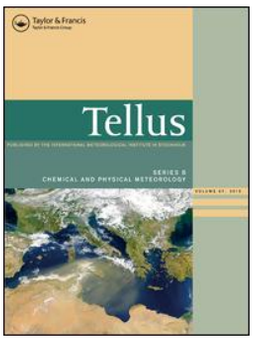
A comparison of multiple regressions and neural network techniques for mapping in situ pCO2 data

Nathalie Lefèvre, Andrew J. Watson, Adam R. Watson



Tellus B: Chemical and Physical Meteorology

ISSN: (Print) 1600-0889 (Online) Journal homepage: https://www.tandfonline.com/loi/zelb20

A comparison of multiple regression and neural network techniques for mapping in situ pCO_2 data

Nathalie Lefévre, Andrew J. Watson & Adam R. Watson

To cite this article: Nathalie Lefévre, Andrew J. Watson & Adam R. Watson (2005) A comparison of multiple regression and neural network techniques for mapping in situ pCO_2 data, Tellus B: Chemical and Physical Meteorology, 57:5, 375-384, DOI: 10.3402/tellusb.v57i5.16565

To link to this article: https://doi.org/10.3402/tellusb.v57i5.16565

Published online: 18 Jan 2017.

Tellus (2005), 57B, 375–384
Printed in Singapore. All rights reserved

TELLUS

A comparison of multiple regression and neural network techniques for mapping *in situ* pCO_2 data

By NATHALIE LEFÈVRE*†, ANDREW J. WATSON and ADAM R. WATSON, *University of East Anglia, School of Environmental Sciences, Norwich NR4 7TJ, UK*

(Manuscript received 20 September 2004; in final form 18 May 2005)

ABSTRACT

Using about 138 000 measurements of surface pCO_2 in the Atlantic subpolar gyre (50–70°N, 60–10°W) during 1995–1997, we compare two methods of interpolation in space and time: a monthly distribution of surface pCO_2 constructed using multiple linear regressions on position and temperature, and a self-organizing neural network approach. Both methods confirm characteristics of the region found in previous work, i.e. the subpolar gyre is a sink for atmospheric CO_2 throughout the year, and exhibits a strong seasonal variability with the highest undersaturations occurring in spring and summer due to biological activity. As an annual average the surface pCO_2 is higher than estimates based on available syntheses of surface pCO_2. This supports earlier suggestions that the sink of CO_2 in the Atlantic subpolar gyre has decreased over the last decade instead of increasing as previously assumed. The neural network is able to capture a more complex distribution than can be well represented by linear regressions, but both techniques agree relatively well on the average values of pCO_2 and derived fluxes. However, when both techniques are used with a subset of the data, the neural network predicts the remaining data to a much better accuracy than the regressions, with a residual standard deviation ranging from 3 to 11 μatm. The subpolar gyre is a net sink of CO_2 of 0.13 Gt-C yr^{-1} using the multiple linear regressions and 0.15 Gt-C yr^{-1} using the neural network, on average between 1995 and 1997. Both calculations were made with the NCEP monthly wind speeds converted to 10 m height and averaged between 1995 and 1997, and using the gas exchange coefficient of Wanninkhof.

1. Introduction

The North Atlantic subpolar gyre (>50°N) is a highly dynamic region where the North Atlantic Deep Water forms, with the main sources of this water mass being located in the Nordic and Labrador seas. This region is also biogeochemically important because it is one of the strongest sinks of carbon in the world's oceans, but it is complex due to its heterogeneity. In winter, deep convective mixing occurs. In spring, when the water stratifies, the spring bloom can cause surface pCO_2 to change rapidly. For example, Schneider et al. (1992) reported an opposite gradient of pCO_2 along 20°W in June 1991 from that of Watson et al. (1991) in May 1989. Shifting patterns of phytoplankton blooms (Lochte et al., 1993) are responsible for the patchiness of the chlorophyll concentrations and hence, the pCO_2 distribution.

In addition to the strong spatial variability of pCO_2, little is known about the variability of the CO_2 sink over time. The climatology of Takahashi et al. (1997) and its update (Takahashi et al., 2002) have been used by modellers to assess model performance in reproducing the seasonal variability and the annual averages in different regions of the oceans (J. Orr, personal communication/unpublished results). In order to produce such climatologies, data collected over several decades have to be combined into a single reference year. In the process, assumptions must be made about how ocean pCO_2 is changing in response to rising atmospheric CO_2. In low-latitude regions oceanic pCO_2 is assumed to follow the atmospheric pCO_2 increase. In subpolar regions it was assumed that surface pCO_2 would increase more slowly, or not at all, because of dilution with older water which has not previously been exposed to the atmosphere due to deeper mixing and wind-driven upwelling. Takahashi et al. (2002) assumed that surface pCO_2 did not change over time in the subpolar regions, based on observations at station P (50°N, 145°W) in the Pacific. No relevant observations were available in the subpolar Atlantic to check this assumption in that location.

There is a recognized need to be able to specify the air–sea flux of CO_2 over large areas of the ocean and on a seasonal or monthly basis. Such estimates are of value both in validating ocean carbon models and as input to atmospheric inversions which aim to constrain both land and ocean fluxes. To achieve accurate estimates of pCO_2 over ocean regions, however, requires some technique

*Corresponding author.
e-mail: nathalie.lefevre@lodyc.jussieu.fr
†Now at: LOCEAN, UMR 7159 CNRS/IRD/UPMC/MNHN, Université Pierre et Marie Curie, Tour 45-46 5ème étage, BP 100, 4 Place Jussieu, 75252 Paris cedex 05, France.

for interpolation of relatively sparse measurements both in time and space. In this paper we present and compare monthly surface pCO_2 maps for the subpolar gyre (restricted to 50–70°N, 60–10°W) for 1995 to 1997, derived by two techniques both using as input a data set of pCO_2, sea surface temperature (SST), time and position. The first technique uses multiple linear regressions of surface pCO_2 on SST and position, with separate regressions for each month; the second is a self-organizing neural network algorithm. Both types of algorithm are then used to produce maps of pCO_2 for each month and over the subpolar gyre, using reanalysed SST fields as input.

2. Materials and methods

2.1. Multiple linear regressions

As part of the CAVASSOO (Carbon Variability Studies by Ships of Opportunity) project, a relational database including all CO_2 data available in the North Atlantic ocean has been constructed (see http://tracer.env.uea.ac.uk/e072). The database contains over 138 000 data points in the subpolar gyre prior to 2001, of which 104 000 measurements were made between 1995 and 1997. As these CO_2 data are unevenly distributed in both time and space, we investigated robust linear regressions that allow interpolation in space and in time. Given the complexity of the oceanic pCO_2 variability these relationships are unlikely to be valid over the entire gyre. Thus, we divided the subpolar gyre into smaller regions where the pCO_2 distribution can be considered relatively uniform. We used the biogeochemical provinces defined by Longhurst et al. (1995) for this purpose. For the subpolar gyre they identify three main provinces: the North Atlantic Drift (NADR, between 44 and 58°N and 42 and 10°W), the sub-Arctic (SARC, between 58 and 66°N and 24 and 10°W) and the Arctic (ARCT, the remaining area between 50 and 70°N and 60 and 10°W) regions (Fig. 1). Processes occurring in the ARCT region are supposed to be valid for a wide range of latitudes (50–70°N) according to the definition of the province by Longhurst et al. (1995). However, the temperature range is quite large. We therefore divided this province into three by defining ARCT1 (50–58,°N, 60–42°W) as an extension in longitude of the NADR province, ARCT2 (58–66°N, 60–24°W) as an extension of SARC, and ARCT3 (>66°N) as the most northern province. The NADR is by far the most sampled region, but in each region most of the data were collected between 1995 and 1997 (Fig. 2). In each biogeochemical province we sought to compute a linear relationship expressing sea water pCO_2 as a function of longitude, latitude, SST and year, for each month. First, the thermodynamic influence on sea surface pCO_2 was removed by normalizing to a constant temperature, $(pCO_2)_T$, using the relationship suggested by Takahashi et al. (1993). This temperature was chosen to be the monthly average temperature. Then we performed the multivariable linear regression

$$(pCO_2)_T = A + B + C + D + E. \quad (1)$$

The coefficients A (intercept), B (longitude), C (latitude), D (SST), E (year), the reference temperature, T, and the square of the correlation coefficient, r^2, are given in Table 1. In some cases there were not enough measurements over time to include the year as a variable but linear regressions were still performed without the year when the data were collected between 1995 and 1997. The seawater pCO_2 was then calculated at the *in situ* temperature and compared with the original data. Figure 3 shows an example of the comparison. The minimum and the maximum of the difference between the calculated pCO_2 and the original data are shown in Table 2. A mean error is also calculated by dividing the sum of the absolute value of the difference by the number of observations. The largest errors and variability are usually in spring and summer. The patchiness of the biological activity and its strong variability observed in this region (Schneider et al., 1992; Watson et al., 1991) makes it more

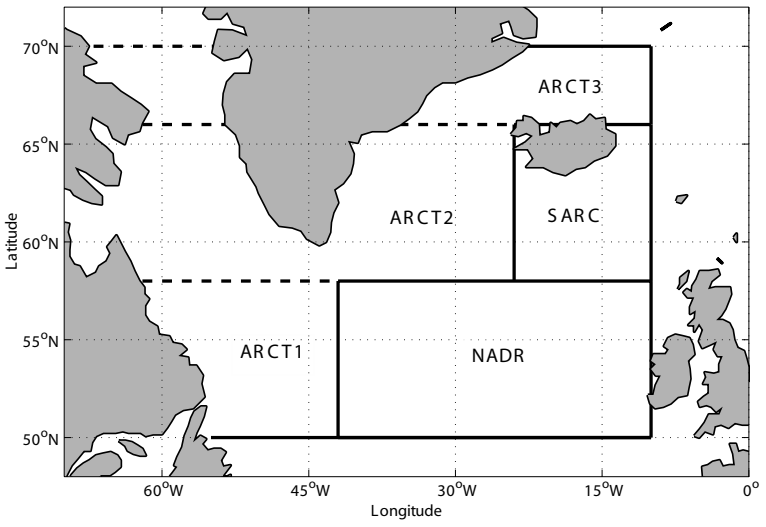


Fig 1. Map of the different regions of the subpolar gyre (>50°N): the North Atlantic Drift Region (NADR), the sub-Arctic region (SARC) and the three Arctic regions (ARCT1, ARCT2, ARCT3).

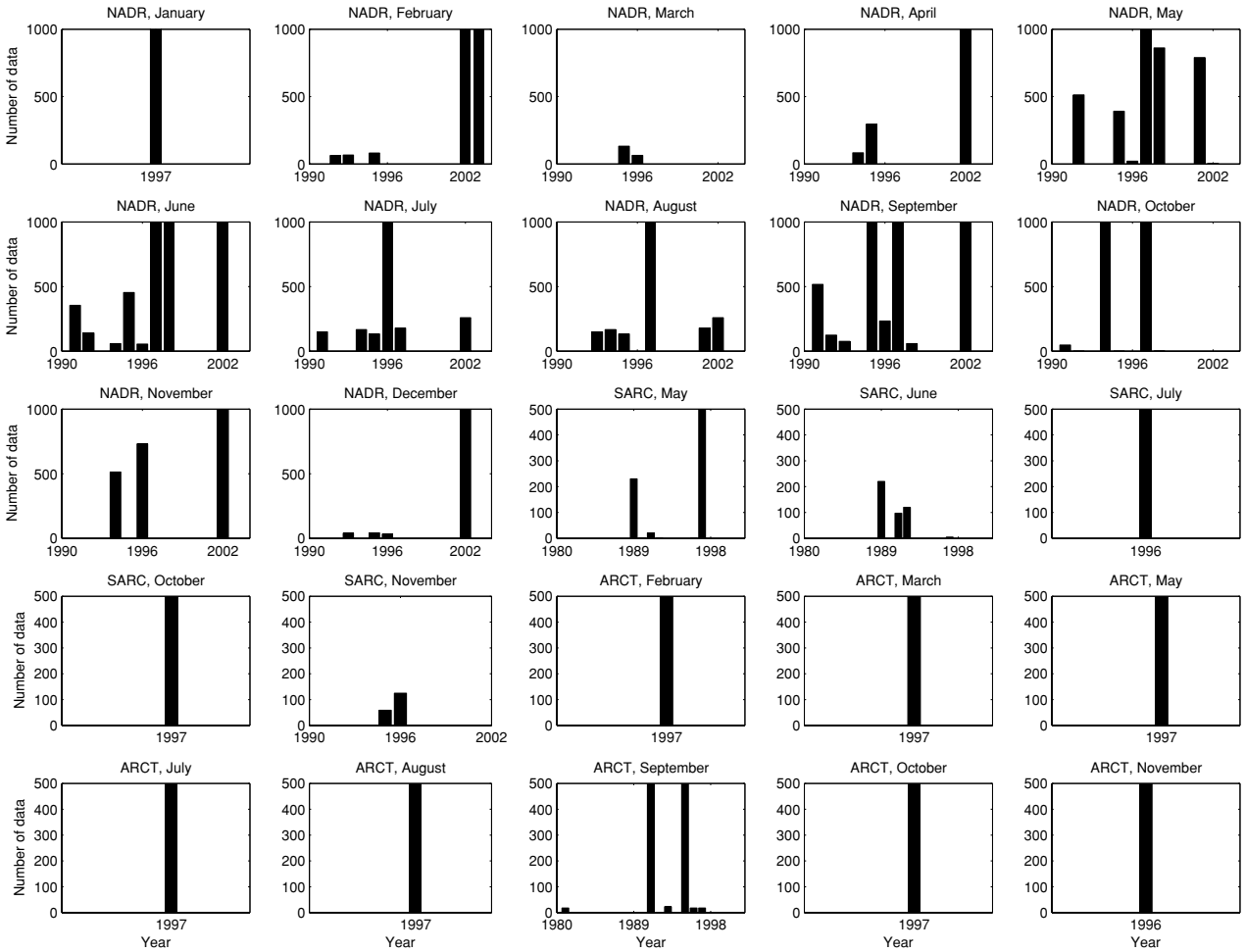


Fig 2. Monthly distribution of pCO_2 data as a function of year for the NADR, SARC and ARCT regions. For better clarity we show the number of data up to 1000 for the NADR region and up to 500 for the SARC and ARCT regions. Data collected in 1981 correspond to the TTO cruises and are located exclusively in the ARCT region. Most of the data were collected in the 1990s with a data record of 75 781 in 1997.

difficult to reproduce pCO_2 with a linear relationship and explains the larger discrepancy between the relationships and the original data for these seasons. Using the empirical relationships developed and the temperature from the NCEP/NCAR reanalysis project (see http://www.cdc.noaa.gov/cdc/reanalysis) from 1995 to 1997, pCO_2 fields were then constructed for each month of the year on a 1° by 1° spatial grid. The climatology is the average pCO_2 from 1995 to 1997. When observations are limited to a small area, the latitude and longitude coefficients are poorly constrained in the regression. In this case, we applied the empirical relationship to a limited location, for example in August the pCO_2 field was limited to the ARCT2 region because no measurements were available in ARCT1 and ARCT3. In order to produce maps for the whole subpolar gyre (50–70°N, 60–10°W) we reconstructed the missing regions by interpolating the pCO_2 distributions of the adjacent months. For example, in August the pCO_2 distribution in ARCT1 is obtained by averaging the pCO_2 distributions for the months July and September. However, we could not interpolate the SARC province in September from adjacent months because of the abrupt change in the pCO_2 due to the breakdown of stratification in October. This would lead to unrealistically high pCO_2 values in September compared with the surrounding provinces. Therefore, the SARC province was masked in September.

2.2. *Neural network: self-organizing maps*

Neural networks are often well-suited to generalizing tasks. The advantage of a neural network approach is that it can recognize and exploit relationships in the data which are not predefined (as in regression techniques) and need not to be expressible by any equation. This makes them particularly suited to mapping relationships that are non-linear and empirical, provided sufficient data are available to 'train' the network. A self-organizing network (Kohonen, 1984) was developed to examine the relationships between temperature, position and time in

Table 1. Coefficients of the regression of $(pCO_2)_T$ at the reference temperature as a function of longitude, latitude, SST and year, and coefficient of determination of the regression, for each month and region. The reference temperature is the monthly average temperature of the observations

Region	Month	T	Intercept A	Longitude B	Latitude C	SST D	Year E	r^2
NADR	Jan	14	565.748	−0.19009	0.540655	−18.5982	0	0.91
	Feb	13	−2488.39	−0.42224	4.976368	−12.2267	1.382684	0.90
	Mar	12	741.9574	−0.11432	−2.21914	−24.1972	0	0.86
	Apr	12	655.6648	−0.42368	−1.47395	−21.6593	0	0.72
	May	12	−7641.63	−0.90473	−1.74153	−20.7719	4.143123	0.90
	Jun	14	−4873.04	−0.8504	1.299685	−15.641	2.660618	0.82
	Jul	15	−7012.59	−0.02493	3.65915	−7.07028	3.635682	0.89
	Aug	15	−3160.22	−0.68692	0.841897	−11.3147	1.799032	0.95
	Sep	16	−1297.09	0.429922	−4.18782	−17.0603	1.054572	0.85
	Oct	15	83.44259	−0.80749	4.810524	−10.9162	0.076253	0.96
	Nov	14	747.2378	0.199037	−0.7298	−17.3013	−0.06186	0.98
	Dec	14	−4306.02	0.381974	−0.21887	−17.1293	2.453443	0.90
SARC	May	10	−2304.21	−1.99439	−7.11173	−27.2612	1.644646	0.88
	Jun	9	−6627.51	−8.05424	−14.9856	−22.7798	3.950624	0.67
	Jul	12	714.0046	0.905402	−2.34237	−19.0055	0	0.31
	Oct	10	917.1003	3.096269	−3.44246	−27.8699	0	0.97
	Nov	8	−15.4489	−1.05598	6.555539	−7.11285	0	0.97
ARCT	Feb	3	453.758	0.055869	−1.44727	−5.99434	0	0.60
	Mar	3	330.6054	−1.05773	0.108788	−11.5004	0	0.59
	May	9	−77.2792	0.318916	8.334795	−8.48854	0	0.76
	Jul	8	217.0376	1.121123	3.135003	−5.79794	0	0.53
	Aug	10	775.5434	−0.46273	−6.23804	−10.8788	0	0.58
	Sep	10	1504.744	0.541596	−2.61486	−14.3116	−0.44159	0.84
	Oct	8	246.2779	−0.87876	2.801472	−11.761	0	0.95
	Nov	5	152.0971	−1.26037	2.721852	−5.05919	0	0.95

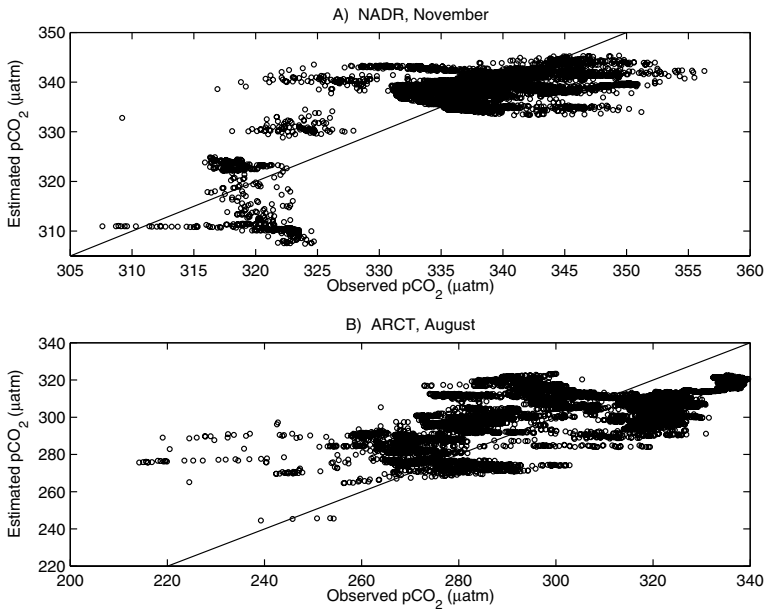


Fig 3. Comparison of estimated pCO_2 from the regressions with the observations for (a) NADR in November and (b) ARCT in August.

Table 2. Minimum and maximum of the difference between the calculated pCO_2 and the original data, number of data used for the regression, sum of the absolute value of the residual and mean error for each month in each province where a regression was performed

Region	Month	Min	Max	No. of data	Sum(|Res|)	Error
NADR	Jan	−21	17	1042	4216	4
	Feb	−16	17	3021	7699	3
	Mar	−37	16	196	2007	10
	Apr	−29	37	378	4549	12
	May	−54	44	12 493	113 319	9
	Jun	−44	91	17 079	265 859	16
	Jul	−31	46	7125	59 784	8
	Aug	−32	71	9646	71 109	7
	Sep	−32	51	16 467	97 226	6
	Oct	−52	35	8936	71 003	8
	Nov	−52	17	9602	30 976	3
	Dec	−30	23	3324	18 412	6
SARC	May	−20	51	1394	4537	3
	Jun	−95	38	442	6228	14
	Jul	−10	11	2569	5908	2
	Oct	−52	18	1732	6546	4
	Nov	−11	16	184	874	5
ARCT	Feb	−24	61	1991	14 595	7
	Mar	−24	17	1387	4060	3
	May	−20	17	2453	13 958	6
	Jul	−128	72	17 756	205 096	12
	Aug	−70	40	7590	120 249	16
	Sep	−74	16	3562	20 439	6
	Oct	−56	27	6855	27 896	4
	Nov	−35	12	781	2987	4

this region, and to assign pCO_2 values to combinations of these inputs.

A self-organizing network is usually composed of a set of neurons or nodes which is organized in a one-layer two-dimensional lattice. The algorithm organizes the nodes in the lattice into local neighbourhoods that act as feature classifiers on the input data. The input data are compared with the vectors stored at each node, and when there is some similarity to the input this similarity is increased at this node and in its neighbourhood. After several cycles, the random set of nodes becomes organized into a feature or topographic map. The algorithm and its equations are briefly described below.

The inputs are all connected to the nodes. The connections between the nodes are associated with weights, where $w_{ij}(t)$ is the weight from input i to node or neuron j at time t. They are initialized from the n inputs to the nodes to small random values, and iteratively adjusted during the training phase of the network. The inputs are normalized to have values between −1 and 1 to avoid unit problems as different variables are considered. For the month we use cosine and sine functions. For example, January is represented by $\cos(1 \times 2\pi/12)$ and $\sin(1 \times 2\pi/12)$. This reinforces the importance of the seasonality (two inputs to represent one month) and also allows January to be close to both February and December. During the training phase, each data point is applied to the network, and the Euclidean distance d_j between the input i and each output node j is calculated as follows:

$$d_j = \sum_{i=0}^{n-1} (x_i(t) - w_{ij}(t))^2 \tag{2}$$

where x_i is the input pattern (temperature, position and month). A "winning neuron" is defined as the one with the shortest Euclidian distance between its weight and the input pattern. The weight of this neuron and those in its neighbourhood are then updated according to the following equation:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)). \tag{3}$$

The gain $\eta(t)$, between 0 and 1, decreases with time to slow the weight adaptation. The size of the neighbourhood decreases with time too to increase the resolution of the feature map.

After several passes of the data through the network, we obtain a classification of the input patterns. This algorithm was written in Fortran 90 and is described in the flow chart of Fig. 4.

2.3. *Assessment of the performance of the neural network*

In order to compare with the regression approach we used month, latitude, longitude and *in situ* temperature as input patterns to the network, made of 70 × 70 neurons. After training with these data, the network classified the input patterns that are then associated with the corresponding pCO_2 values. To measure the performance of the network, we ran it with all the data we used to train it, and calculated the residual standard deviation (RSD) between the data and the network output:

$$\mathrm{RSD} = \sqrt{\frac{\sum_i (Y_i - X_i)^2}{N-2}}$$

where N is the number of data, Y is the network pCO_2 and X is the associated pCO_2 observation. The NADR province is the most sampled region of the subpolar gyre and the network was trained both in the NADR and in the subpolar gyre (50–70°N, 60–10°W) where the data distribution is not so good (Fig. 5). The monthly RSD values for each month (Table 3) were used as an estimate of the network error. The errors were usually larger in the subpolar gyre than in NADR but they still remained acceptable, so maps could be generated for the subpolar gyre. The monthly means were exactly reproduced by the network (Table 3).

Using this trained network, the NCEP SST dataset was presented to the network as new input patterns to provide monthly maps of pCO_2 in the subpolar gyre.

1. Initialise the network

Let $w_{ij}(t)$ $0 \le i \le n-1$ be the weight from input i to neuron j at time t (n inputs)

Initialise the weights to small random values between 0 and 1

Set the initial radius of the neighbourhood around neuron j, $N_j(0)$ to be large

Set the initial gain term $\eta(0)$, $\eta(t)$ is between 0 and 1 and decreases with time

2. Present input to the network

At time step t, present input $x_0(t)$, $x_1(t)$,…, $x_{n-1}(t)$ where $x_i(t)$ is the input at neuron i

3. Calculate distances

Compute the distance dj between the input and each output neuron j:

$$d_j = \sum_{i=0}^{n-1} (x_i(t) - w_{ij}(t))^2$$

4. Select the minimum distance

Find the best-matching neuron ("winning neuron") j* at time step t by using the minimum distance criterion

5. Update weights, neighbourhood size and gain term

Update the weights for the winning neuron and its neighbours, defined by the neighbourhood size $N_{j^*}(t)$, using the formula:

$$\forall\ j \in N_{j^*}(t) \quad w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)) \quad 0 \le i \le n-1$$

The gain term $\eta(t)$ decreases in time, so slowing the weight adaptation.

The neighbourhood $N_{j^*}(t)$ decreases in size with time to localise the area of maximum activity

6. Repeat until convergence

Fig 4. Flow chart of the algorithm of the neural network.

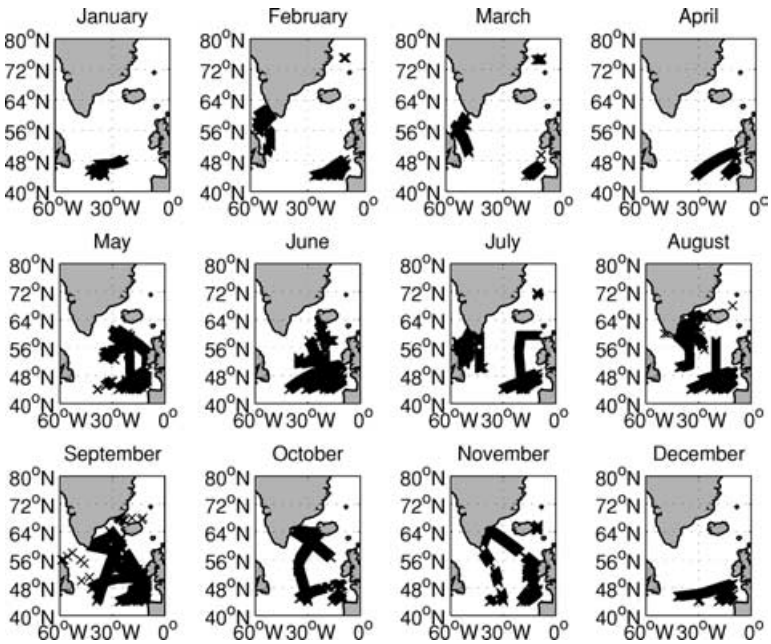


Fig 5. Monthly data distribution of pCO_2 in the subpolar gyre.

Table 3. Comparison of the network pCO_2 output with the data. The network was trained with 100% of the data used for determining the algorithms in the NADR province (44–58°N, 42–10°W, 1.86×10^{12} m^2). The residual standard deviation (RSD) is used as a measure of the performance of the network. The monthly means of pCO_2 (in μatm) calculated from the observations (Obs. mean) and the neural network (NNet mean) are also compared. *N* is the number of observations. The neural network was also trained with all the subpolar region data and RSD and *N* are given for the subpolar gyre (50–70°N, 60–10,°W, 6.15×10^{12} m^2) in bold.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
RSD NADR	6.5	4.3	6.3	14	6.4	7.1	5.4	4.9	4.8	4.2	2.6	5.6
Obs. mean	336.9	355.3	352.9	329.6	315.9	307.7	320.0	327.1	311.0	331.6	337.7	345.3
NNet mean	336.9	355.3	352.9	329.6	315.9	307.7	320.0	327.1	311.0	331.6	337.7	345.3
N	1042	3021	196	378	12 493	17 079	7125	9646	16 467	8936	9602	3324
RSD subp.	**6.9**	**9.8**	**4.8**	**15.8**	**8**	**9**	**8.9**	**6.9**	**5.3**	**3.9**	**3.7**	**7**
N subpol.	**1042**	**5012**	**1583**	**378**	**16 340**	**17 521**	**27 450**	**17 236**	**20 029**	**17 523**	**10 567**	**3324**

Table 4. Residual standard deviation (RSD) of the network output and the regression equations. NNET is the RSD of the network, REG is the RSD of the regressions and *N* is the number of observations. A subset of the data was used to train the network and to determine the regressions. The predicted pCO_2 form both techniques is compared with the remaining dataset

	Jan	Feb	Mar	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
NNET	6.3	10.6	4.3	7.5	9.9	9.0	6.2	5.3	4.5	2.9	6.4
REG	5.4	88.1	4.1	53	19.4	22.0	48.9	27.3	19.2	23.1	7.0
N	521	2505	693	8053	8539	13 724	8604	9986	8761	4796	1662

2.4. *Prediction of pCO_2 using regressions and the neural network*

In order to assess which technique would predict pCO_2 more accurately, we use the same subset of data to determine multiple regressions and to train the neural network. We then compared the pCO_2 predicted by both techniques for the remaining data. The working data set was chosen by randomly sampling the total data set to divide the data into two roughly equal sets: one set of 69 611 data values was used while the remaining 66 182 data values were the values for prediction. The coefficients of the regressions were not very different from the ones determined on the total dataset. Using the new regression equations, pCO_2 was calculated at the temperature, year and location of the remaining data. The temperature, month and location of these data were presented to the neural network to determine pCO_2. The RSD was calculated for both techniques to estimate how well they predicted the pCO_2 of the remaining data (Table 4). This test was not done in April because there were too few data to make a subset. The neural network predicts the remaining pCO_2 data with a better accuracy than the regressions. Except in January, March and December, the RSD given by the regressions is significantly higher. The regressions are much more dependent on the data than the neural network.

3. Monthly CO_2 climatology

Monthly pCO_2 maps based on linear regressions and using the neural network for the years 1995 to 1997 in the subpolar gyre are shown in Figs 6 and 7. Both methods reproduce the broad features seen in the data. The highest pCO_2 are observed in winter and autumn, and are associated with the lowest temperatures. The vertical mixing breaks the summer stratification down from October, and brings more CO_2 to the surface explaining the abrupt change of the pCO_2 distribution between September and October. The highest pCO_2 are in February and March. In April, at the onset of stratification and biological activity, the surface pCO_2 starts decreasing. Although the water warms up during that time to reach a maximum temperature in August, the CO_2 drawdown due to biological activity dominates the pCO_2 distribution and the lowest pCO_2 values are usually encountered during the summer in this region (Peng et al., 1987; Takahashi et al., 1993).

One problem of the multiple linear regression approach is that the ocean must be divided into subregions and no attempt has been made to remove discontinuities between them, therefore the transition between the different provinces is not smooth, i.e. in January and February the pCO_2 changes quite abruptly at 24°W between SARC and ARCT. By contrast, when the neural network was run in the subpolar gyre all the data were presented together regardless of the province. The maps obtained by the neural network approach show more small-scale features than the maps based on multiple regression relationships. For example the pCO_2 distribution in June is patchier with a small area of pCO_2 of less than 280 μatm. Also, it tends to produce southeast to northwest bands of uniform pCO_2 (particularly pronounced in March and October) which corresponds to the temperature patterns. However, the pCO_2 distribution is more complex than

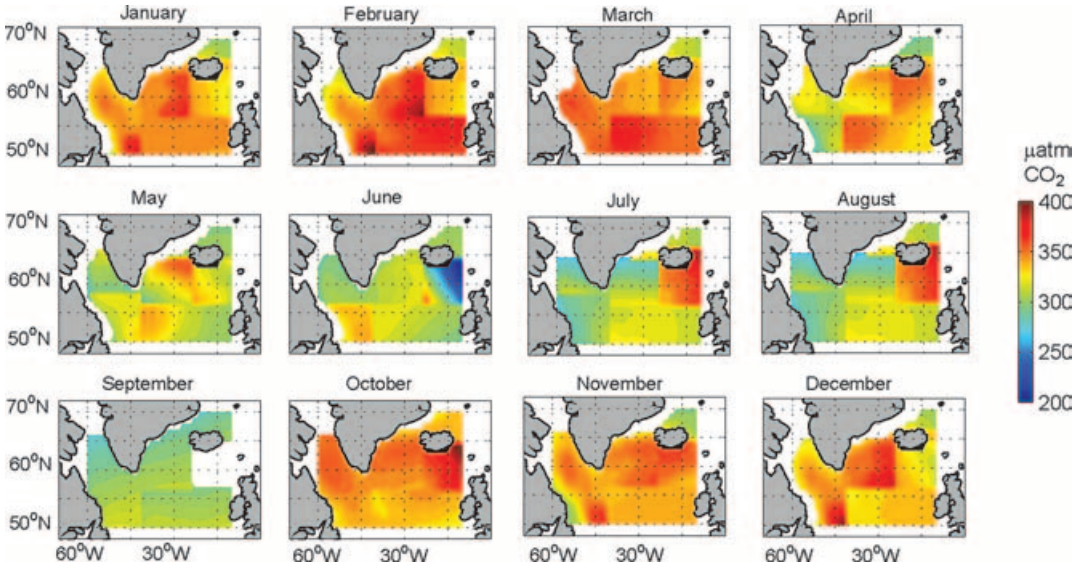


Fig 6. Monthly surface pCO_2 maps for the years 1995 to 1997 in the subpolar gyre obtained using empirical relationships.

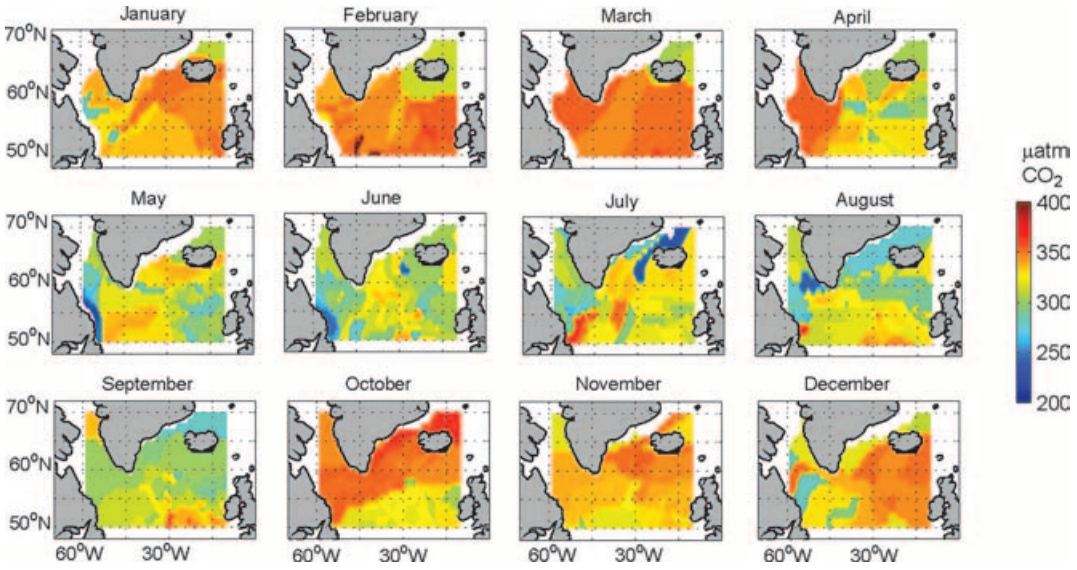


Fig 7. Monthly surface pCO_2 maps for the years 1995 to 1997 in the subpolar gyre obtained using the neural network.

the temperature field, as similar temperatures can be associated with very different pCO_2 values.

In a few regions the neural net and linear regression approaches differ substantially. For example, in July the *James Clark Ross* 1996 cruise data strongly affected the pCO_2 distribution north of Iceland when used in the neural network. This cruise was not used in the regression approach because of the paucity of data in this region, so the pCO_2 distribution was obtained by interpolation between March and September north of 66°N (ARCT3). In this region the temperature fields are similar in July, August and September so the impact of the *James Clark Ross* 1996 data also affects the neural network maps for these months. In June the temperature field is different, so the neural network maps this region with higher pCO_2 values similar to those in February and March where temperatures are closer to the June values. 1996 was an anomalous year in the Greenland and Icelandic seas, with extensive summer-time sea ice originating from the Arctic. The current distribution includes very few data north of 66°N (Fig. 5).

Another region of high discrepancy between the two approaches is found in the SARC province, south of Iceland, in August. Again, there are not enough data available for determining a regression so the pCO_2 distribution is an interpolation between July and October where pCO_2 is around 340 μatm. However, there is a single, restricted set of data in the region in August, which shows pCO_2 values down to 300 μatm and

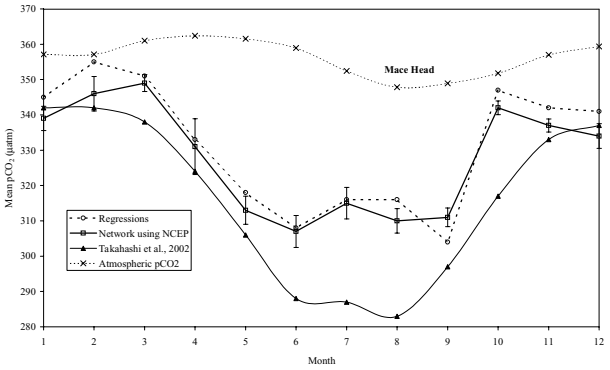


Fig 8. Mean surface pCO_2 in the subpolar gyre plotted as a function of month. Open circles are estimates from the empirical relationships maps, open squares are estimates from the neural network, black triangles are the estimates of Takahashi et al. (2002).

which can be included in the neural network approach. Accordingly, the neural network shows a large undersaturation, although the temperature reaches its maximum in August.

4. Mean seasonal cycle of surface pCO_2 and CO_2 flux in the subpolar gyre

Figure 8 shows the monthly mean pCO_2 calculated in the subpolar gyre (50–70°N, 60–10°W) from the maps based on algorithms and neural networks compared with the climatology of Takahashi et al. (2002). The error bars on the neural network estimates are the RSD given in Table 3 for the subpolar gyre.

Despite the spatial differences between the regression and the neural network based maps, the monthly means agree well. The regressions give higher pCO_2 values than the other approaches but the subpolar gyre remains below the atmospheric level throughout the year. Winter is the least-sampled season, and the linear relationships are based on few data that might not be representative of the region. Thus, if some heterogeneity in the pCO_2 distribution occurs during these months it is not captured by the empirical relationships, whereas the network will infer it from the temperature distribution. This might explain the highest discrepancy observed at this season between the two techniques. As an annual average, the regressions give a value of 331 μatm whereas the neural network gives 328 μatm, both higher than the climatology of Takahashi et al. (2002) (316 μatm). The main differences between our estimates and those of Takahashi et al.'s climatology are in summer when both the neural network and the regressions give significantly higher pCO_2 with a difference >10 μatm from the monthly mean. The surface pCO_2 has increased significantly in this season in recent decades (Lefèvre et al., 2004), perhaps because of the decrease in primary production at high latitudes (Gregg et al., 2003) so the discrepancy between the regressions and the climatology of Takahashi et al. could be due to the use of more recent measurements. Takahashi et al. assume no change in surface pCO_2 which would affect the pCO_2 distribution more strongly in summer as the increase of pCO_2 is >2 μatm yr^{-1} for this season (Lefèvre et al., 2004).

As the methods applied to map pCO_2 are different, it is possible that other factors could contribute to our weaker sink than solely the year correction. However, not correcting the data for the year of measurement, and hence assuming no increase of sea water pCO_2 over time, will lead to an overestimate of the carbon uptake in the Atlantic subpolar gyre. The neural network technique appears promising. Furthermore, it should be applicable to other types of input data as well, notably ocean colour, which we would expect to have some predictive power in relation to surface pCO_2, but will not be linearly related to it. However, validating the output of a neural network is a difficult task which we have only just begun here, and which requires independent data with wide spatial and temporal coverage. The neural network technique can, in principle, be used to interpolate from very scattered data, extracting the maximum information from a sparse database. This same property may be a disadvantage, however, if some of the input data are in fact unrepresentative of normal conditions.

The increase in sea water pCO_2 means has implications for the carbon uptake in this region. The higher pCO_2 values observed with the regressions and the neural network suggest that the carbon uptake is lower than previous estimates have implied. However, the winds are stronger in winter and, since the main differences in pCO_2 are observed in summer, this might not lead to a significantly different CO_2 uptake for the region. Using the surface pCO_2 maps we calculated the CO_2 flux as follows. Atmospheric xCO_2 (molar fraction of CO_2) measured at the Mace Head Station (53°N, 9°W) were downloaded from the World Data Centre for Greenhouse Gases (http://gaw.kishou.go.jp/wdcgg). CO_2 fluxes were calculated using the gas exchange coefficient of Wanninkhof (1992) and the average monthly wind speed between 1995 and 1997 calculated from the NCEP monthly surface wind speed converted to a 10 m height. The neural network gives an annual oceanic uptake of CO_2 of 5.68 mmol m^{-2} d^{-1}, which corresponds to 0.15 Gt-C yr^{-1} for the region 50–70°N, 60–10°W. The maps based on the linear relationships give a smaller uptake of 4.72 mmol m^{-2} d^{-1} (0.13 Gt-C yr^{-1}). We obtain similar results when using wind speed for 1995 only so the average between 1995 and 1997 is not significantly different from the wind field in 1995 for this region.

In order to compare with the estimates of Takahashi et al. (2002) we also computed the CO_2 flux using their 41-yr wind field at 10 m height and calculated the average for the three techniques in the area 50–70°N, 60–10°W. Although the 41-yr wind speed is stronger than the mean wind speed averaged between 1995 and 1997, this does not lead to a significantly stronger uptake with 0.14 versus 0.13 Gt-C yr^{-1} for the regressions and 0.17 versus 0.15 Gt-C yr^{-1} for the neural network. As our annual means, compared for the same region with the same wind field,

are significantly lower than the Takahashi et al. (2002) average of 0.22 Gt-C yr^{-1}, this means that the discrepancy between the flux estimates is caused by the difference in the pCO_2 distributions.

5. Conclusions

We have constructed a sea water pCO_2 climatology for the subpolar gyre (50–70°N, 60–10°W) for the years 1995 to 1997 by developing multiple linear regression relationships based on over 108 000 measurements. A self-organizing neural network approach has also been investigated. Most of the data were collected between 1995 and 1997. Only 35%, 11% and 3% of the data were collected in the NADR, SARC and ARCT regions respectively, at other periods. The subpolar gyre is a sink of CO_2 in agreement with previous estimates. However, the mean surface pCO_2 is higher than in the climatology of Takahashi et al. (2002) implying that the CO_2 sink is actually weaker. The neural network and the regression relationships provide higher pCO_2 in summer, which is consistent with the recent observation of a decadal decrease of primary productivity at high latitudes. Both approaches, regressions and neural network, suggest a weaker sink of CO_2 in the subpolar gyre compared with the climatology of Takahashi et al. (2002). A correction for the increase of sea water pCO_2 with year should be applied to estimate the oceanic sink of CO_2 in the Atlantic subpolar gyre. The encouraging results of the neural network approach open the prospect of using temperature and ocean colour from satellite data to generate maps of pCO_2. However, although it is a good tool for generalizing a limited set of pCO_2 observations, the technique will still require careful validation using comprehensive *in situ* pCO_2 data.

6. Acknowledgments

This work has been funded by the European Commission under the programme Environment and Sustainable Development, contracts EVK2-CT-2000-00088 (CAVASSOO) and EVK2-CT-2001-00115 (NOCES). It was additionally funded by the UK NERC's CASIX project. The atmospheric measurements are from T. J. Conway and P. P. Tans [Climate Monitoring and Diagnostics Laboratory, National Oceanic and Atmospheric Administration (http://www.cmdl.noaa.gov/ccgg/index.html)].

References

Gregg, W. W., Conkright, M. E., Ginoux, P., O'Reilly, J. E. and Casey, N. W. 2003. Ocean primary production and climate: global decadal changes. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL016889.

Kohonen, T. 1984. *Self-organization and Associative Memory*. Springer-Verlag, Berlin.

Lefèvre, N., Watson, A. J., Olsen, A., Rios, A. F., Perez, F. F. and co-author 2004. A decrease in the sink for atmospheric CO_2 in the North Atlantic. *Geophys. Res. Lett.* **31**(L07306), doi:10.1029/2003GL018957.

Lochte, K., Ducklow, H. W., Fasham, M. J. R. and Stienen, C. 1993. Plankton succession and carbon cycling at 47°N 20°W during the JGOFS North Atlantic Bloom Experiment. *Deep Sea Res.* **40**, 91–114.

Longhurst, A., Sathyendranath, S., Platt, T. and Caverhill, C. 1995. An estimate of global primary production in the ocean from satellite radiometer data. *J. Plankton Res.* **17**(6), 1245–1271.

Peng, T. H., Takahashi, T. and Broecker, W. S. 1987. Seasonal variability of carbon dioxide, nutrients and oxygen in the northern North Atlantic surface water: observations and a model. *Tellus* **39B**, 439–458.

Schneider, B., Kremling, K. and Duinker, J. C. 1992. CO_2 partial pressure in Northeast Atlantic and adjacent shelf waters: processes and seasonal variability. *J. Marine Syst.* **3**, 453–463.

Takahashi, T., Feely, R. A., Weiss, R. F., Wanninkhof, R. H., Chipman, D. W. and co-authors 1997. Global air-sea flux of CO_2: an estimate based on measurements of sea-air pCO_2 difference. *Proc. Natl. Acad. Sci. USA* **94**, 8292–8299.

Takahashi, T., Olafsson, J., Goddard, J. G. and Chipman, D. W. 1993. Seasonal variation of CO_2 and nutrients in the high-latitude surface oceans: a comparative study. *Global Biogeochem. Cycles* **7**(4), 843–878.

Takahashi, T., Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N. and co-authors 2002. Global sea–air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects. *Deep Sea Res.* **49**(9–10), 1601–1622.

Wanninkhof, R. H. 1992. Relationship between wind speed and gas exchange over the ocean. *J. Geophys. Res.* **97**(C5), 7373–7382.

Watson, A. J., Robinson, C., Robertson, J. E., Williams, P. J. B. and Fasham, M. J. R. 1991. Spatial variability in the sink for atmospheric carbon dioxide in the North Atlantic. *Nature* **350**, 50–53.