlorophyll *a* data in both data sets are affected by the lack of satellite coverage north of ∼45° N in December and January.

2.4.2 The labelling procedure

Each data point from the labelling set is presented to the already trained SOM as an input vector (Fig. 1c). The winning neuron is found according to Eq. (2). 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 the in situ pCO_2 measurements and each individual component (SST, MLD and CHL) of the associated vectors are strongly non-linear from the basin-wide, year-long perspective. Figure 3d–f shows how the density distribution of the SOM neurons follows the density distribution of the pCO_2 data. Neurons concentrate where data density is highest and there is little data space

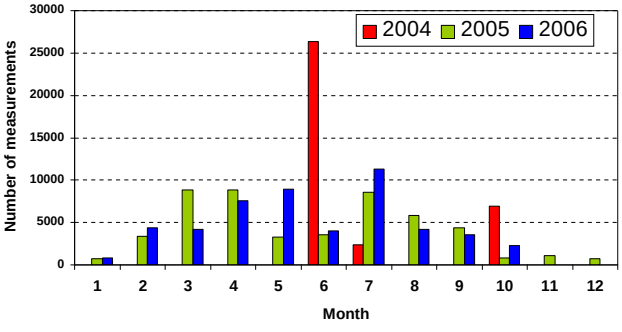

Fig. 5. Number of in situ pCO_2 measurements in the North Atlantic used for labelling the preconditioned SOM versus month for 2004, 2005 and 2006.

(grey colour) not accounted for during the labelling (red diamonds), meaning that SOM produces 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 could suggest that parameters additional to those considered in this study control the distribution of pCO_2 in the North Atlantic.

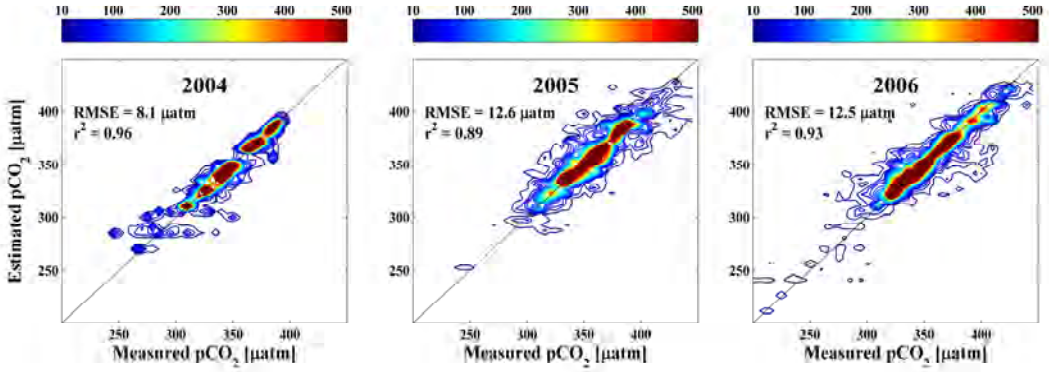

Fig. 6. Comparison of the SOM estimates with the measured pCO_2 for 2004, 2005 and 2006 as data density contours.

2.5 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 from June 2005 are presented to the preconditioned (trained and labelled) SOM. Each input vector is labelled with the pCO_2 label of the winning-neuron. 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, 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 Uncertainty estimate

For each in situ pCO_2 measurement, the corresponding SOM pCO_2 estimate was found based on spatial (1° longitude×1° latitude grid) and temporal (8 day intervals between 1 January 2004 and 31 December 2006) coordinates. The residual r value was calculated as a difference between the two. The Root mean-square error (RMSE) of the residuals calculated as:

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

for the whole dataset (n), provides an estimate of the uncertainty of the method in reproducing the available in situ measurements, and equals 11.6 μatm, or 3.2% of average pCO_2 in the in situ dataset.

Contour plots (one for each year) of estimated values, versus measured pCO_2 are shown in Fig. 6. The RMSE varies from 8.1 μatm in 2004 to 12.6 μatm in 2005 and 12.5 μatm in 2006. The distribution of the data density in all three years indicates that no systematic bias exists in the method. The values are scattered around the identity line for which the correlation coefficient is 1.

In a recent study, Friedrich and Oschlies (2009) derived the basin-wide monthly maps of pCO_2 in the North Atlantic for 2005 from modelled pCO_2 distribution using the SOM approach (they call it KFM). These authors report the basin-wide RMSE of 21.1 μatm. Such a relatively high error results mainly from the employed SOM training procedure, which is fundamentally different to that used in our study. As described in Sect. 2.3 we use three years of the whole grid data (SST, MLD and CHL) to train the SOM. This way the SOM "sees" the relationships between the training parameters in every grid point in the North Atlantic, with weekly frequency for the three years. This enables maximum SOM efficiency, regardless of the spatio-temporal cover of the in situ measurements used for labelling, and ensures that the SOM has been preconditioned with comprehensive, basin-wide training knowledge with regards to the relevant biogeochemical processes. Friedrich and Oschlies (2009) decided to train the SOM only with values (SST, CHL) collected along the VOS lines in 2005 (2005 VOS coverage metadata was used to extract the values from the model output). Such a small data set carries limited training knowledge, despite the very successful data gathering campaign in 2005. Processes occurring in the vast extent of the basin are never sampled (and therefore not included in the training), and when they are sampled, it very often happens only a few times during the year (Friedrich and Oschlies, 2009; their Fig. 2 for monthly cover and Fig. 6 for seasonal cover). It is not surprising that such trained SOM produces poor estimates for regions biogeochemically different than those sampled for the training data (their Fig. 6).

Moreover, our RMSE of 11.6 μatm relates the SOM estimates to data points along the pCO_2 sampling network (VOS lines), whereas in Friedrich and Oschlies (2009), the basin-wide pCO_2 distribution in the North Atlantic is known

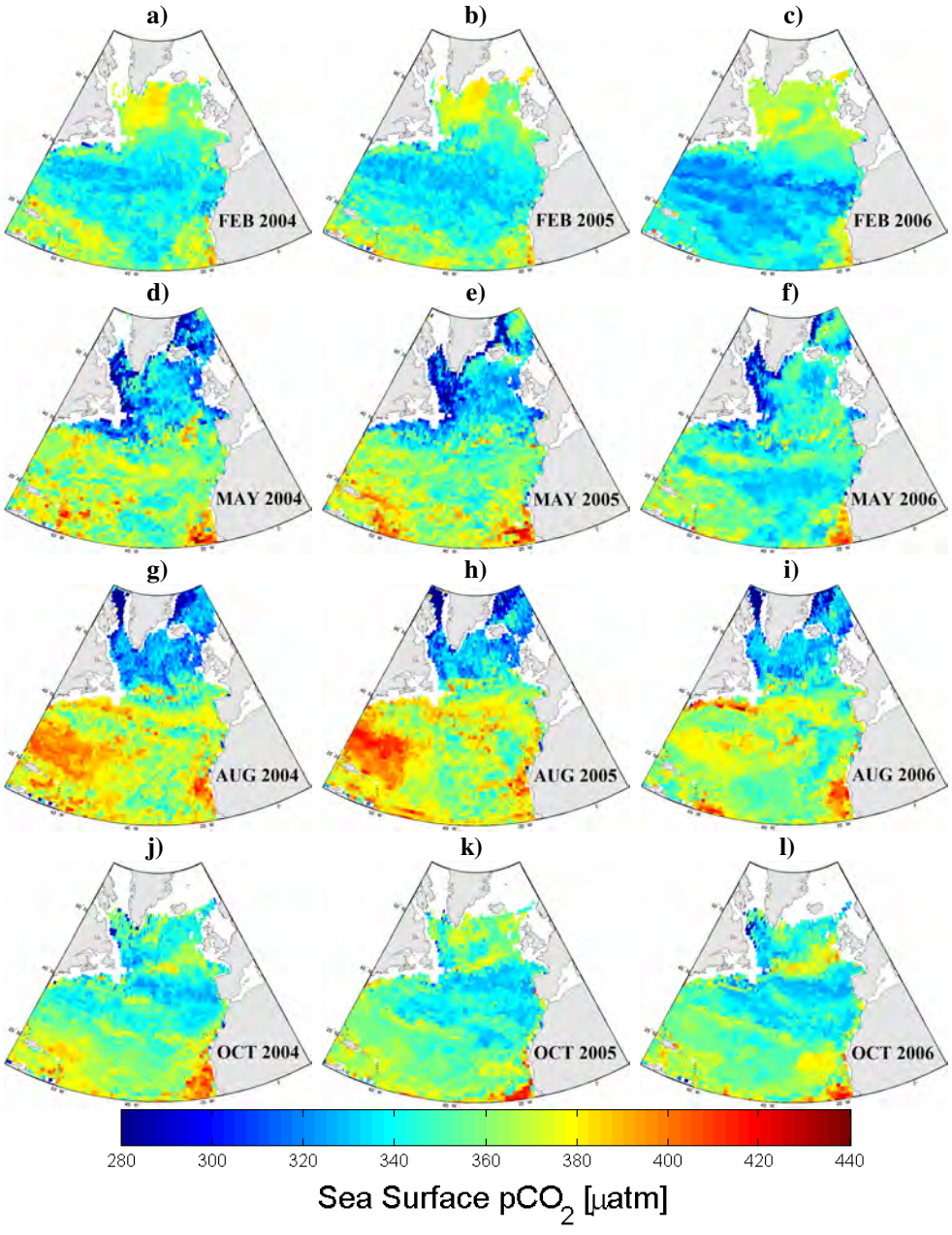

Fig. 7. Seasonal (in columns) and interannual (in rows) variability of the sea surface pCO_2 in the North Atlantic for years 2004 to 2006.

(since it is coming from the model) and therefore, they can report a basin-wide RMSE. Their equivalent of our "along the lines" RMSE equals 6.3 μatm from which they conclude that "along the VOS lines" RMSE is not representative of the basin-wide error (their basin-wide error is around three times higher than "along the lines" error). We suggest that the two RMSE estimates (along the VOS lines and basin-wide) are much more closely related if the training scheme employed in this study is used. The SOM, by definition, cannot reliably estimate output values for input values from outside the training data range, and this is essentially what Friedrich and Oschlies (2009) tested. This shows that the SOM can be applied to the data in at least two very different ways and careful choice of the training procedure is crucial for successful application.

3.2 Monthly pCO_2 maps

Out of the 36 monthly pCO_2 maps, one representing each season for the three years is shown in Fig. 7. In the columns are three seasonal cycles and in the rows maps for 2004, 2005 and 2006 showing SOM

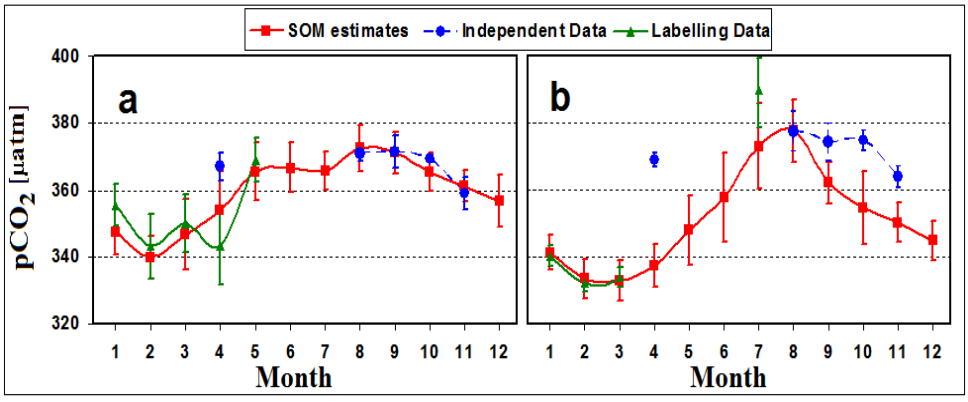

Fig. 8. Monthly mean pCO_2 versus month in 2006 estimated by SOM for two regions where independent (not used in SOM labelling) data from *MV Santa Maria* are available. The regions are well within two biogeochemical provinces: **(a)** the tropical North Atlantic (15° N to 25° N and 50° W to 60° W) and **(b)** the western subtropical North Atlantic (26° N to 38° N and 35° W to 60° W). For comparison, monthly means of pCO_2 from the labelling data in 2006 are also shown. The vertical bars extend from -1σ to $+1\sigma$ of the area weighted distribution for the given region and month.

estimates of the interannual variability. For the full set of monthly maps see Supplement http://www.biogeosciences.net/6/1405/2009/bg-6-1405-2009-supplement.pdf.

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. 7d–i), while values are around 20 μatm lower during fall and winter (Fig. 7a–c and j–l), confirming 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. 7d–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. 7j–l). Relatively high pCO_2 values in the northern part of the basin in winter (Fig. 7a–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 current SOM set up (Fig. 7a–c and j–l). The phytoplankton activity during late fall and winter in the northern 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. The pCO_2-SST/MLD relationship is strong in regions and periods where the productivity is low (Jamet et al., 2007), therefore pCO_2 estimates computed using SST/MLD trained SOM could be used to fill the gaps in the northern North Atlantic with coherent values during fall and winter.

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 increase of pCO_2 values in this region induced by this upwelling (Pelegri et al., 2005) is recognized by high pCO_2 estimates at around 20° N to 25° N in summer (Fig. 7g–i), and at around 10° N to 15° N in winter (Fig. 7a–c).

3.3 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 cycle of pCO_2 in its most prominent regions. We first compare SOM estimates in two regions within two biogeochemical provinces (Longhurst, 2007) to an independent in situ data set. We then compare the SOM results to climatological results in five major biogeochemical provinces of the basin.

3.3.1 Comparison to an independent data set

Five months of pCO_2 data collected between the UK and the Caribbean on board the *MV Santa Maria* during 2006 were not included in the labelling data set. Monthly means of this independent data for two regions are shown in Fig. 8. The data used for labelling the SOM for these regions in 2006 are

also plotted. SOM estimates are the area weighted means for the regions.

In the tropical North Atlantic (15° N to 25° N and 50° W to 60° W, Fig. 8a), the RMSE between the SOM estimates and the independent data equals 6.3 μatm, which relates to RMSE of 7.9 μatm for all the labelling data in the region. The independent data for one month (April) of the five months falls outside the 1-σ standard deviation of the SOM-predicted values for the region. Given that, based on Gaussian probability distribution, 68% of independent values should fall within such error bars, the SOM performs well in this direct validation exercise. Interestingly, the labelling data for April is on average 25 μatm lower than the independent data. Both voyages (one in the labelling data set and one in the independent data set) used for calculating averages took place between 18 and 30 April 2006, within the same 10° longitude by 10° latitude region. The tracks crossed (4 days apart) and measured pCO_2 values differed by around 20 μatm at the crossing point. Such high spatial and temporal variability complicates comparing a small number of in situ data from a specific sampling region and period to results obtained with interpolation methods designed for a much larger area.

In the western subtropical North Atlantic (26° N to 38° N and 35° W to 60° W, Fig. 8b), the RMSE between the SOM estimates and the independent data equals 19.3 μatm. The RMSE for all the labelling data in this region is also relatively high (14.3 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 high April values for the region. The sparseness of the in situ data (not shown) in the box may introduce sampling bias and may explain some of these differences. In addition, we suggest that SOM based maps should be used with caution when analyzing fine scale features and processes. Additionally, the RMSE for labelling and independent data in these regions are similar to each other. This confirms our hypothesis that the RMSE which relates the SOM estimates to data points used for labelling is representative of the basin-wide error, providing that sufficient data is available.

3.3.2 Comparison to the climatology

The robustness of SOM estimates is further assessed for five biogeochemical regions similar to those proposed by Longhurst (2007), as shown in Fig. 9. 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) ranges from 46° N to 58° N and from 10° W to 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

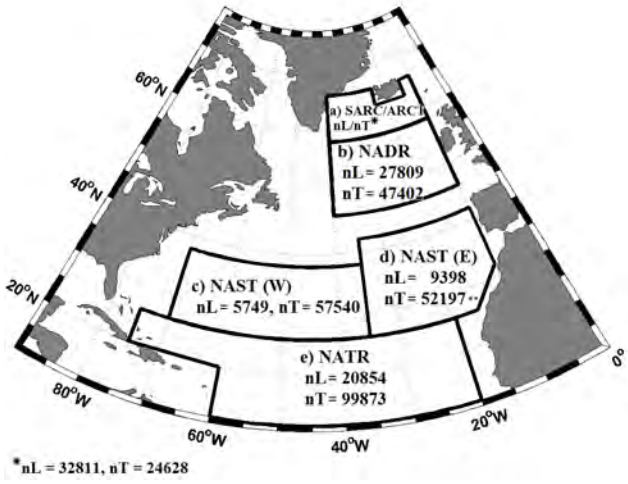

Fig. 9. The biogeochemical provinces of the North Atlantic proposed by Longhurst (2007) as used here for an analysis of SOM estimates. nT and nL represent the number of data points available for training and labelling of the SOM respectively.

35° W. The North Atlantic Tropical Gyre (NATR) stretches from 10° N to 26° N and from 20° W to 60° W (plus the region from 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 (Fig. 9). The mid-latitude North Atlantic has the smallest number of in situ measurements, whereas the northern provinces were by far the most sampled. The RMS errors for provinces do not correspond to such sampling distributions. The RMSE for the most sampled SARCT/ARCT amounts to 9.7 μatm, whereas almost four times less sampled NAST(E) has an RMSE of 7.2 μatm. In situ measurements are not used during the training of the SOM and their spatial distribution is irrelevant to the performance of the method. The number of points in the training data depends mainly on the size of the province. The chosen data sources offer year-round coverage except for the occasional lack of chlorophyll measurements in the SARC/ARCT region in winter.

In Fig. 10, 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 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., 2009). For comparison purposes we adjust the climatological distribution of Takahashi et al. (constructed for a reference year 2000) to a reference year 2005 assuming an annual rate of increase of 1.8 μatm as proposed in Takahashi et al. (2009). The original and adjusted distributions are plotted.

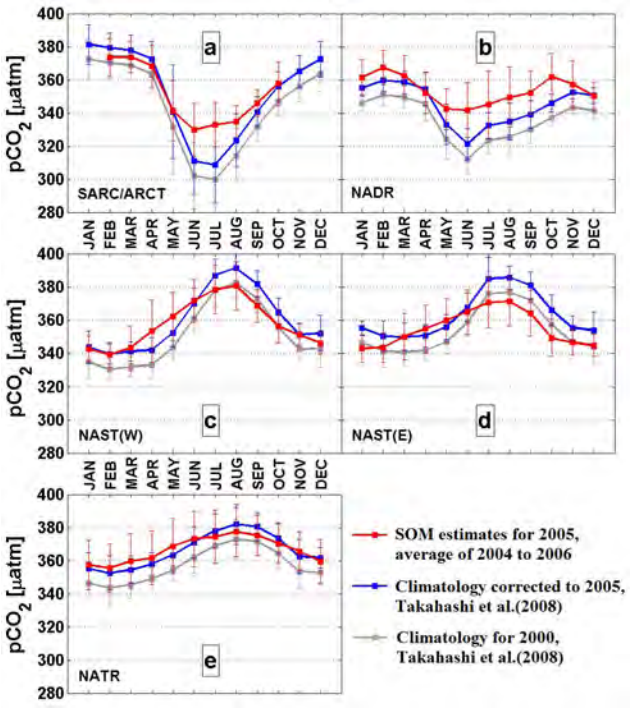

Fig. 10. Seasonal cycle of the sea surface pCO_2 in five biogeochemical provinces of the North Atlantic. The vertical bars extend from -1σ to $+1\sigma$ of the area weighted distribution for the given region and month.

3.3.3 SARC/ARCT region

In the SARC/ARCT region (Fig. 10a), the SOM estimates pCO_2 values of 330–340 μatm during late spring and summer and of around 370 μatm during fall and winter. These estimates agree with earlier findings showing that the 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), and higher winter values (caused by deepening of the mixed layer supplying CO_2 rich waters to the surface) dominate the seasonal cycle. The SOM estimates resolve such a pattern for the region. SOM values for the summer months are around 20 μatm higher than the long term climatology (Fig. 10a). 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 biological carbon uptake and thus underestimate the pCO_2 values in summer.

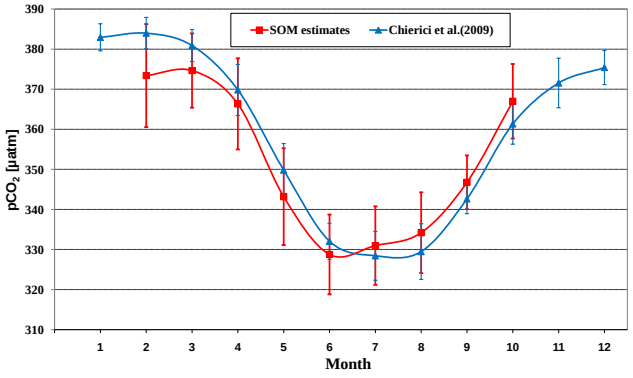

Fig. 11. Monthly, area weighted pCO_2 fields during 2005 in the subpolar North Atlantic (58° N to 63° N, 10° W to 40° W) estimated by the SOM, compared to the multiple regression estimates for 2005 in the same region after Chierici et al. (2009). Note that this region does not extend as far north as the SARC/ARCT region.

The SOM estimates for the region are consistent with those of Chierici et al. (2009) (Fig. 11). These authors estimated sea surface pCO_2 for 2005 in a region with a slightly smaller latitudinal extent than our SARC/ARCT region (Fig. 11 shows SOM estimates for the region used by Chierici et al.). Using pCO_2 measurements obtained onboard *MV Nuka Arctica*, together with remotely sensed data, they applied algorithms based on multiple linear regression. The in situ pCO_2 measurements used by Chierici and coworkers represent a fraction of the dataset used in this study. The resulting seasonal cycle for 2005 agrees well with the SOM estimates for 2005. 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. The similar results of the multiple linear regression (designed for regional, high-resolution estimates) and the SOM, increase our confidence in the basin-wide SOM estimates, despite the fact that the pCO_2 data used have a large overlap.

3.3.4 NADR region

In the NADR (Fig. 10b), 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 northwest relative to that in Schuster and Watson (2007).

3.3.5 Subtropical North Atlantic

The seasonal cycle in the subtropical North Atlantic, represented here by two provinces (Fig. 10c–d), has an opposite shape to that further north. SOM pCO_2 estimates in the NAST(W) are characterized by a strong summer maximum in August, which corresponds to 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 thermodynamically driven variability in surface water pCO_2 (Bates, 1998, 2001).

The SOM estimates for the NAST(E) have a relatively low annual amplitude mainly due to an 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. (2009) 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., 2009) This is in line with the Takahashi climatology adjusted to 2005, which estimates a summer maximum of 386 μatm (Fig. 10d). The SOM however, estimates 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. Altering the shape of the neighbourhood function during the training phase, slightly increases the ability of the SOM to better mimic the extreme pCO_2 values. According to sensitivity studies on the choice of neighbourhood function in extracting the known patterns, the Gaussian neighbourhood function (used in this study) returns the smoothest SOM patterns, while the Epanechikov (ep) neighbourhood function reproduces the most extreme values the most accurately (Liu and Weisberg, 2005; Liu et al., 2006b). However, the real benefit of using "ep" neighbourhood function in the current study, although not negligible, is relatively minor. Our simulations suggest that the monthly pCO_2 values for July and August in the eastern subtropics increase by 1–2 μatm when an "ep" neighbourhood function is used instead of the Gaussian neighbourhood function. The SOM estimates are still more than 10 μatm lower than other reports suggest for July through September in the NAST(E), and other causes for the SOM to underestimate the highest values will be investigated. 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 variability in sea surface pCO_2 not determined by changes in SST, MLD and CHL. Additionally, an increase in the spatial and temporal resolution of the training data to reduce the effect of averaging on SOM estimates may help overcome this local "smoothing" effect of the SOM.

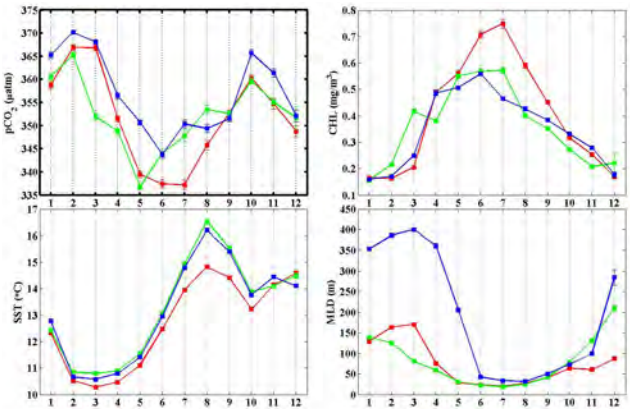

Fig. 12. SOM estimates of the interannual variability of pCO_2 during 2004 (red), 2005 (green) and 2006 (blue) in the NADR province (46° N to 58° N and 10° W to 40° W) compared to the variability of SST, CHL and MLD.

3.3.6 NATR region

In the NATR (Fig. 10e) the SOM estimates a fairly flat seasonal cycle coupled to temperature variability, which agrees well with the climatology. These warm, relatively oligotrophic waters do not support much primary production (Longhurst, 2007), and lack of strong winds reduces mixed layer deepening as a control on the variability of pCO_2. Values are relatively high throughout the year and vary between 355 and 380 μatm. The West African upwelling brings CO_2-rich waters to the surface, thus increasing the sea surface pCO_2 values, especially during summer.

Overall the SOM proves a robust method for reconstructing seasonal pCO_2 cycles in a diverse suite of biogeochemical provinces in the North Atlantic.

3.4 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 location, by interannual variability. The SOM's basin-wide estimates of such variability for each season are presented in Fig. 7. Visual inspection of this three-year period reveals apparent year-to-year changes. In the western subtropics August values (Fig. 7g–i), are highest for 2005. Also the region of high (∼400 μatm) pCO_2 values covers a larger area than in either 2004 or 2006. Similarly, in the North Atlantic Drift Region, October pCO_2 values (Fig. 7j–l) in 2006 are higher and more extensive than in the two previous years.

According to Eq. (1), the SOM predictions are entirely data-based, and therefore the interannual variability in the SOM estimates can only be forced by the interannual

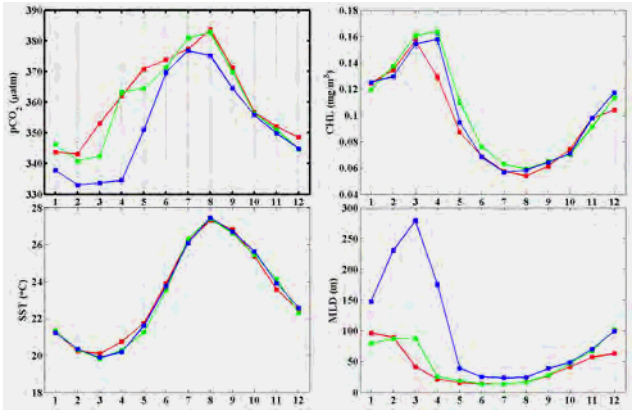

Fig. 13. SOM estimates of the interannual variability of pCO_2 during 2004 (red), 2005 (green) and 2006 (blue) in the NAST(W) province (26° N to 38° N and 35° W to 70° W) compared to the variability of SST, CHL and MLD.

variability of the three training parameters: SST, CHL and MLD. Their variability affects the pCO_2 distribution in a non-uniform manner, varying with the spatio-temporal dependence of pCO_2 on the given parameter in each region. Figures 12 and 13 show these relationships for the 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 the standard error of the mean calculated as:

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

where σ represents the standard deviation of the sample, and n is the sample size.

In the NADR (Fig. 12), the interannual variability of CHL controls the variability in the estimated pCO_2 field during months with high mean CHL concentrations. High mean CHL values (0.55–0.75 mg/m^3) during June–August 2004 correspond to pCO_2 values which are 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 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 into similarly substantial increase in pCO_2. This may be explained by the non-linear relationship between sea surface pCO_2 and MLD in the subpolar North Atlantic proposed by Olsen et al. (2008) and confirmed by our data (not shown). Olsen et al. (2008) found that MLDs deeper than 500m have no or little influence on the sea surface pCO_2 (their Fig. 9a), whereas an MLD increase between 0 and 200m corresponds to a pCO_2 increase of up to 100 μatm. This relationship is influenced by the variability in SST and CHL, but the SOM estimates appear coherent with findings by Olsen et al. It is also worth noting that the 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 on pCO_2 during the bloom period (Olsen et al., 2008), and SOM estimates follow this relationship.

In the NAST(W) (Fig. 13), the estimated interannual variability in pCO_2 for the years 2004 to 2006 appears to be controlled by the variations in the MLD. The seasonal pCO_2 cycle in this region is controlled by the combination of SST and MLD. However, for the years 2004–2006 the interannual variability of SST is too weak to contribute to the interannual variability in pCO_2. For all three years, the MLD strongly contributes to the parameterization in winter with the mean regression coefficient of −0.42. It is, however, remarkable that the deep MLD in the winter of 2006 does not translate into a relative change of SST. Also, the lower pCO_2 for the higher MLD in January–March 2006 is difficult to explain without changes in temperature (despite the negative coefficient for MLD). The non-linear relationship between the sea surface pCO_2 and MLD found in the subpolar North Atlantic (Olsen et al., 2008), occurring between the deepening of MLD in fall and the beginning of the bloom in spring, is also apparent in the subtropical part of the basin. The SOM estimates resolve this relationship, which can be observed especially during late winter and early spring, when SST and CHL vary relatively little (Fig. 13). During the January–March period of 2004, the MLD was variable and shallow in the subtropics (a decrease from 96 m in January to 41 m in March), and the SOM predicts variable pCO_2 as a result (an increase of 9.3 μatm). During February to April of 2005 the MLD was similarly shallow and variable (62m decrease), and the SOM predicts a 12.3 μatm increase in the pCO_2. Contrary to that, much greater variability of the much deeper MLD during the February–April period of 2006 translates to 1.5 μatm variability in the predicted pCO_2. Changes in deep MLDs result in less pCO_2 variability than similar or smaller changes in the shallow MLDs.

4 Summary and conclusions

A self organizing neural network has been applied to construct 36 basin-wide, monthly pCO_2 maps for 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 agree 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., 2009) indicates. The difference is especially profound in the June–September phytoplankton bloom period. 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, one with and one without CHL as training parameter, thus 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) mission, planned between July and October 2009. However, the influence of the smoothing effect on the overall performance seems to be local and mainly related to the analyses of sub-mesoscale features. Very high spatial and temporal variability of in situ pCO_2 makes the current SOM estimates too coarse for small-scale analyses and they should only be considered for analyses over larger regions.

The pioneering data-based, basin-wide estimates of the interannual pCO_2 variability provide 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. The method can be used, therefore, to examine the interannual variability in the North Atlantic over the last decade or so, during which the region has 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 in pCO_2 or disagree with in situ measurements. The sparse nature of in situ observations is often given as an explanation; therefore monthly basin-wide maps of pCO_2 as created by the SOM could serve as a better input for models.

Additionally, our SOM maps are a very promising input for monitoring the changing ocean CO_2 sink and source regions. Watson et al. (2009) use the SOM output together with geostatistical techniques to constrain the CO_2 uptake by the North Atlantic. They define annual uptake to an unprecedented precision of about 10%. Our future plans include extending their effort to the northern hemisphere and ultimately to provide SOM-based global flux estimates.

As a whole, the SOM approach presented 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 individual biogeochemical relationships. The SOM's ability to extract numerous existing relationships simultaneously provides a 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 of pCO_2 in the real ocean, and allow us to assess the quality of pCO_2 estimates with more confidence. It is our strong recommendation 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. We are grateful to three reviewers and Jack J. Middelburg for their constructive and critical comments. 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 Quéré (University of East Anglia, UK) for helpful comments and discussions. This work was funded by the European Commission CARBOOCEAN project GOCE 511176-2; the European Commision Marie Curie RTN Greencycles project (MRTN-CT-2004-512464); the Spanish project ICCABA CTM2005-03893/MAR; the Norwegian Research Council through A-CARB (178167) and CARBON-HEAT (185093) and the Swedish National Space Board through RESCUE – II (d.nr 62/07:1).

Edited by: J. Middelburg

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. Pt. 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., and 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. Pt. II, 56(8–10), 630–639, 2009.

Cooper, D. J., Watson, A. J., and Ling, R. D.: Variation of pCO_2 along a North Atlantic shipping route (UK 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.

Friedrich, T. and Oschlies, A.: Neural network-based estimates of North Atlantic surface pCO_2 from satellite data: A methodological study, J. Geophys. Res., 114, C03020, doi:10.1029/2007JC004646, 2009.

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, 2247–2257, 2007, http://www.ann-geophys.net/25/2247/2007/.

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. Meteorol. 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., 752 pp., Wiley-Interscience, 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., andSutherland, 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. H.: 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., Weisberg, R. H., and He, 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. H., 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., and 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. Pt. 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., 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. Pt. II, 56(8–10), 620–629, 2009.

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. Pt. 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. Pt. II, 56(8–10), 554–577, 2009.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankagas, J.: SOM Toolbox for Matlab 5., Libella Oy, Espoo, 59 pp., 2000.

Watson, A. J., Schuster, U., Bakker, D .C. E., Bates, N., Corbière, A., Friedrich, T., González-Dávila, M., Hauck, J., Heinze, C., Johannessen, T., Körtzinger, A., Metzl, N., Olafsson, J., Olsen, A., Oschlies, A., Padin, X. A., Pfeil, B., Santana-Casiano, M., Steinhoff, T., Telszewski, M., Ríos, A., and Wallace, D. W. R., Wanninkhof, R.: Accurately tracking the variation in the North Atlantic sink for atmospheric CO_2, Science, in review, 2009.