Biogeosciences, 10, 4319–4340, 2013
www.biogeosciences.net/10/4319/2013/
doi:10.5194/bg-10-4319-2013

Biogeosciences Open Access

A novel method for diagnosing seasonal to inter-annual surface ocean carbon dynamics from bottle data using neural networks

T. P. Sasse[1], **B. I. McNeil**[1], **and G. Abramowitz**[1,2]

[1]Climate Change Research Centre, Faculty of Science, University of New South Wales, Sydney, Australia
[2]ARC Centre of Excellence for Climate Systems Science and Climate Change Research Centre, UNSW, Sydney, Australia

Correspondence to: T. P. Sasse (t.sasse@unsw.edu.au)

Received: 8 October 2012 – Published in Biogeosciences Discuss.: 1 November 2012
Revised: 8 April 2013 – Accepted: 26 May 2013 – Published: 27 June 2013

Abstract. The ocean's role in modulating the observed 1–7 Pg C yr^{-1} inter-annual variability in atmospheric CO_2 growth rate is an important, but poorly constrained process due to current spatio-temporal limitations in ocean carbon measurements. Here, we investigate and develop a non-linear empirical approach to predict inorganic CO_2 concentrations (total carbon dioxide (C_T) and total alkalinity (A_T)) in the global ocean mixed layer from hydrographic properties (temperature, salinity, dissolved oxygen and nutrients). The benefit of this approach is that once the empirical relationship is established, it can be applied to hydrographic datasets that have better spatio-temporal coverage, and therefore provide an additional constraint to diagnose ocean carbon dynamics globally. Previous empirical approaches have employed multiple linear regressions (MLR) and relied on ad hoc geographic and temporal partitioning of carbon data to constrain complex global carbon dynamics in the mixed layer. Synthesizing a new global C_T/A_T carbon bottle dataset consisting of ~33 000 measurements in the open ocean mixed layer, we develop a neural network based approach to better constrain the non-linear carbon system. The approach classifies features in the global biogeochemical dataset based on their similarity and homogeneity in a self-organizing map (SOM; Kohonen, 1988). After the initial SOM analysis, which includes geographic constraints, we apply a local linear optimizer to the neural network, which considerably enhances the predictive skill of the new approach. We call this new approach SOMLO, or self-organizing multiple linear output. Using independent bottle carbon data, we compare a traditional MLR analysis to our SOMLO approach to capture the spatial C_T and A_T distributions. We find the SOMLO approach improves predictive skill globally by 19 % for C_T, with a global capacity to predict C_T to within 10.9 µmol kg^{-1} (9.2 µmol kg^{-1} for A_T). The non-linear SOMLO approach is particularly powerful in complex but important regions like the Southern Ocean, North Atlantic and equatorial Pacific, where residual standard errors were reduced between 25 and 40 % over traditional linear methods. We further test the SOMLO technique using the Bermuda Atlantic time series (BATS) and Hawaiian ocean time series (HOT) datasets, where hydrographic data was capable of explaining 90 % of the seasonal cycle and inter-annual variability at those multi-decadal time-series stations.

1 Introduction

The ocean's role in modulating rising atmospheric carbon dioxide (CO_2) levels has been found to be very important (Khatiwala et al., 2012; Sabine et al., 2004). A variety of data-based estimates suggest net oceanic uptake for CO_2 to be 2.1 ± 1.0 Pg C yr^{-1} (1 Pg = 10^{15} g) since the year 2000, or about 25–30 % of anthropogenic CO_2 emissions over that period (Jacobson et al., 2007; Khatiwala et al., 2009; Manning and Keeling, 2006; McNeil et al., 2003; Mikaloff-Fletcher et al., 2006; Takahashi et al., 2009). Between 1990 and 2009, atmospheric CO_2 accumulation rates vary between 1 and 7 Pg C yr^{-1}, indicating large inter-annual variability from both the terrestrial and oceanic reservoirs (Sarmiento et al., 2010). Although our long-term, decadal-scale understanding of oceanic CO_2 uptake has advanced, our shorter-term understanding (seasonal to inter-annual) of ocean carbon dynamics remains poorly constrained due to current data limitations.

Atmospheric CO_2 observations, inversion techniques and ocean models suggest a large range for inter-annual variability in oceanic CO_2 uptake (0.1–1.5 Pg C yr^{-1}) (Bender et al., 2005; Le Quéré et al., 2003; Patra et al., 2006; Rayner et al., 2008). However, from an oceanic perspective, our understanding of natural variability of ocean carbon has come about sporadically, dominated by regional time-series measurement programs (e.g. Bermuda Atlantic time series (BATS) and Hawaiian ocean time series (HOT)). Without a better understanding of shorter-scale natural variability, the ability to constrain and understand the time-evolving capacity for the ocean to absorb atmospheric CO_2 in a high-CO_2 world will be limited, particularly since some evidence suggests the ability for the ocean to absorb CO_2 has slowed since the late 1980s as a consequence of decadal-scale trends in winds and oceanic circulation (Le Quéré et al., 2010; Sarmiento et al., 2010).

Standard hydrographic measurements in the ocean (temperature, salinity, dissolved oxygen and nutrients) are sampled and analysed much more frequently than inorganic carbon. With the deployment of satellites, gliders and ARGO floats providing an immense capacity for capturing short-term seasonal to inter-annual variability in the oceans, the question is, can this new information be used to help infer and diagnose short-term carbon dynamics in the ocean?

The oceans inorganic carbon system can be fully constrained by knowing any two measurements within its inorganic carbon constituents; partial pressure of CO_2 (pCO_2), total dissolved carbon dioxide (C_T), total alkalinity (A_T) or pH. Significant time and resources have been devoted on national and international levels to survey the global oceanic C_T and A_T distribution. However, even with approximately 330 000 bottle measurements taken sporadically over the past 30 years, our ability to globally understand natural seasonal C_T and A_T dynamics has been hindered due to the large spatio-temporal limitations in this current accumulated dataset (Key et al., 2004).

Autonomous pCO_2 measuring devices mounted mainly onto commercial shipping vessels has resulted in a global network of approximately 6.4 million ocean surface pCO_2 measurements (Takahashi et al., 2012). This pCO_2 dataset has given us the best idea of seasonal (Takahashi et al., 2009; herein after referred to as T-09) to inter-annual (McKinley et al., 2011; Park et al., 2010; Telszewski et al., 2009) CO_2 variability within the ocean. However, the global pCO_2 dataset cannot inform us on some very important processes and biogeochemical dynamics that modulate atmospheric CO_2. The ocean's biological carbon export flux has been estimated to be between 11 and 16 Pg C yr^{-1} from satellite chlorophyll *a* measurements (Falkowski et al., 2000), some 5–8 times the net oceanic CO_2 absorption from the atmosphere. Small changes in the biological carbon flux have large and important implications for atmospheric CO_2. However, this large signal is yet to be constrained from inorganic carbon data itself, since it requires constraints on mixed-layer carbon dynamics rather than just sea-surface constraints like the pCO_2 climatology. Secondly, without equivalent A_T or C_T measurements, pCO_2 by itself cannot provide insights into partitioning the biological carbon pump into both organic and calcification components, particularly important with regard to future ocean acidification impacts. Previous estimates on this "rain ratio" (organic/calcifier export flux) have needed to assume a constant redfield ratio on nutrient changes in the oceans mixed layer (Sarmiento et al., 2002). Finally, spatio-temporal deficiencies in the pCO_2 dataset in regions like the Southern Ocean introduce uncertainties in the direct evaluation of short-term variability. To understand seasonal to inter-annual variability in these regions requires methods that have better spatio-temporal coverage than is constrained by historical pCO_2 sampling. Here, we seek to diagnose seasonal to inter-annual C_T and A_T concentrations in the mixed layer that provide independent, but important additional constraints to the global sea-surface pCO_2 climatology.

To varying degrees, concentrations of C_T and A_T are influenced by the solubility of CO_2, biological processes, vertical and lateral water transport and direct CO_2 exchange with the atmosphere (Sarmiento and Gruber, 2006). Ocean mixing is largely controlled by density dynamics via temperature (T) and salinity (S) variations in the ocean, which also regulate the solubility of CO_2 (Weiss, 1974). Information on nitrate (N), silicate (Si), phosphate (P) and dissolved oxygen (DO) variations provide insight into the biological influences on oceanic inorganic carbon (Anderson and Sarmiento, 1994). From this, it should be implicit that we can derive empirical relationships between these standard hydrographical parameters and the carbon constituents. If a robust empirical relationship is established, we could apply our model to the order of magnitude more in situ measurements of these standard hydrographic parameters (Boyer et al., 2009) or the objectively analysed $1^\circ \times 1^\circ$ climatologies (e.g. Locarnini et al., 2010) to give us new constraints on seasonal to inter-annual carbon dynamics in the mixed layer.

The use of the global sea-surface pCO_2 dataset would be ideal to develop such empirical algorithms. However, these continuous pCO_2 measurements generally have no coinciding biogeochemical information (i.e. DO or nutrients) that could help establish an empirical relationship. Some have used satellite chlorophyll *a* measurements to help constrain ocean surface pCO_2 with varying degrees of success (Chen et al., 2011; Chierici et al., 2009; Telszewski et al., 2009). The benefits of using ship-based bottle measurements of C_T and A_T, is that they are almost always complemented by a suite of hydrographic and biogeochemical parameters (T, S, DO and nutrients) that can be used to help derive empirical relationships.

Wallace (1995) verified a multiple linear regression (MLR) concept by successfully capturing C_T using T, S, Si and apparent oxygen utilization (AOU) in the North Atlantic. Several studies have since investigated this MLR approach in capturing the surface distribution of C_T and A_T (see Table 1).

Table 1. Previous empirical approaches to constrain surface A_T and C_T distributions. T is temperature, S is salinity, DO is dissolved oxygen, AOU is apparent oxygen utilization, N is nitrate (NO_3^-), Si is silicate (SiO_4), P is phosphate (PO_4^{3-}), Chl a is chlorophyll a, Lat is latitude, and Long is longitude.

Study Region	Response	Predictors	N[a]	RSE[b] ($\mu mol\,kg^{-1}$)	Author
Global	NA_T[c]	T	1740	5	Millero et al. (1998)
Global	A_T	T, T^2, S, S^2, Long	5692	8.1	Lee et al. (2006)
Indian Ocean	A_T	T, S, N, AOU, Depth, Lat, P	2363	4.5–6.4[d]	Bates et al. (2006)
Southern Ocean	A_T	S, N, Si	1200	8.1	McNeil et al. (2007)
Arctic Ocean	A_T	T, S	853	26.9, 75	Arrigo et al. (2010)
Global	NC_T[c]	T, T^2, N	$\sim$4900	7	Lee et al. (2000)
Indian Ocean	C_T	T, S, N, AOU, Depth, Lat, P	2395	4.4–6.0[d]	Bates et al. (2006)
Southern Ocean	C_T	T, S, DO, N, Si	1032	8	McNeil et al. (2007)
Arctic Ocean	C_T	Chl a, T, S	853	33.4, 61.6, 17.3	Arrigo et al. (2010)

[a] Number of measurements used in the study.
[b] Residual standard error as quoted by the authors.
[c] Salinity normalized concentrations of C_T and A_T. $\left(\times \frac{35}{S}\right)$.
[d] Range of RSE values presented for the four monsoonal/inter-monsoonal seasons.

Divergent biological and mixing regimes throughout the ocean have made it difficult to use linear empirical techniques on a global scale. Researchers have traditionally partitioned the global bottle dataset geographically, hydrographically and temporally in an attempt to improve the ability of linear approaches to model the non-linear relationship between inorganic carbon and the standard hydrographic parameters. Here we use a non-linear empirical modelling approach to avoid this ad hoc partitioning and show that it delivers considerable improvements in predictability. We use a self-organizing map (SOM; Kohonen, 1988) to classify or cluster measurements of hydrographic parameters into groups and then establish the relationship between these parameters and C_T/A_T separately for each group. SOMs have already been found to be well suited in extracting features of the ocean surface pCO_2 dataset in the North Atlantic using a combination of modelled and remotely sensed parameters to constrain the system, (Friedrich and Oschlies, 2009a, b; Lefèvre et al., 2005; Telszewski et al., 2009).

To contextualize this work, we firstly explore the use of the traditional MLR approach to diagnose global seasonal carbon dynamics in the ocean. To do this, we employ the MLR approach on a newly synthesized C_T/A_T bottle dataset of $\sim$33 000 mixed-layer samples. Next, we present our SOM-based approach to diagnose seasonal carbon dynamics on a global scale, which better accounts for non-linearities that would limit the ability of the MLR approach. To compare the MLR and our SOM approach, we develop an independent test approach to assess the model's skill. We then use the BATS and HOT in situ time series as an explicit test for our new approach and finally show the capacity of the model to capture coherent, spatial and temporal carbon fields over the global ocean.

2 Global carbon measurements and training dataset

The extraordinary effort to collate and synthesize the bottle hydrographic and biogeochemical data has been conducted by several groups; including GLODAP (GLobal Ocean Data Analysis Project; Key et al., 2004), CARINA (CARbon dioxide IN the Atlantic Ocean; CARINA Group, 2009a, b, 2010) and PACIFICA (PACIFic Ocean Interior CArbon project; Suzuki et al., 2013).

Precision in measuring bottle C_T and A_T samples has consistently improved over the past 30 yr as a result of advances in techniques and apparatus (Bradshaw et al., 1981; Johnson et al., 1987). However, it was not until the introduction of standard operating procedures and certified reference materials (Department of Energy, 1994; Dickson et al., 2003; Dickson et al., 2007) that the quality consistency of independent laboratory measurements was achieved and is currently estimated to be $\pm 2\,\mu mol\,kg^{-1}$ (Dickson et al., 2007). To account for any systematic measurement biases between independent laboratories when combining data, a secondary quality control (QC) method was incorporated by the project groups to identify and smooth out any inconsistencies, as outlined in Tanhua et al. (2010). The internal consistency of the CARINA C_T/A_T dataset has been estimated to $\pm 2.5\,\mu mol\,kg^{-1}$ (Tanhua et al., 2010). More recent additional measurements we included in the global dataset underwent a 1^{st} QC check to remove measurements that were flagged as bad or questionable under the World Ocean Circulation Experiment (WOCE) convention (Joyce and Corry, 1994).

For this work, 470 cruises from GLODAP, PACIFICA, CARINA, CLIVAR and miscellaneous sources were merged with the BATS and HOT measurements to form the global carbon training dataset, as shown in Table 2. We refined the global data to be within the mixed layer (Supplement A),

Table 2. Data sources of our global merged dataset.

Source	Number of Measurements
CARINA	12 599
PACIFICA	9690
GLODAP	6674
CLIVAR[a]	1689
AAIW[b]	755
BATS[c]	705
HOT[d]	540
NACP[e]	291
Miscellaneous	192
Total	33 135

[a] Climate Variability and Predictability.
[b] Antarctic Intermediate Cruise.
[c] Bermuda Atlantic time series.
[d] Hawaiian ocean time series.
[e] North Atlantic Carbon Program.

non-coastal (Supplement B) and data post-1980 due to large uncertainties in early measuring techniques. The final number of usable C_T/A_T discrete measurements in the global mixed layer was $\sim 33\,000$.

Whilst the spatial coverage of the refined data is consistent over all major ocean basins (Fig. 1a), there are approximately 45 % less wintertime measurements than were collected during summertime (Fig. 1b), which we examine here as a potential cause for bias when applying our approach.

Normalization of C_T measurements

Global atmospheric CO_2 concentrations during the 1980s, 1990s and 2000s have increased at 1.60 ± 0.56, 1.47 ± 0.66 and 1.90 ± 0.38 ppm yr^{-1}, respectively (Thomas Conway and Pieter Tans, NOAA/ESRL, www.esrl.noaa.gov/gmd/ccgg/trends). Mixed-layer measurements of C_T were corrected for temporal anthropogenic CO_2 uptake to the reference year 2000 by calculating the change in mixed-layer C_T in equilibrium with the atmospheric CO_2 increase using observed Revelle factors (see supplement material C for details). This approach is somewhat equivalent to that of T-09 where all pCO_2 measurements values were corrected to the year 2000 using a rate of 1.5 µatm yr^{-1}.

There are regions of the ocean where upwelling and sea–ice inhibit air-sea gas exchange, resulting in considerable CO_2 disequilibrium (e.g. Southern Ocean, equatorial Pacific). The anthropogenic CO_2 correction technique used here, like those for T-09 and Lee et al. (2000), will be biased in these regions. However, by performing a test using no anthropogenic CO_2 correction (Supplement D), we demonstrate the very low impact this anthropogenic correction has to our final result. This is in part due to the large natural fingerprint of C_T (± 50 µmol kg^{-1}) relative to the small changes (~ 1 µmol kg^{-1} yr^{-1}) resulting from anthropogenic CO_2 uptake.

3 Testing algorithm skill: a systematic independent test (SIT) approach

Most empirical studies report statistical errors calculated as the residual standard error (RSE) from linear regressions. For example, C_T in the Indian Ocean was reported to be predicted to within ± 5 µmol kg^{-1} using a suite of hydrographic parameters (Bates et al., 2006), ± 8 µmol kg^{-1} for the Southern Ocean (McNeil et al., 2007) and ± 7 µmol kg^{-1} for a global dataset (Lee et al., 2000). However, an independent dataset not used in the regressions is needed to accurately report true statistical uncertainty for any empirical approach.

Here, we developed a "systematic independent test" (SIT) approach in order to compare the MLR and NN empirical approaches consistently. The SIT method evaluates the algorithm's skill through an independent test of each cruise or time series without using it in the training or regression dataset. This implies that for a training data pool consisting of n cruises and i time series, $n+i$ unique algorithms with identical model configurations are used to predict the excluded cruise or time series measurements. Calculating the residual standard error (RSE; Eq. 1) using all (or a subset) of the cruises and time-series independent predictions then provides a better and accurate estimate of the algorithms global (or regional) skill. In Eq. (1), the independent predictions and in situ measurements are represented by $y_{\mathrm{indp-pred}}$ and $y_{\mathrm{in-situ}}$ respectively, while N defines the number of discrete samples.

$$\mathrm{RSE} = \sqrt{\frac{\sum \left(y_{\mathrm{indp-pred}} - y_{\mathrm{in-situ}}\right)^2}{N-2}} \quad (1)$$

The reason we independently test each cruise dataset individually, rather than a randomly selected subset of the data, is due to similar concentrations of carbon and auxiliary measurements within local casts of the same cruise. As there are typically two to three measurements within each cast of the training dataset, the independent prediction of one of these measurements will give a misleading representation of the model's true skill, as the remaining two measurements with a very similar "biogeochemical fingerprint" will be used to train the algorithm. The prediction of an entire independent cruise is a more robust measure of the algorithm's skill.

4 Traditional MLR approach

4.1 Method description

Multiple linear regression is a numerical estimation of the linear relationship between a set of predictor variables, $\boldsymbol{x} = (x_1, \ldots, x_n, \ldots, x_N)$, and response variable, y, (Eq. 2).

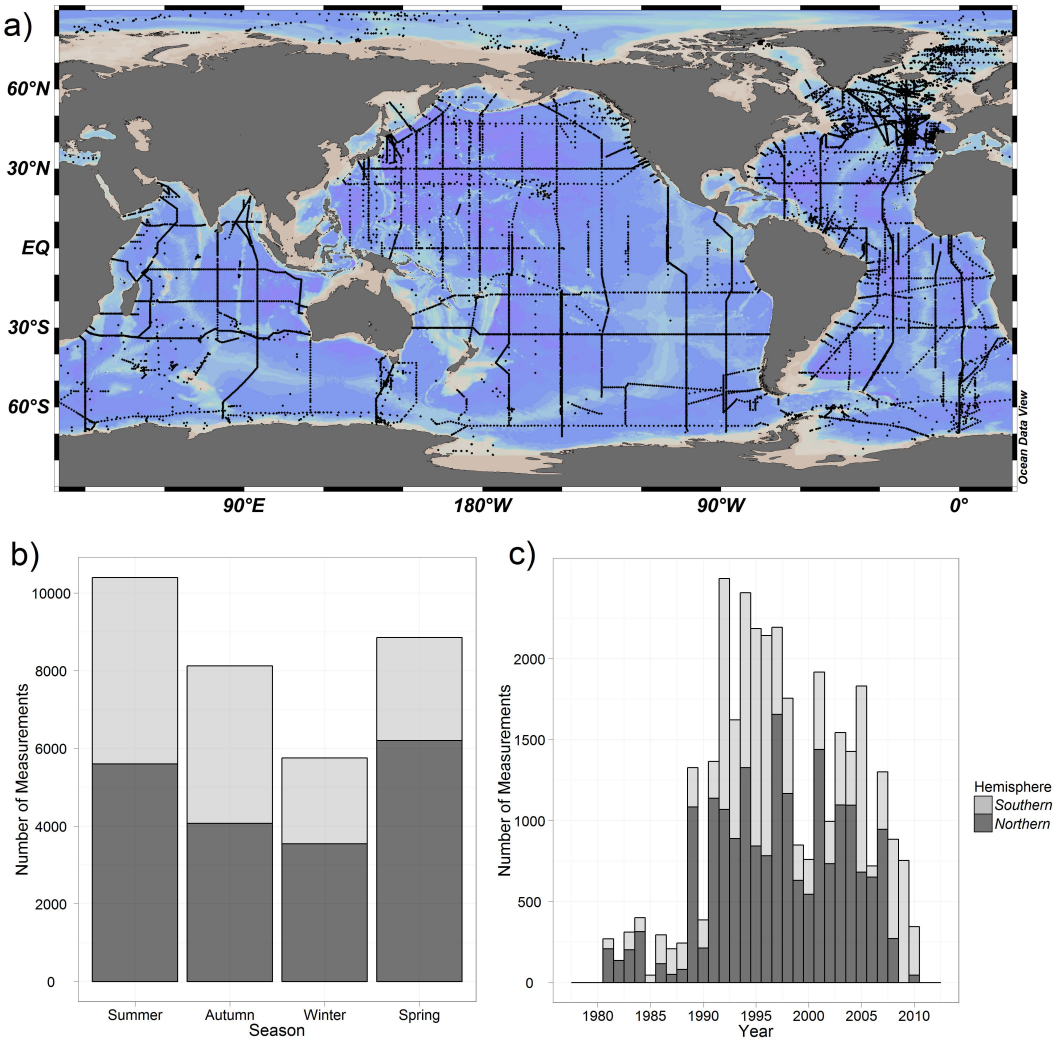

Fig. 1. (a) Global distribution of the training dataset, **(b)** seasonal and **(c)** yearly histograms of the training dataset separated into Southern (light shade) and Northern (dark shade) Hemispheres. Southern Hemisphere seasons are defined as summer (Dec–Feb), autumn (Mar–May), winter (Jun–Aug) and spring (Sept–Nov), Northern Hemisphere seasons are opposite.

$$y = \beta_0 + \sum_{n=1}^{N} \beta_n x_n \ , \tag{2}$$

where β_0 and β_n represent the intercept and empirically derived coefficients respectively. Multi-collinearity (MCL) between predictor variables and non-normality of the residual errors are both issues that may affect the predictive and diagnostic ability of a MLR. To minimize the effect of these issues, the empirical relationships between C_T/A_T and the standard hydrographic parameters were constrained using a forward stepwise robust MLR routine.

Following the schematic in Fig. 2, the routine initiates by ranking predictor variables $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n, \ldots, \boldsymbol{p}_N$ according to their degree of linear correlation to the response variable, $\boldsymbol{y}$; where $\boldsymbol{p}_{n,1}$ represents the parameter with the highest correlation. The primary model ($\boldsymbol{M}_1$) is then established by applying a least-squares MLR between the top ranked predictor variable ($\boldsymbol{p}_{n,1}$) and $\boldsymbol{y}$ to constrain the regression coefficients β_0 and $\beta_{n,1}$. The routine then expands on $\boldsymbol{M}_1$ in step 3 by regressing the top two ranked predictor variables ($m = 2$); where m represents the modelled predictor variable with the lowest correlation to y.

To determine if MCL exists in the expanded model ($\boldsymbol{M}_m$), we calculate the variance inflation factor (VIF) for each modelled variable in $\boldsymbol{M}_m$ and compare them to VIF values calculated for the same variables modelled in $\boldsymbol{M}_{m-1}$. The existence of MCL is identified if the VIF value for any predictor variable $\boldsymbol{p}_{n,i}$ (where $i < m$) increased by 5. For the scenario when MCL is detected, the model is updated with interaction terms between the newly added predictor variable ($\boldsymbol{p}_{n,m}$) and any modelled variable with a VIF increase greater than 5. An

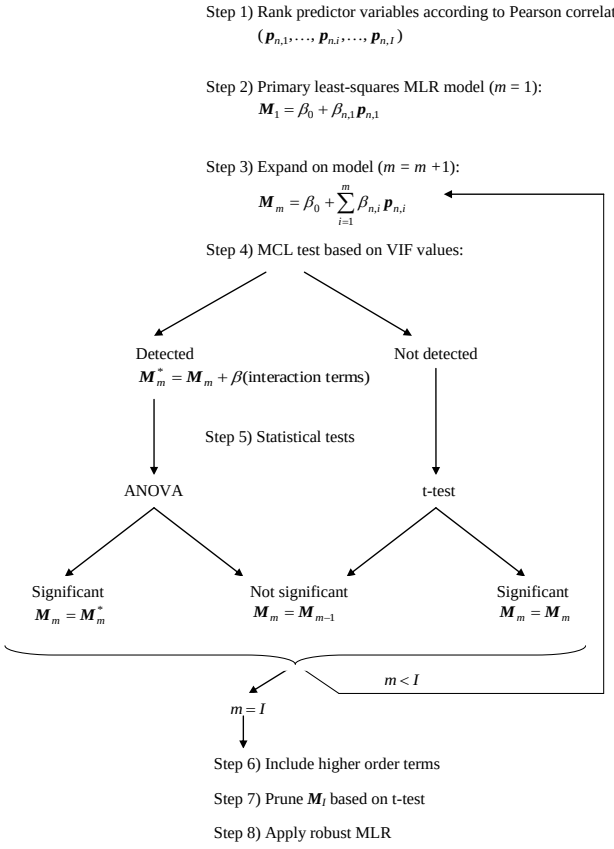

Fig. 2. Schematic diagram of our robust forward MLR routine.

analysis of variance (ANOVA) between the previous model ($\boldsymbol{M}_{m-1}$) and expanded model (M^*_m) is then applied to evaluate the significance of the newly added predictor variable and interaction terms. If the expanded model is found to statistically constrain the system with a higher degree of skill with a 95% confidence interval, the updates are accepted and the routine returns to step 3 to incorporate the next lowest ranked predictor variable (i.e. $m = m + 1$).

If MCL is not detected, a null-hypothesis test based on the t statistic is applied to determine if the coefficient of the new predictor variable is significantly different from 0 (i.e. the new predictor is important in constraining the system). If it does not differ from 0 with a 95 % confidence interval, the new predictor variable is defined as insignificant and is subsequently rejected from the model. The routine then returns to step 3 to again expand $\boldsymbol{M}_m$ with the next lowest ranked predictor variable.

Once each predictor variable has had an opportunity to update the model (i.e. $m = I$), any desired higher order variable terms are incorporated into the model on the provision the first order term was found to be statistically significant. The routine then prunes the model through an iterative process that removes insignificant terms based on the t test. Once all terms are statistically significant, the final stage of the routine applies a robust MLR to the set of significant terms to reduce potential influences from outliers.

This MLR routine is well suited for optimizing the model and dampening the influence of outliers that cannot be reasonably identified as bad measurements. This aspect is particularly important when the global dataset is subject to ad hoc geographical and/or temporal separation methods, where measurements not consistent with the bulk biogeochemical dynamics within a region have the potential to affect the model.

4.2 Ad hoc vs. universal MLR

To investigate the application of the traditional MLR method, we compared the skill of using one single regression globally (universal MLR) to an ad hoc approach that partitions the dataset into regions (ad hoc MLR). We based the ad hoc approach on dividing the global carbon dataset on the geographical and temporal guidelines outlined by Lee et al. (2006, 2000) and Bates et al. (2006). In this way, the global dataset was subset into 5 geographic regions to constrain the A_T system, and 11 geographic regions, 8 of which were subjected to further separation into summer and winter months to constrain C_T (see Fig. 3). The universal method simply uses the entire global dataset without division.

4.3 MLR results

When universally applying the traditional MLR on the $\sim 33\,000$ global mixed-layer C_T measurements, the statistical regression RSE is 15.1 µmol kg^{-1} when using T, S, DO, P, N and Si as predictors (Table 3). If applying the ad hoc geographical and temporal separations, the statistical regression RSE reduces to 13.2 µmol kg^{-1}. However, when the independent test (SIT) is used to evaluate the regressions, errors increase to be 16 µmol kg^{-1} for the ad hoc approach and 15.6 µmol kg^{-1} for the global regression. For A_T, optimal predictors were found to be T, S, S^2, DO, P and Si, while a global MLR algorithm captured the signal to within 11 µmol kg^{-1} using the SIT approach. All empirical relationships for the global and ad hoc MLR models can be found in Supplement Tables T1 and T2.

The MLR approach and results give us a framework to attempt to develop a better method that captures any potential non-linear biases that are contributing to errors of $\pm$16 µmol kg^{-1} in C_T predictions and $\pm$ 11µmol kg^{-1} for A_T on a global scale.

5 Neural network approach

5.1 Overview of the neural network approach

A self-organizing map (SOM) is an algorithm that uses an iterative approach to classify multi-dimensional data into discrete groups, or neurons, usually arranged in a 2-dimensional

Table 3. Universal and ad hoc MLR results for (a) C_T and (b) A_T.

					RSE (μmol kg^{-1})			
					Regression		Independent test (SIT)	
	Region	Zone[a]	N[b]	N cruises[c]	Ad hoc	Universal	Ad hoc	Universal
(a)	Subtropical	1	5388	109	11.9	17.1	15.2	17.3
	Eq. Pacific	2	752	14	11.3	16.8	18.9	17.7
	North Atlantic	3	4626	69	13.2	15.5	15.5	16.2
	North Pacific	4	2344	112	17.7	17.2	16.8	17.5
	Southern Ocean	5	7856	75	12.5	12.4	16.4	12.8
	Global		20 966	289	13.2	15.1	16.0	15.6
(b)	Subtropical	1	4917	94	10.2	10.2	11.0	10.4
	Eq. Pacific	2	513	7	6.9	12.4	9.4	13.0
	North Atlantic	3	3181	53	7.7	10.0	7.9	10.1
	North Pacific	4	1956	88	14.3	16.4	14.8	16.6
	Southern Ocean	5	6084	58	8.0	9.1	9.4	9.8
	Global		16 651	224	9.5	10.8	10.4	11.1

[a] Corresponding geographical region in Fig. 3.
[b] Number of measurements in the corresponding region.
[c] Number of unique cruises/time series in the region.

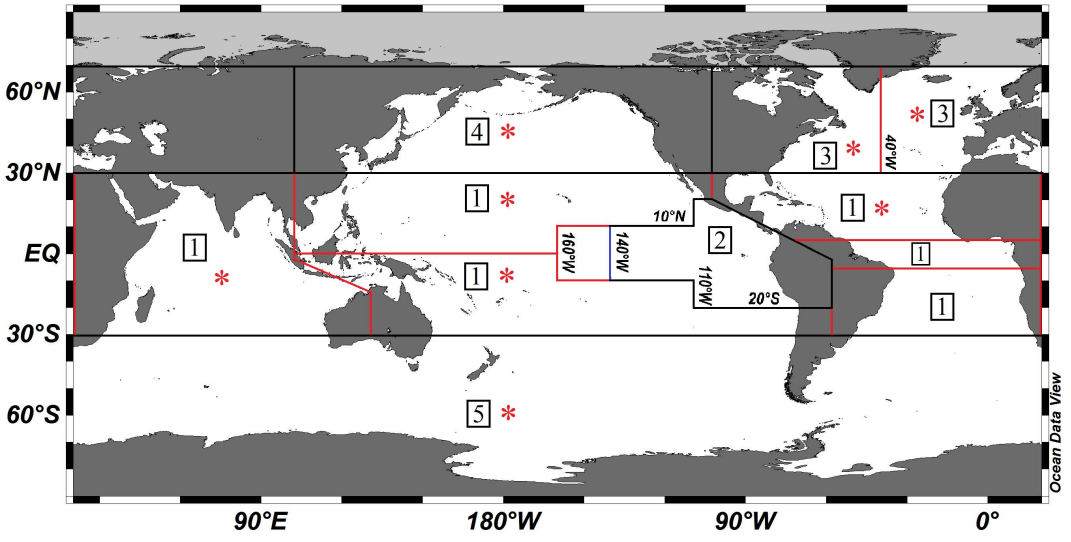

Fig. 3. Spatio-temporal division of the global training dataset for the ad hoc MLR approach. Black boundaries are common for both C_T and A_T models, while red boundaries are for C_T only and blue for A_T only. A red asterisk indicates that MLRs were developed for both summer (Nov–Apr for austral hemisphere) and winter (May–Oct for austral hemisphere) periods to constrain C_T. Boreal summer/winter seasons are opposite.

grid. Using an algorithm that employs discrete clustering is appealing, as it removes the need for the type of ad hoc partitioning we discussed in Sect. 4.2. This has led to application of SOMs in a wide range of disciplines (Abramowitz, 2005; Hsu et al., 2002; Pöllä et al., 2009).

Figure 4 illustrates the routine of SOM training and prediction. For a training dataset of P samples consisting of predictor variables $\boldsymbol{x}$ and response variable y, the SOM clustering process allocates each sample to one of J neurons (sometimes also called clusters, nodes or groups). The neurons are typically arranged in a 2 dimensional $\mathbf{A} \times \mathbf{B}$ matrix so that we represent a node as $j_{a,b}$. The clustering algorithm aims to ensure that nodes that are nearby in this matrix contain samples that have similar values of the predictor variables $\boldsymbol{x}$. The $y = f(\boldsymbol{x})$ input–output mapping is then completed by

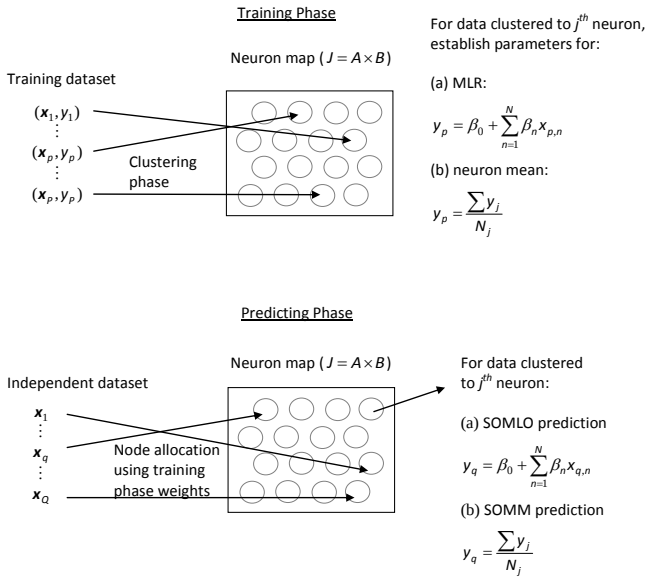

Fig. 4. Schematic diagram of neural network training and prediction phases.

performing a linear regression between $\boldsymbol{x}$ and y separately for each neuron.

These SOM and regression parameters can then be used to make predictions of y for an independent set of Q predictor samples $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q, \ldots, \boldsymbol{x}_Q)$. First, each $\boldsymbol{x}_q$ is allocated to a SOM neuron, based on its similarity to the SOM weights from the training dataset. This is the "winning neuron" for a particular sample $j(\boldsymbol{x}_q)$. Then the regression parameters for $j(\boldsymbol{x}_q)$ are used to predict y_q.

Here we explore two variants to this approach. The first, as described above, uses a multiple linear regression at each neuron, which we describe here as self-organizing multiple linear output (SOMLO). The second takes the mean of all response values belonging to a node, which we will call self-organizing map mean (SOMM). We now describe both in more detail.

5.2 Initialization of the model constraints

For our implementation, the input–output pairs $(\boldsymbol{x}_p, y_p)$, $1 \leq p \leq P$ in the training dataset are some subset of $\boldsymbol{x} = (T, S,$ DO, N, Si, P), and $y = C_\mathrm{T}$ or A_T. To ensure each predictor variable has an equal opportunity to define the features of the SOM during the training routine, we zero-mean and scale the variables by their standard deviation so that their distribution and range are similar. For nitrate, phosphate and silicate, due to the exponential distribution of their measurements, we first $\log_{10}$ scale their measurements.

The J-neuron SOM we use here is structured in a hexagonal topology for the current study (Fig. 4). Careful consideration needs to be exercised when defining the size of the SOM, as too few neurons will not capture all important features, while too many will over-fit the training dataset. Each neuron $(j_{a,b})$ is then assigned an initial weighting vector $(\boldsymbol{\omega})$ of length equal to the number of input variables, and whose values are randomly selected from the input variable range.

5.3 SOM training routine

Once all the neuron weights have been initialized, training is an iterative process designed to cluster the P samples into J neurons. For each iteration step of the model (τ), the input data samples are individually presented to the SOM in a random order and the neuron whose weights are closest to the current input sample is declared the "winning neuron" for that sample, using

$$\text{distance}\,(\boldsymbol{x}_p, \boldsymbol{\omega}_j) = \left[\sum_{n=1}^{N} \left(x_{p,n} - \omega_{j,n}\right)^2\right]^{0.5}. \quad (3)$$

That is, the "winning neuron", $j(\boldsymbol{x}_p)$, for sample $\boldsymbol{x}_p$ is simply the neuron that minimizes this distance. Once the winning neuron is established, the weights of the winning neuron, as well as those neurons in its topological neighbourhood in the SOM, are then adjusted towards the value of the current sample value $(\boldsymbol{x}_p)$ via

$$\boldsymbol{\omega}_j(\tau + 1) = \boldsymbol{\omega}_j(\tau) + h_{j,j(x_p)}\left(\boldsymbol{x}_p - \boldsymbol{\omega}_j(\tau)\right). \quad (4)$$

In this expression, $h_{j,j(x_p)}$ determines the extent to which a node's weight is brought closer to the current sample value (termed a "learning rate", $h \leq 1$). It also determines the size of the neighbourhood around the winning node that receives a significant adjustment. We use

$$h_{j,j(x_p)} = \eta(\tau) \exp\left(-\frac{d_{j,j(x_p)}}{2\sigma^2(\tau)}\right), \quad (5)$$

where $d_{j,j(x_p)}$ represents the discrete distance in the SOM topology between the winning neuron $j(\boldsymbol{x}_p)$ and an arbitrary neuron j, and $\sigma^2(\tau)$ and $\eta(\tau)$ are the neighbourhood width and learning rate respectively. As the model progresses through iterations, $\sigma^2(\tau)$ ensures that the neighbourhood width shrinks from a value that significantly adjusts most of the neurons to finish with only adjusting the winning neuron. Similarly, the learning rate $\eta(\tau)$ decreases with iterations, so that regional features of the SOM gradually develop as iterations continue.

The form of the model used here is known as a supervised SOM, whereby distributional information of the response parameter (C_T or A_T) is used as an additional constraint beyond the hydrographic information (T, S, DO, etc.) in clustering the global dataset into the set of J neurons. For more detail see Supplement E.

5.4 Completing the input–output mapping

We complete the $y = f(\boldsymbol{x})$ in one of two ways. First, the mean of all output values y_p belonging to a node is used

– the SOMM. Alternatively, we use MLRs with the training data assigned to the winning neuron to establish this relationship (see Fig. 4). Here we use MLRs after the SOM training through the application of either a principal component regression (PCR; see supplement F for details) or our forward stepwise robust MLR routine (see Sect. 4.1). To ensure confidence in regression coefficients, a minimum threshold value of 10 times the number of predictor parameters was implemented. If the number of data points assigned to the winning neuron is below this threshold value, data from the second most similar neuron is merged with the winner, and then third, until the data pool reaches the threshold limit.

5.5 Predicting with the SOMLO / SOMM system

For any independent input data vector ($\boldsymbol{x}_q$), we can predict the output value (y_q) using the SOM trained above via a two-step process. First, determine which neuron in the SOM each new data sample is closest to using the distance measure in Sect. 5.3 (Eq. 3). Then the output value (of C_T or A_T) is determined using either the mean value of the winning neuron's training output values (using the SOMM) or the regression parameters established with training data.

6 Application to the global ocean

6.1 Optimization of the global model

To converge on the optimal SOMLO approach for the ocean carbon mixed-layer dataset, we employed a two-phase process. Firstly, three unique subsets of ocean carbon data were extracted to ascertain which hydrographic parameter combination worked best. In the second phase we applied the SIT approach to make an out-of-sample assessment of the global skill of the model.

6.1.1 Defining optimal predictor parameters

Correlations between hydrographic parameters may lead to redundancy in the information predictor variables provide. To investigate the importance of each variable in informing the SOM or constraining the MLR, we perform tests that exclude the variables one at a time (Fig. 6). These test the ability of the models to capture three unique independent datasets that each represent about 10 % of the global carbon dataset (Table 4). As an example, Fig. 5 presents the spatial distribution of the T1 independent dataset, constituting 11.4 % of the global training dataset.

To explore the optimal SOM configuration, 800 iteration steps were used to train the SOM, using neuron map sizes ranging from 9 to 529 for every different input variable combination, with the ultimate aim to converge on the model with the lowest RSE.

Salinity was found to be the most important parameter for capturing the mixed-layer carbon signal, followed by temperature then nutrients (Fig. 6). The final optimal parameter set and SOM neuron size using the three independent tests were (SOPSi, 25) and (TSPO, 56) for the global A_T and C_T models respectively (Fig. 7). For C_T, the SOMLO model using PCR constrained the system with a higher skill than the robust MLR, whilst A_T was better constrained using the robust MLR model.

Table 4. Summary of the three independent datasets used to constrain the general configuration of the SOMLO model.

Independent dataset	Number of measurements	Percentage of global dataset
T1	3769	11.4
T2	2919	8.8
T3	3391	10.2
Total	10 079	30.4

The addition of phosphate beyond temperature, salinity and dissolved oxygen improved the prediction of C_T by $\sim 27\%$ or $5.1\,\mu mol\,kg^{-1}$ (Fig. 7). Without air–sea gas exchange modulating its behaviour, phosphate likely provides clearer constraints on organic matter production and respiration than dissolved oxygen alone. The redundancy of nitrate for both C_T and A_T (Fig. 7) is likely due to the near constant stoichiometric uptake rate of phosphate and nitrate by photosynthesizing organisms. The preference of phosphate over nitrate may be a result of the continual production of organic matter by nitrogen fixers after the nitrate pool is completely depleted (Gruber and Sarmiento, 1997). Furthermore, the renaming of samples where only "nitrate + nitrite" was listed to nitrate in the GLODAP and CARINA products (Key et al., 2004) may serve to introduce additional biases in using nitrate.

Precipitation and dissolution of calcium carbonates ($CaCO_3$) affects the concentration of A_T twice as much as C_T (Sarmiento and Gruber, 2006). As waters high in silicate tend to relate to high biological respiration by diatoms (a non-calcifying organism), and waters of low silicate foster a more conducive environment for calcifying organisms (such as coccolithophores) (Kirchman, 2012), silicate helps constrain the spatial patterns of $CaCO_3$ cycling which influence A_T.

Salinity's significant importance in constraining the A_T system is likely due to the known high correlation between these two parameters (Millero et al., 1998), whereas the addition of temperature to the parameter set is redundant, as pointed out by some earlier studies (e.g. McNeil et al., 2007).

6.1.2 Importance of geography in the model

Carbon data from geographically diverse ocean regions will be clustered into the same neuron when input–output concentrations are similar. For example, a cluster of similar biogeochemical data in the North Atlantic Ocean can be equally

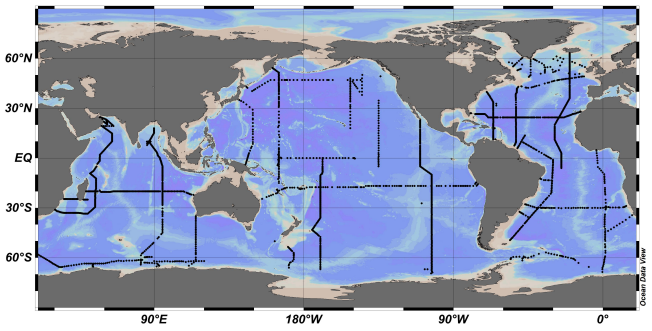

Fig. 5. Distribution of the T1 independent dataset.

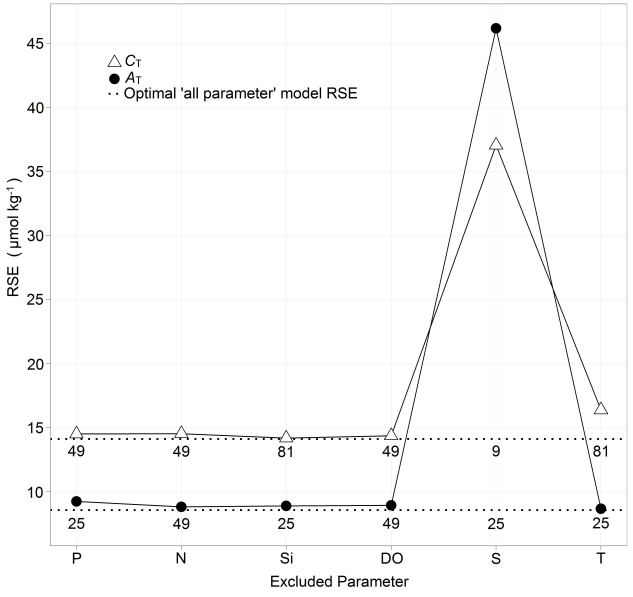

Fig. 6. RSE results for the C_T (open triangles) and A_T (black circles) SOMLO models when applied to the three independent datasets. Numbers under the dotted lines represents the optimal number of neurons to constrain the system. Excluded parameter represents the variable not used in the SOMLO training and testing.

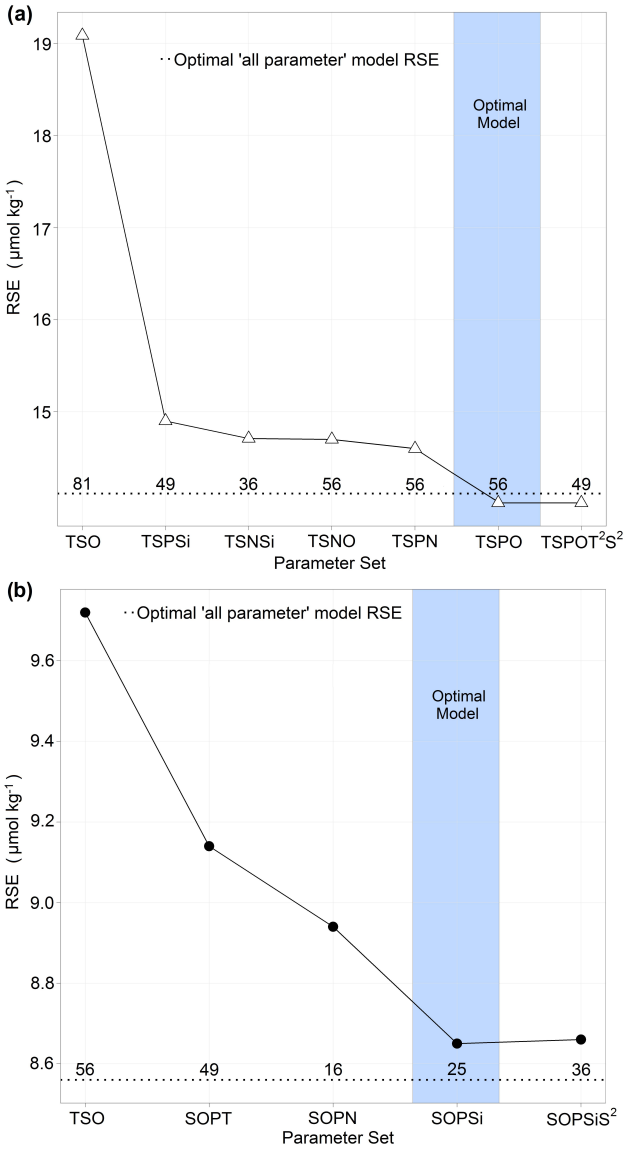

Fig. 7. Optimal RSE values for **(a)** C_T and **(b)** A_T SOMLO models. Numbers above the line represent the optimal number of neurons. Parameter set represents the combination of parameters used to train and test the SOMLO model, where O represents dissolved oxygen.

represented by those in some parts of the North Pacific Ocean, despite there being little ocean inter-connectedness between these two carbon datasets on shorter timescales. Spatial length scales of variability are known to be within ocean basins, not between them, especially those constrained by land. Without applying geographical boundary conditions, non-linearities may be introduced into the final MLR, which would limit the models predictive skill. To test this hypothesis, optimal model configurations were trained with the inclusion of geographical input parameters during the training of the SOM, but were excluded as predictor parameters in the linear regressions.

To reduce the influence of longitudinal discontinuity at ±180° in the mid-Pacific, we shifted all longitude values by 160° W (or 20° E), thereby setting the 180° discontinuity at a position that bisects continental Africa and Europe (see Supplement Fig. F1). We also tested a normal vector to the Earth ellipsoid (n-vector) that transforms the 2-D latitude/longitude position system into a 3-D vector while maintaining unique vectors for every geographical position. Employing a version of the n-vector presented by Gade (2010), we transformed latitude and longitude values using

$$n = \begin{bmatrix} \sin(\text{latitude}) \\ \sin(\text{longitude})\cos(\text{latitude}) \\ \cos(\text{longitude})\cos(\text{latitude}) \end{bmatrix}. \quad (6)$$

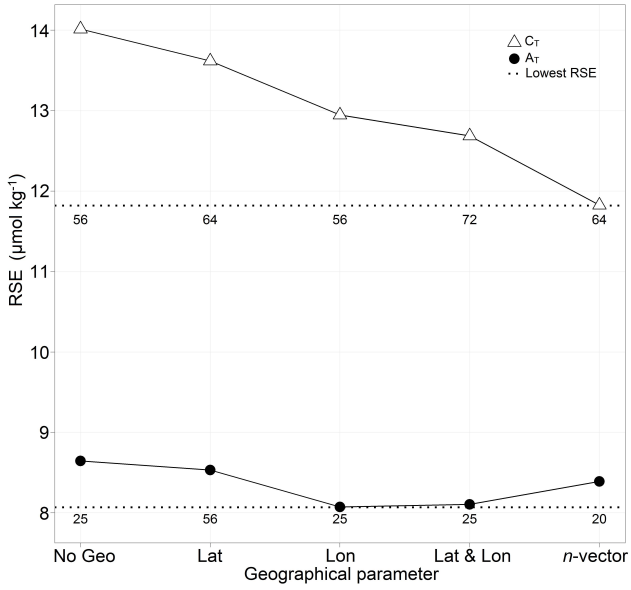

Fig. 8. Skill of optimal SOMLO models with geographical constraints. Numbers below dashed line represent the optimal number of neurons.

We found introducing geographical information to be a powerful addition in improving the skill of the method for C_T by 16 % or 2.2 µmol kg^{-1}, however there was little improvement for A_T (Fig. 8). The optimal SOMLO configuration additionally incorporates longitude and n-vector geographical inputs in constraining A_T and C_T respectively, and increased the optimal number of neurons to 64 for C_T.

To better understand and visualize why geography is important, we compare the spatial distribution of neurons for C_T models trained with only biogeochemical information, and both biogeochemical and geographical information (Fig. 9a, b). To illustrate the spatial distribution of the assigned neurons for the global carbon dataset, we plot the neurons using different colours. Here, each colour represents a neuron, while shades of colours indicate close similarity in the weighting vectors. The broad regions of similarity that are captured when the SOM is constrained by only biogeochemical properties include the Southern Ocean, sub-tropical gyres, North Pacific and North Atlantic (Fig. 9a). However, these ocean "fingerprints" extend beyond the known spatial length scales, for example linking features in the Southern Ocean to those of the North Atlantic, while zonal bands stretch across ocean basins (Fig. 9a). When biogeochemical and geographical information are incorporated into the SOM training routine, the resulting distribution preserves the neuron boundaries at known frontal zones, such as the sub-tropical convergence zone, but is able to constrain the classification of data to be within each ocean basin (Fig. 9b). Using geography is an important additional constraint that implicitly shortens the length scales of variability which dominate seasonal mixed-layer dynamics in the ocean.

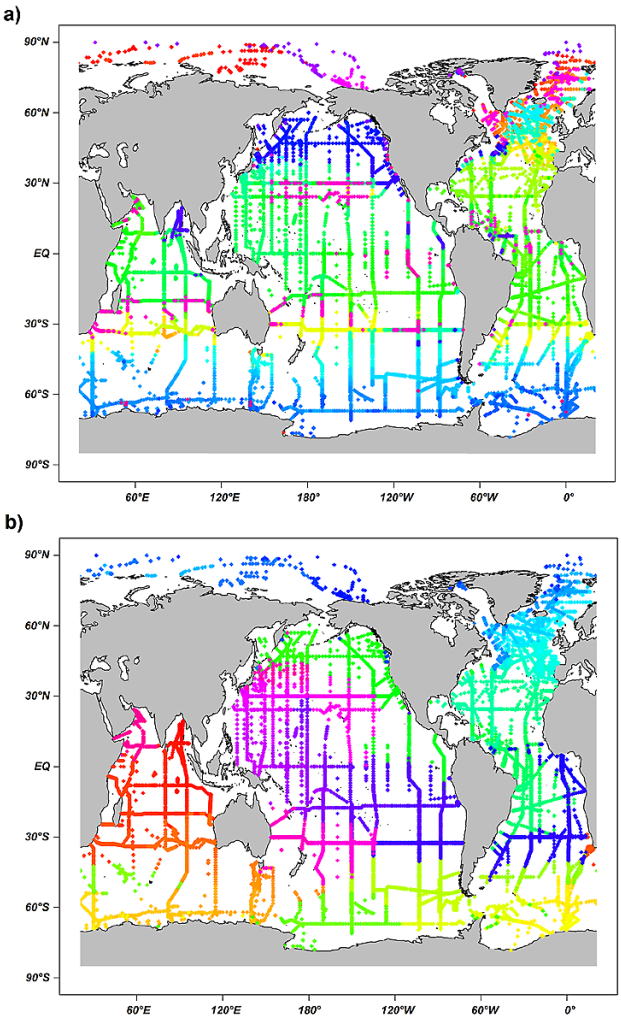

Fig. 9. Distribution of assigned neurons for optimal C_T SOM models trained with **(a)** biogeochemical information only and **(b)** biogeochemical and geographical information.

It is important to note that the addition of geography did not alter the optimal parameter set for the technique.

6.1.3 SOMM/SOMLO comparison

Optimal model configurations were tested with neuron sizes extending up to 2500 to explore the ability of the SOMM model in constraining the three independent datasets (Fig. 10). Using all data the SOMM model converged on an RSE value of 16 µmol kg^{-1} in constraining C_T. Although the SOMM is powerful in constraining complex non-linear datasets, spatio-temporal limitations in the current global carbon dataset hamper the SOM's mean-mode ability to predict C_T on a global scale.

We found using a local multiple-linear optimizer (i.e. the MLR) in addition to the global SOM optimizer to significantly improve the model's ability to constrain global C_T by

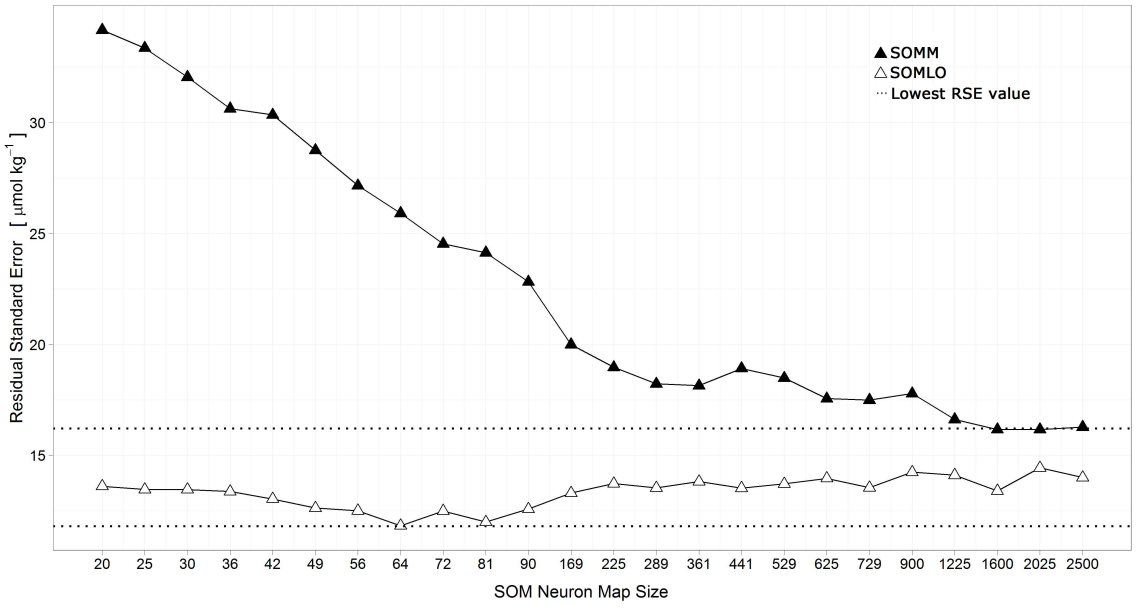

Fig. 10. Skill comparison between the SOMLO and SOMM models in capturing C_T.

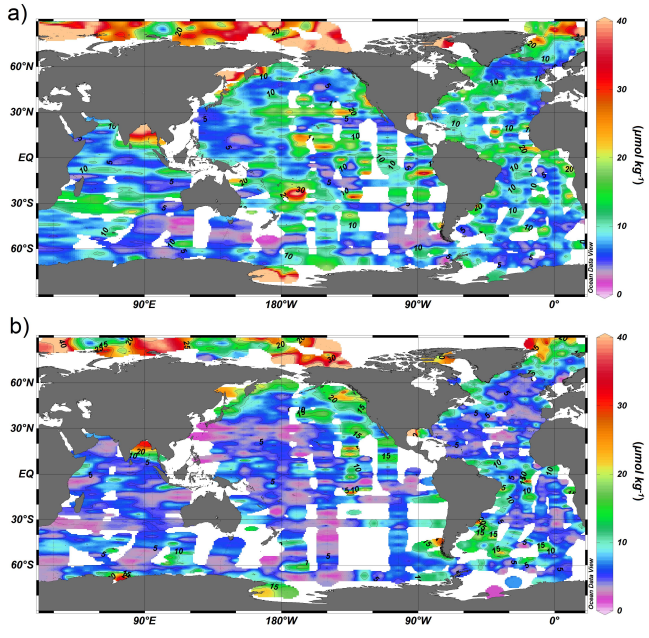

Fig. 11. Distribution of global systematic independent test absolute residual errors for **(a)** C_T and **(b)** A_T.

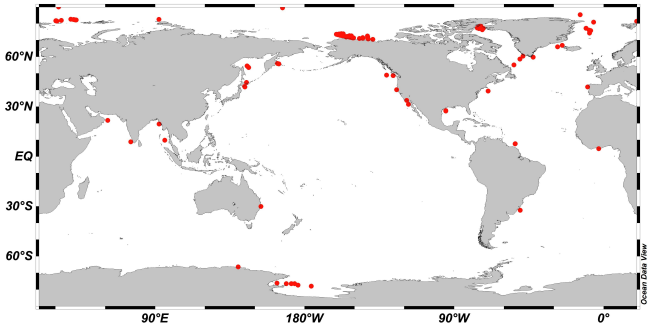

Fig. 12. Geographical distribution of the 277 samples located within 300 km of a major coastline and with a SIT residual error greater than ±50 µmol kg^{-1} for C_T and/or A_T.

$\sim$27% or 4.4 µmol kg^{-1}. Similar findings are found for the A_T model.

6.2 Measuring the improvement over traditional MLR

To evaluate the skill of the two independent approaches used here (MLR versus SOMLO), we tabulated the results of each technique based on the global SIT predictions divided into 5 geographical regions and evaluated globally (Table 5). The SOMLO approach improves the predictive skill of C_T by between 11 and 30 % for all 5 regions (Table 5). In particular, known complex dynamical regions with global CO_2 importance like the equatorial Pacific, Southern Ocean and North Atlantic are where the non-linear SOMLO approach excelled, improving the prediction of C_T by between 23 and 30 % (or 4–6 µmol kg^{-1}). From a global point of view, SOMLO improves the predictive skill of C_T in the mixed layer by $\sim$ 19 %.

For A_T, the benefits of using SOMLO are much weaker, with only a marginal global improvement by 6.7 % (or 0.7 µmol kg^{-1}) and even deterioration of detection in the equatorial Pacific and North Atlantic. This is most likely a result of the carbonate system being less prone to non-linearities and complexity, thereby limiting the benefits of

Table 5. Skill comparison between the traditional MLR and SOMLO approaches for (a) C_T and (b) A_T.

	Region	Zone [b]	N[c]	RSE[a] (μmol kg^{-1}) Ad hoc MLR	SOMLO	% Improvement
(a)	Subtropical	1	5388	15.2	13.5	11.2
	Eq. Pacific	2	752	18.9	13.3	29.7
	North Atlantic	3	4626	15.5	11.7	24.5
	North Pacific	4	2344	16.8	14.3	14.9
	Southern Ocean	5	7856	16.4	12.7	22.6
	Global		20 966	16.0 (15.6)[d]	12.9	19.4 (17.4)[d]
(b)	Subtropical	1	4917	11.0	9.2	16.4
	Eq. Pacific	2	513	9.4	9.6	−2.1
	North Atlantic	3	3181	8.0	8.5	−6.3
	North Pacific	4	1956	14.8	14.4	2.7
	Southern Ocean	5	6084	9.4	8.8	6.4
	Global		16 651	10.4 (11.1)[d]	9.7	6.7 (12.6)[d]

[a] Calculated using the SIT predictions.
[b] Corresponding geographical region in Fig. 3.
[c] Number of measurements.
[d] Universal MLR.

Table 6. Regional and global SOMLO skill evaluation (see supplementary Fig. F2 for map of spatial division).

Region	Zone[a]	RSE[b] C_T	RSE[b] A_T	N[c] C_T	N[b] A_T
Arctic Ocean	1	26.6	22.1	782	795
Subpolar North Atlantic	2	11.6	9.0	4425	2641
Subtropical North Atlantic	3	9.1	6.6	1481	1254
Equatorial Atlantic	4	13.7	13.0	654	582
Subtropical South Atlantic	5	10.6	8.7	659	551
Subpolar North Pacific	6	11.2	14.7	2053	1615
Subtropical North Pacific	7	11.1	8.2	2367	1446
Equatorial Pacific	8	11.2	8.3	1524	802
Subtropical South Pacific	9	12.3	7.7	1824	1404
Subtropical North Indian (Exc. Bay of Bengal)	10	22.1 (13.9)	13.4 (7.5)	143 (111)	168 (136)
Equatorial Indian	11	11.8	7.7	512	500
Subtropical South Indian	12	11.5	5.6	1411	1388
Southern Ocean	13	8.7	8.8	3950	3088
Sub-Antarctic waters	14	9.5	8.5	2250	1474
Global		11.8	10.2	24 035	17 708
Global (below 70° N)		10.9	9.2	23 253	16 913

[a] Corresponding geographical region in Supplement Fig. F2.
[b] Residual standard error (μmol kg^{-1}).
[c] Number of measurements in the region.

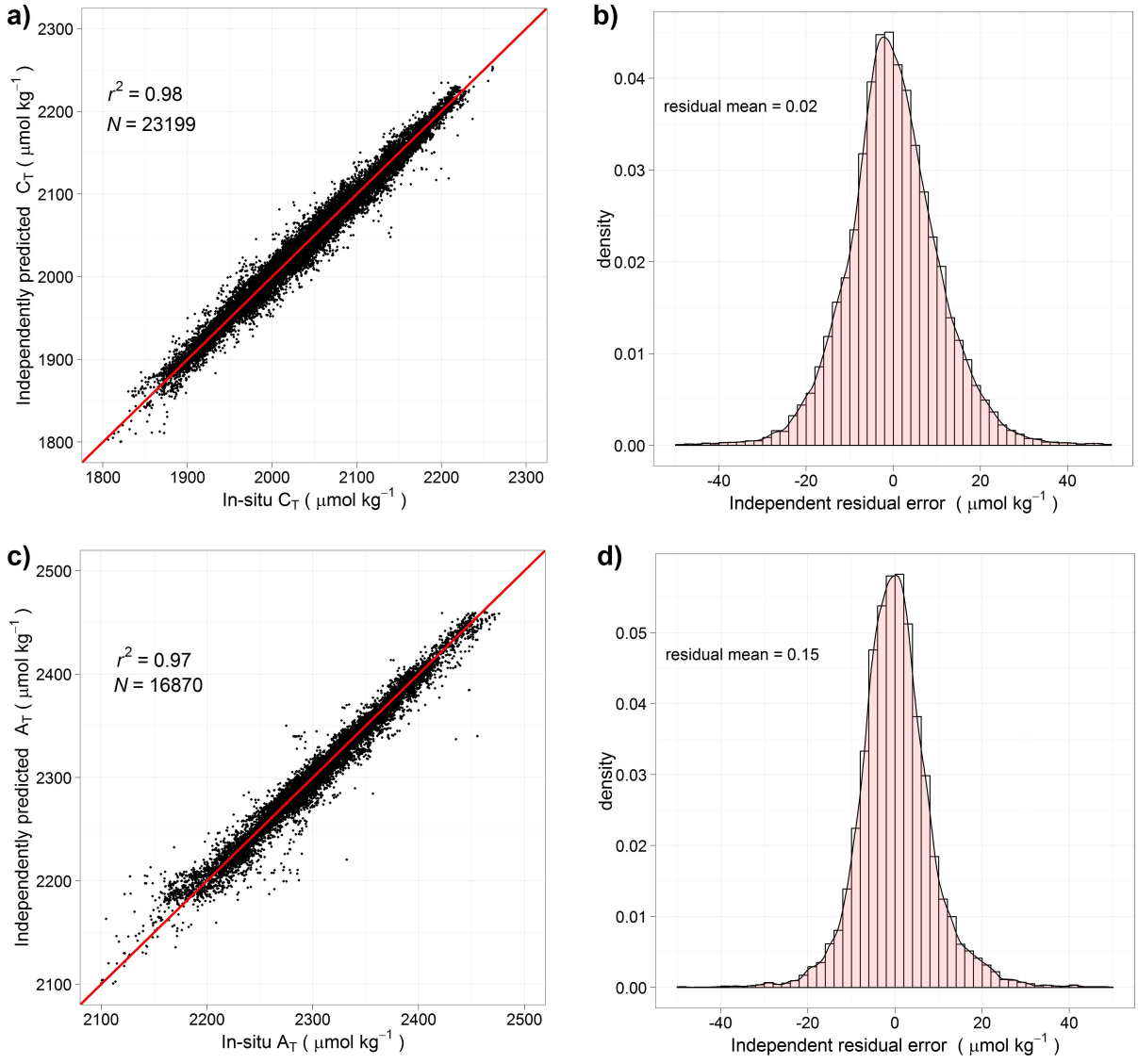

Fig. 13. Global independent test (SIT) predictions versus in situ measurements and residual error density distribution for optimal **(a–b)** C_T and **(c–d)** A_T SOMLO configurations. r^2 is r-squared correlation, and N is number of samples.

SOMLO, since it better constrains more complex non-linear systems.

6.3 SOMLO regional error assessment

To scrutinize the spatial skill of the SOMLO model, absolute values of the global SIT residual errors were interpolated around the in situ sample locations using VG gridding software in the Ocean Data View (ODV) program (Schlitzer, R.: Ocean Data View, http://odv.awi.de, 2011). Although the Arctic Ocean, Bay of Bengal and Sea of Okhotsk are regions not well constrained by the technique, the majority of the ocean maintains a relatively homogenous residual error range (Fig. 11). These unconstrained regions are either coastal or marginal seas with known locally complex biogeochemical regimes, so it is understandable that a trained global open-ocean technique will poorly constrain these local regions.

Further investigation of the 395 samples with a SIT residual error greater than ±50 µmol kg^{-1} for C_T and/or A_T, revealed that 70 % (277) are located within 300 km of a major coastline (Fig. 12). Since a study by Gibbs et al. (2006) identified terrestrial influences extend up to 345 km from land and well beyond our bathymetric defined coastal ocean limit of 500 m (Supplement B), these anomalous independently predictions are likely the result of land–ocean interactions affecting the carbon and SHP concentrations. Separating the SIT predictions into 14 different regions and removing these anomalous coastal samples then provides the most accurate constraint on the models regional open-ocean skill (Table 6). Again we find the Arctic Ocean and Bay of Bengal are the

two regions were the model's skill is poorest. Through the exclusion of Arctic Ocean measurements (North of 70°), the final estimate for the global open-ocean accuracy for C_T and A_T is 10.9 and 9.2 µmol kg^{-1} respectively.

To investigate skewness, we plot the SOMLO global SIT predictions versus the in situ measurements (Fig. 13a, c). For C_T, skewness is limited ($R^2 = 0.98$), giving us confidence in the model's ability to accurately capture the concentrations of C_T and A_T for any given set of temperature, salinity, dissolved oxygen, (silicate for A_T), and phosphate measurements in the open ocean mixed layer.

Finally, we found no strong seasonal bias in our SOMLO predictions (Fig. 14).

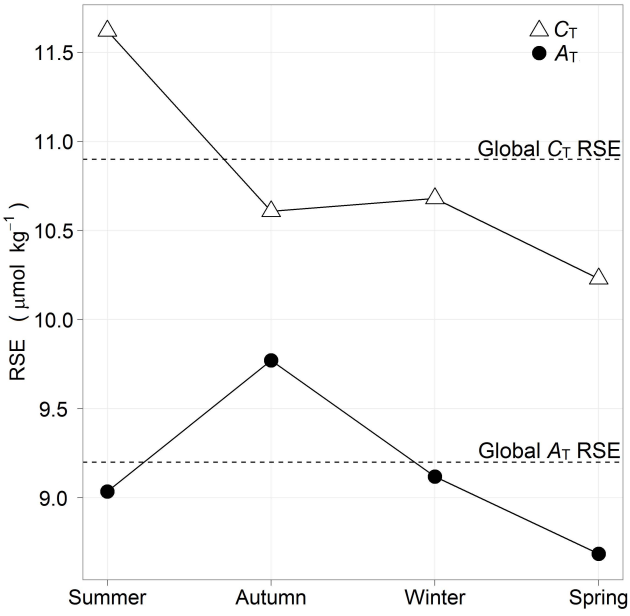

Fig. 14. SOMLO C_T/A_T seasonal independent test RSE values. Southern Hemisphere seasons are defined as summer (Dec–Feb), autumn (Mar–May), winter (Jun–Aug) and spring (Sept–Nov). Northern Hemisphere seasons differ by 6 months.

7 Application to the Bermuda Atlantic and Hawaiian ocean time-series sites

The SOMLO technique was trained on a global C_T and A_T dataset that consisted mostly of sporadic one-time cruises in time. To test how well seasonal to inter-annual variability is captured using our technique, we use carbon time-series data from the BATS and HOT stations as a test bed.

7.1 Predicting the North Atlantic seasonal cycle for inorganic carbon (BATS)

Located in the Sargasso Sea, the BATS hydrographic site is a high frequency measurement program of carbon and auxiliary parameters that has been ongoing since 1989. To test the global SOMLO model in reconstructing the BATS seasonal cycle, we firstly re-trained the global algorithm without using the BATS 1989–2007 carbon time-series dataset. We then use the measured monthly hydrographic properties between 1987 and 2007 to independently predict C_T and A_T concentrations at the BATS site and finally compare our predicted carbon values to the in situ measurements to investigate the skill of the technique. We also independently predict C_T/A_T values with the traditional MLR approach as a further test.

Figure 15a and b shows the measured versus predicted C_T and A_T annual cycles at BATS. Within the uncertainty of the SOMLO prediction, both the magnitude and structure of the seasonal C_T cycle at BATS is well constrained, capturing 90 % of the signal (Fig. 15a). For a global MLR approach, the seasonal cycle is overestimated significantly by ∼ 50 %. For A_T, the small seasonality is captured by both techniques (Fig. 15b).

To gain better insight into how the SOMLO substantially improves the prediction of the BATS seasonal cycle from the traditional MLR analysis, we investigate the neuron distribution for C_T in the northwestern Atlantic (Fig. 16). Applying a traditional ad hoc MLR analysis requires defining somewhat subjective longitude and latitude boundaries for the data to be used in the linear regressions. Here, as an illustration we use the spatial boundaries of 30 to 70° N and 40 to 85° W that were also used by Lee et al. (2000) in their MLR approach. The traditional MLR explicitly uses all carbon data within the prescribed region, whilst the SOMLO approach partitions the data into neurons without any prior geographic constraints. The benefit in this approach is that when we are applying the SOMLO to a new dataset (in this case BATS) the SOM only uses neurons (data) most consistent with its "biogeochemical fingerprint", and therefore reduces the potential bias that would be introduced from including all data in the regression.

7.2 How well does SOMLO capture inter-annual signals?

Inter-annual variability of C_T at BATS is captured to within the uncertainty of the SOMLO technique over the 18 yr period (Fig. 17). This illustrates a new potentially powerful way to diagnose year-to-year carbon variability in the ocean by using the many more long-term hydrographic time series that are available in the ocean (McNeil, 2010). To further test the SOMLO approach in capturing inter-annual variability of C_T, we predict the C_T signal at the HOT time series as reported by Brix et al. (2004). The SOMLO prediction captures the smoothed inter-annual trend line at the HOT site to within 85 % (Fig. 18).

The BATS and HOT comparisons provide additional confidence that the SOMLO approach provides good constraints on both seasonal and inter-annual variability of C_T, so that it

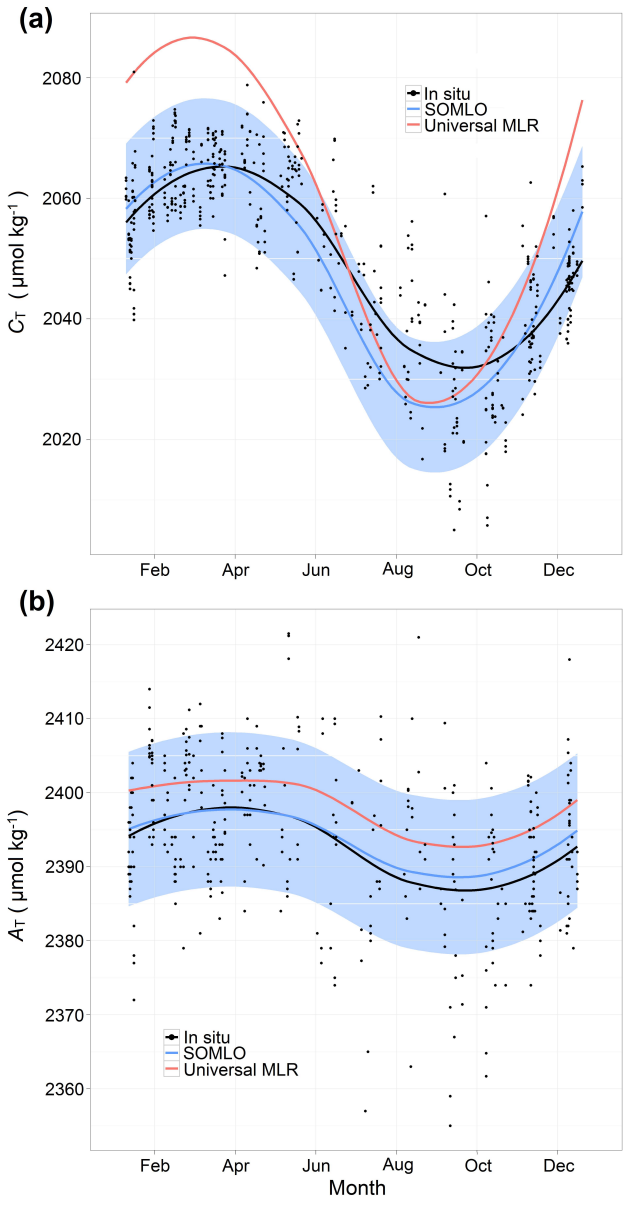

Fig. 15. BATS in situ and independently predicted seasonal cycles for **(a)** C_T and **(b)** A_T. Black dots show the in situ measurements and blue shaded region represents the uncertainty in SOMLO predictions.

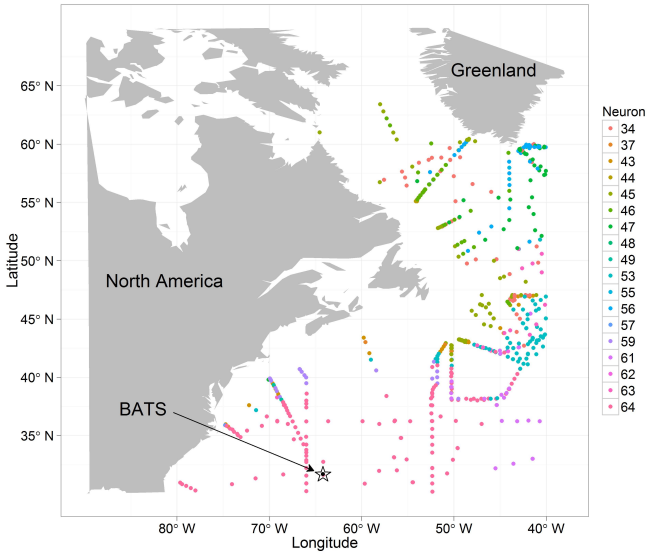

Fig. 16. Distribution of assigned neurons in the northwestern Atlantic region for optimal C_T SOMLO model (30 to 70° N and 40 to 85° W). Numbers represent the neuron each measurement was assigned to (maximum of 64).

could be used on a wider scale to help understand the ocean's role in modulating atmospheric CO_2.

8 Comparison to previous techniques

It is important to emphasize reported error estimates of previous empirical studies to those calculated here. RSE values presented by previous empirical studies (see Table 1) are calculated from the regression's residual error rather than independent tests as done here, so direct comparisons between previous studies and our results are not valid. We use the systematic independent test approach (see Sect. 3) in order to accurately report the differences between our results and previous traditional MLR results.

We conduct two sets of calculations as shown in Table 7. The first set of calculations (MLR_{old}) involves taking the regressions from a suite of prior work (Bates et al., 2006; Lee et al., 2006, 2000; McNeil et al., 2007) and applying it to the new larger dataset within each region. The second set of calculations (MLR_{new}) involved developing our own set of regressions using the same geographical and temporal boundaries and predictor variables as the previous authors within the much larger dataset. Using the SIT predictions, the skill of the models were calculated (RSE) and could then be directly compared to our SOMLO values (see Table 7).

The SOMLO, as shown at BATS/HOT, improves the predictive skill of C_T and A_T in most regions by between 10 and 40 %. Globally for C_T, the SOMLO reduces the error by 28 % beyond the MLR method that was used to conduct the only global analysis (Lee et al., 2000).

9 Diagnosing global C_T and A_T distributions

Large historical and recent datasets up until 2008 of temperature, salinity, dissolved oxygen and nutrients has allowed researchers to objectively interpolate monthly $1° \times 1°$ global climatologies (Antonov et al., 2010; Garcia et al., 2010a, b; Locarnini et al., 2010). Here, we employed the WOA09 ocean surface climatologies in conjunction with our SOMLO

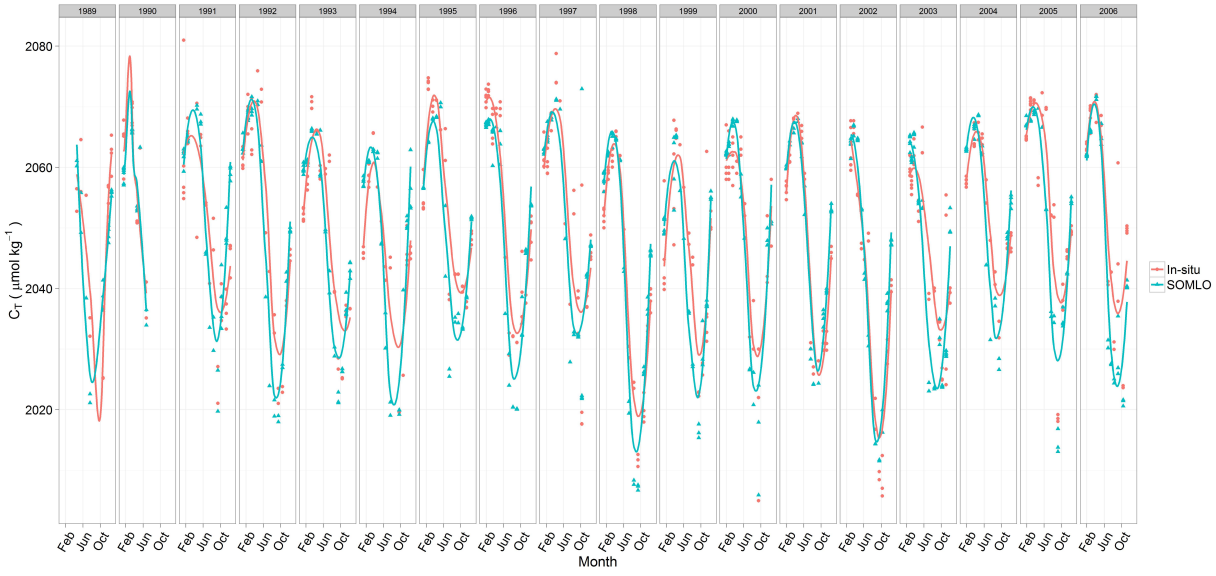

Fig. 17. In situ and independently predicted BATS C_T measurements partitioned into years with loess fit (locally weighted scatter-plot smoothing).

Fig. 18. In situ and independently predicted HOT C_T measurements with loess fit.

model to diagnose monthly C_T and A_T distributions for the nominal year 2000.

Large-scale features in our estimated annual mean C_T and A_T distributions exhibit strong agreement with bottle measurements and follows our broader understanding of spatial carbon variability (Figs. 19, 20). In the Southern Ocean, for example, we find longitudinally homogenous bands driven by the Antarctic Circumpolar Current (ACC), and higher C_T concentrations relative to the global-mean due to strong upwelling of CO_2-enriched subsurface waters and cooler surface temperatures enhancing CO_2 solubility (McNeil et al., 2007; Metzl et al., 2006). In equatorial upwelling regions, cold waters enriched with remineralized organic material are brought to surface resulting in elevated C_T and A_T concentrations (Feely et al., 2002). As the surface water is then transported laterally from the site of upwelling, biological processes and loss of CO_2 to the atmosphere reduces C_T to some of the lowest concentrations observed globally. For A_T, maxima concentrations are found in the central subtropical gyres ($\sim 25°$), where stronger evaporation relative to precipitation drives higher ocean surface salinity and therefore A_T concentrations (Lee et al., 2006; Millero et al., 1998). Conversely,

Table 7. Comparison to previous empirical approaches.

Study	Response variable	N	RSE (μmol kg^{-1})			% Improvement[d]	Author
			MLR_{old}	MLR_{new}	SOMLO		
Global[a]	C_T	13881	22.0	17.8	12.8	28	Lee et al. (2000)
Indian[b]	C_T	2052	15.2	21.4	13.0	39	Bates et al. (2006)
Southern Ocean	C_T	4196	17.3	8.8	9.0	−2	McNeil et al. (2007)
Global (exc. North Pacific)[c]	A_T	10360 (8995)	11.7 (10.3)	10.9 (10.4)	10.7 (9.9)	2	Lee et al. (2006)
Indian[b]	A_T	2042	9.4	11.8	7.1	40	Bates et al. (2006)
Southern Ocean	A_T	4196	10.3	10.3	9.3	10	McNeil et al. (2007)

[a] Using only surface data (above 30 m).
[b] Only measurements from within our defined mixed layer were used to constrain new regressions and test previous regressions.
[c] The North Pacific empirical regression of Lee et al. (2006) included an interaction term between temperature and longitude. Here, longitude values were taken to range from 0 to 360°.
[d] Calculated using $((MLR_{new} - SOMLO)/MLR_{new}) \times 100$.

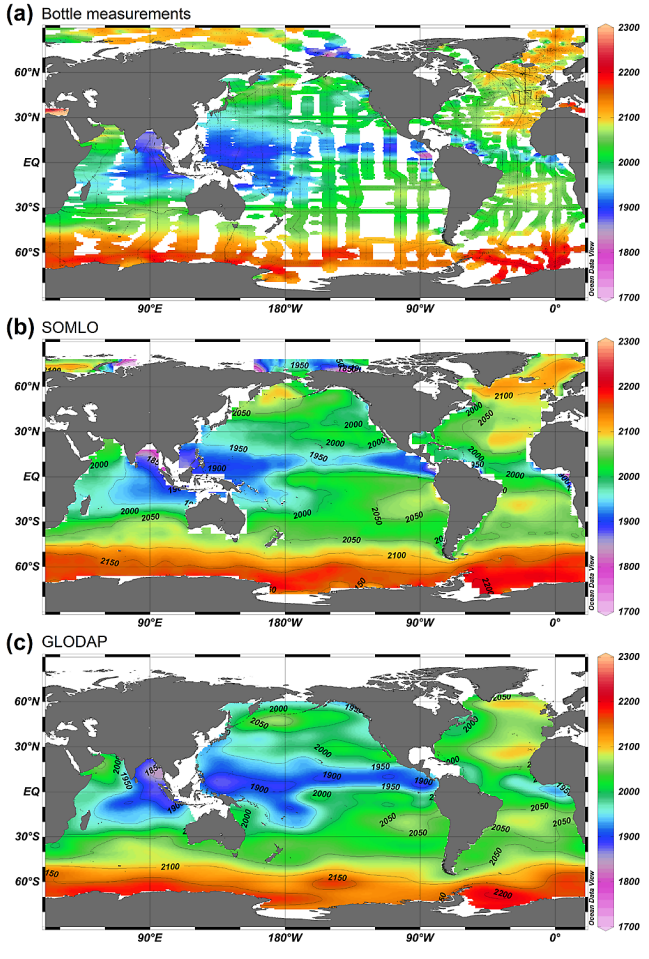

Fig. 19. Global distributions of **(a)** bottle C_T measurements corrected to the year 2000 **(b)** annual-mean SOMLO C_T predictions for the nominal year 2000 **(c)** GLODAP-v1.1 ocean surface C_T distribution of Key et al. (2004).

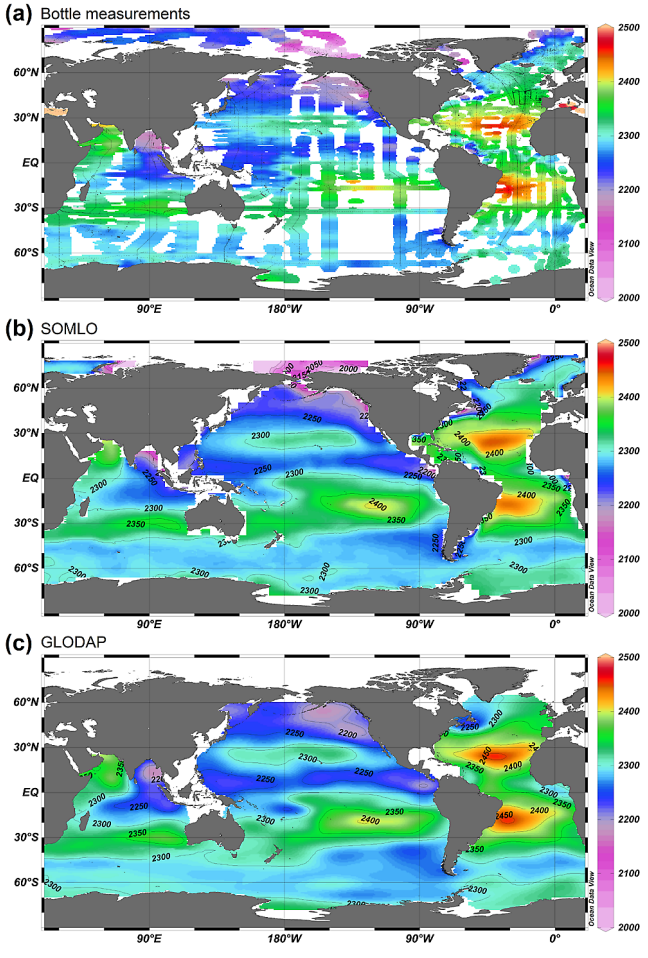

Fig. 20. Global distributions of **(a)** bottle A_T measurements **(b)** annual-mean SOMLO A_T predictions **(c)** GLODAP-v1.1 ocean surface A_T distribution of Key et al. (2004).

freshwater input from rivers and seasonal ice melt lowers A_T concentrations in regions like the Bay of Bengal (George et al., 1994) and Arctic marginal waters.

Key et al. (2004) interpolated bottle measurements collected between 1985 and 1999 to diagnose $1° \times 1°$ resolution global climatologies for C_T and A_T at 33 depth surfaces (GLODAP-v1.1; available at: http://cdiac.ornl.gov/oceans/glodap/Glop_grid_OV.html). Comparison between the GLODAP-v1.1 0 m carbon distributions and our results shows good general agreement (Figs. 19, 20). However, our average C_T concentration between 65° N and 77° S is 14 µmol kg^{-1} higher than the GLODAP-v1.1 average of 2033 µmol kg^{-1}. In particular, the Southern Ocean and equatorial Pacific are where we find the largest discrepancies. This could either reflect the uptake of anthropogenic CO_2 that was not accounted for in the GLODAP study (Key et al., 2004), or result from a 30 % improvement in Southern Ocean data coverage since 1999. However, it's likely that the large spatial and temporal bias within the GLODAP dataset plays the largest role in this discrepancy.

10 Conclusions

Here, we have exploited the global carbon C_T/A_T mixed-layer bottle database ($\sim$ 33 000) to investigate two different empirical approaches that diagnose mixed-layer carbon dynamics from standard hydrographic parameters. Using independent data as a test, the traditional multiple linear regression approach constrains the C_T system to 15.6 µmol kg^{-1} and A_T to 10.4 µmol kg^{-1}. We then deploy a new non-linear neural network based approach that improves the predictive skill by 2.7–3 µmol kg^{-1} for C_T, or $\sim$ 19 % over the MLR, and 0.7–1.4 µmol kg^{-1} for A_T or $\sim$ 10 %. In particular, regions of known complexity and importance to carbon cycling like the Southern Ocean, North Atlantic and equatorial Pacific are where the new non-linear approach excels, reducing errors by up to 35 % over traditional linear approaches. We further test our neural network technique and find it to predict both seasonal and inter-annual variability of carbon at BATS and HOT very well.

The predictive skill of the neural network approach is shown to be spatially and temporally robust, making the model a powerful tool for diagnosing carbon dynamics in the ocean. In reality, the intensity of a sampling regime needed to constrain seasonal to inter-annual variability for carbon is so great that it will always be difficult to achieve on a global scale. We demonstrate here, that the use of non-linear empirical techniques on a global scale could potentially advance our understanding of oceanic carbon variability, particularly in a future where the amount of autonomous hydrographic data is increasing exponentially.

Supplementary material related to this article is available online at: http://www.biogeosciences.net/10/4319/2013/bg-10-4319-2013-supplement.zip.

Acknowledgements. We thank all captains, crew and researchers who helped to collect, analyse and synthesize the in situ bottle dataset used in this study. All training of the model was conducted using the R-statistical project software (R Development Core Team: R: A Language and Environment for Statistical Computing, http://www.R-project.org, 2012), with kohonen, ggplot2 and akima packages developed by Wehrens and Buydens (2007), Wickham (2009) and Akima et al. (akima: Interpolation of irregularly spaced data, http://CRAN.R-project.org/package=akima, 2012), respectively. We also thank the developers of the Ocean Data Viewer (ODV) program (Schlitzer, R.: Ocean Data View, http://odv.awi.de, 2011), from which many figures presented in this paper were developed.

Edited by: L. Cotrim da Cunha

References

Abramowitz, G.: Towards a benchmark for land surface models, Geophys. Res. Lett., 32, L22702, doi:10.1029/2005gl024419, 2005.

Anderson, L. A. and Sarmiento, J. L.: Redfield ratios of remineralization determined by nutrient data analysis, Global Biogeochem. Cy., 8, 65–80, doi:10.1029/93gb03318, 1994.

Antonov, J. I., Seidov, D., Boyer, T. P., Locarnini, R. A., Mishonov, A. V., Garcia, H. E., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 2: Salinity, in: NOAA Atlas NESDIS 69, edited by: Levitus, S., US Government Printing Office, Washington DC, 184, 2010.

Arrigo, K. R., Pabi, S., van Dijken, G. L., and Maslowski, W.: Air-sea flux of CO_2 in the Arctic Ocean, 1998–2003, J. Geophys. Res., 115, G04024, doi:10.1029/2009jg001224, 2010.

Bates, N. R., Pequignet, A. C., and Sabine, C. L.: Ocean carbon cycling in the Indian Ocean: 1. Spatiotemporal variability of inorganic carbon and air-sea CO_2 gas exchange, Global Biogeochem. Cy., 20, GB3020, doi:10.1029/2005gb002491, 2006.

Bender, M. L., Ho, D. T., Hendricks, M. B., Mika, R., Battle, M. O., Tans, P. P., Conway, T. J., Sturtevant, B., and Cassar, N.: Atmospheric O_2/N_2 changes, 1993–2002: Implications for the partitioning of fossil fuel CO_2 sequestration, Global Biogeochem. Cy., 19, GB4017, doi:10.1029/2004gb002410, 2005.

Boyer, T. P., Antonov, J. I., Baranova, O. K., Garcia, H. E., Johnson, D. R., Locarnini, R. A., Mishonov, A. V., O'Brien, T. D., Seidov, D., Smolyar, I. V., and Zweng, M. M.: World Ocean Database 2009, in: NOAA Atlas NESDIS 66, edited by: Levitus, S., US Gov. Printing Office, Washington DC, 216, 2009.

Bradshaw, A. L., Brewer, P. G., Shafer, D. K., and Williams, R. T.: Measurements of total carbon dioxide and alkalinity by potentiometric titration in the GEOSECS program, Earth Planet. Sci. Lett., 55, 99–115, doi:10.1016/0012-821x(81)90090-x, 1981.

Brix, H., Gruber, N., and Keeling, C. D.: Interannual variability of the upper ocean carbon cycle at station ALOHA near Hawaii, Global Biogeochem. Cy., 18, GB4019, doi:10.1029/2004gb002245, 2004.

CARINA Group: Carbon in the Arctic Mediterranean Seas Region – the CARINA project: Results and Data, Version 1.2, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, http://cdiac.ornl.gov/ftp/oceans/CARINA/CARINA_Database/CARINA.AMS.V1.2/, doi:10.3334/CDIAC/otg.CARINA.AMS.V1.2, 2009a.

CARINA Group: Carbon in the Atlantic Ocean Region – the CARINA project: Results and Data, Version 1.0, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, http://cdiac.ornl.gov/ftp/oceans/CARINA/CARINA_Database/CARINA.ATL.V1.0/, doi:10.3334/CDIAC/otg.CARINA.ATL.V1.0, 2009b.

CARINA Group: Carbon in the Southern Ocean Region – the CARINA project: Results and Data, Version 1.1, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, http://cdiac.ornl.gov/ftp/oceans/CARINA/CARINA_Database/CARINA.SO.V1.1/, doi:10.3334/CDIAC/otg.CARINA.SO.V1.1, 2010.

Chen, L., Xu, S., Gao, Z., Chen, H., Zhang, Y., Zhan, J., and Li, W.: Estimation of monthly air-sea CO_2 flux in the southern Atlantic and Indian Ocean using in-situ and remotely sensed data, Remote Sens. Environ., 115, 1935–1941, doi:10.1016/j.rse.2011.03.016, 2011.

Chierici, M., Olsen, A., Johannessen, T., Trinañes, J., and Wanninkhof, R.: Algorithms to estimate the carbon dioxide uptake in the northern North Atlantic using shipboard observations, satellite and ocean analysis data, Deep-Sea Res. Pt. II, 56, 630–639, doi:10.1016/j.dsr2.2008.12.014, 2009.

Department of Energy: Handbook of methods for the analysis of the various parameters of the carbon dioxide system in sea water, Version 2, edited by: Dickson, A. G., and Goyet, C., ORNL/CDIAC-74, Carbon Dioxide Inf. and Anal. Cent., Oak Ridge, Natl. Lab., Oak Ridge, Tenn., 1994.

Dickson, A. G., Afghan, J. D., and Anderson, G. C.: Reference materials for oceanic CO_2 analysis: a method for the certification of total alkalinity, Mar. Chem., 80, 185–197, doi:10.1016/s0304-4203(02)00133-0, 2003.

Dickson, A. G., Sabine, C. L., and Christian, J. R. (Eds.), Guide to best practices for ocean CO_2 measurements, PICES Special Publication 3, 191 pp., 2007.

Falkowski, P., Scholes, R. J., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Högberg, P., Linder, S., Mackenzie, F. T., Moore III, B., Pedersen, T., Rosenthal, Y., Seitzinger, S., Smetacek, V., and Steffen, W.: The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System, Science, 290, 291–296, doi:10.1126/science.290.5490.291, 2000.

Feely, R. A., Boutin, J., Cosca, C. E., Dandonneau, Y., Etcheto, J., Inoue, H. Y., Ishii, M., Le Quéré, C., Mackey, D. J., McPhaden, M. J., Metzl, N., Poisson, A., and Wanninkhof, R.: Seasonal and interannual variability of CO_2 in the equatorial Pacific, Deep-Sea Res. Pt. II, 49, 2443–2469, doi:10.1016/s0967-0645(02)00044-9, 2002.

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, 2009a.

Friedrich, T. and Oschlies, A.: Basin-scale pCO_2 maps estimated from ARGO float data: A model study, J. Geophys. Res., 114, C10012, doi:10.1029/2009jc005322, 2009b.

Gade, K.: A Non-singular Horizontal Position Representation, J. Navigation, 63, 395–417, doi:10.1017/S0373463309990415, 2010.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 3: Dissolved Oxygen Apparent Oxygen Utilization, and Oxygen Saturation, in: NOAA Atlas NESDIS 70, edited by: Levitus, S., US Government Printing Office, Washington DC, 344, 2010a.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Zweng, M. M., Baranova, O. K., and Johnson, D. R.: World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate), in: NOAA Atlas NESDIS 71, edited by: Levitus, S., US Government Printing Office, Washington DC, 398, 2010b.

George, M. D., Dileep Kumar, M., Naqvi, S. W. A., Banerjee, S., Narvekar, P. V., de Sousa, S. N., and Jayakumar, D. A.: A study of the carbon dioxide system in the northern Indian Ocean during premonsoon, Mar. Chem., 47, 243–254, doi:10.1016/0304-4203(94)90023-X, 1994.

Gibbs, M. T., Hobday, A. J., Sanderson, B., and Hewitt, C. L.: Defining the seaward extent of New Zealand's coastal zone, Estuar. Coast. Shelf Sci., 66, 240–254, doi:10.1016/j.ecss.2005.08.015, 2006.

Gruber, N. and Sarmiento, J. L.: Global patterns of marine nitrogen fixation and denitrification, Global Biogeochem. Cy., 11, 235–266, doi:10.1029/97gb00077, 1997.

Hsu, K., Gupta, H. V., Gao, X., Sorooshian, S., and Imam, B.: Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis, Water Resour. Res., 38, 1302, doi:10.1029/2001wr000795, 2002.

Jacobson, A. R., Mikaloff Fletcher, S. E., Gruber, N., Sarmiento, J. L., and Gloor, M.: A joint atmosphere-ocean inversion for surface fluxes of carbon dioxide: 1. Methods and global-scale fluxes, Global Biogeochem. Cy., 21, GB1019, doi:10.1029/2005gb002556, 2007.

Johnson, K. M., Sieburth, J. M., Williams, P. J. l., and Brändström, L.: Coulometric total carbon dioxide analysis for marine studies: Automation and calibration, Marine Chemistry, 21, 117–133, doi:10.1016/0304-4203(87)90033-8, 1987.

Joyce, T. and Corry, C. (Eds.), Requirements for WOCE Hydrographic Programme Data Reporting, 90-1 Rev. 2, WOCE Hydrographic Programme Office, La Jolla, California, 145 pp., 1994.

Key, R. M., Kozyr, A., Sabine, C. L., Lee, K., Wanninkhof, R., Bullister, J. L., Feely, R. A., Millero, F. J., Mordy, C., and Peng, T. H.: A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP), Global Biogeochem. Cy., 18, GB4031, doi:10.1029/2004gb002247, 2004.

Khatiwala, S., Primeau, F., and Hall, T.: Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean, Nature, 462, 346–349, doi:10.1038/nature08526, 2009.

Khatiwala, S., Tanhua, T., Mikaloff Fletcher, S., Gerber, M., Doney, S. C., Graven, H. D., Gruber, N., McKinley, G. A., Murata, A.,

Ríos, A. F., Sabine, C. L., and Sarmiento, J. L.: Global ocean storage of anthropogenic carbon, Biogeosciences Discuss., 9, 8931–8988, doi:10.5194/bgd-9-8931-2012, 2012.

Kirchman, D. L.: Processes in Microbial Ecology, Oxford University Press, 368 pp., 2012.

Kohonen, T.: Self-organization and associative memory, Springer-Verlag Berlin Heidelberg New York, Also Springer Series in Information Sciences, 8, 312 pp., 1988.

Le Quéré, C., Aumont, O., Bopp, L., Bousquet, P., Ciais, P., Francey, R., Heimann, M., Keeling, C. D., Keeling, R. F., Kheshgi, H., Peylin, P., Piper, S. C., Prentice, I. C., and Rayner, P. J.: Two decades of ocean CO_2 sink and variability, Tellus B, 55, 649–656, doi:10.1034/j.1600-0889.2003.00043.x, 2003.

Le Quéré, C., Takahashi, T., Buitenhuis, E. T., Rödenbeck, C., and Sutherland, S. C.: Impact of climate change and variability on the global oceanic sink of CO_2, Global Biogeochem. Cy., 24, GB4007, doi:10.1029/2009gb003599, 2010.

Lee, K., Wanninkhof, R., Feely, R. A., Millero, F. J., and Peng, T. H.: Global relationships of total inorganic carbon with temperature and nitrate in surface seawater, Global Biogeochem. Cy., 14, 979–994, doi:10.1029/1998GB001087, 2000.

Lee, K., Tong, L. T., Millero, F. J., Sabine, C. L., Dickson, A. G., Goyet, C., Park, G.-H., Wanninkhof, R., Feely, R. A., and Key, R. M.: Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans, Geophys. Res. Lett., 33, L19605, doi:10.1029/2006gl027207, 2006.

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, doi:10.1111/j.1600-0889.2005.00164.x, 2005.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 1: Temperature, in: NOAA Atlas NESDIS 68, edited by: Levitus, S., US Government Printing Office, Washington DC, 184, 2010.

Manning, A. C. and Keeling, R. F.: Global oceanic and land biotic carbon sinks from the Scripps atmospheric oxygen flask sampling network, Tellus B, 58, 95–116, doi:10.1111/j.1600-0889.2006.00175.x, 2006.

McKinley, G. A., Fay, A. R., Takahashi, T., and Metzl, N.: Convergence of atmospheric and North Atlantic carbon dioxide trends on multidecadal timescales, Nat. Geosci., 4, 606–610, doi:10.1038/ngeo1193, 2011.

McNeil, B. I.: Diagnosing coastal ocean CO_2 interannual variability from a 40 year hydrographic time series station off the east coast of Australia, Global Biogeochem. Cy., 24, GB4034, doi:10.1029/2010gb003870, 2010.

McNeil, B. I., Matear, R. J., Key, R. M., Bullister, J. L., and Sarmiento, J. L.: Anthropogenic CO_2 uptake by the ocean based on the global chlorofluorocarbon data set, Science, 299, 235, doi:10.1126/science.1077429, 2003.

McNeil, B. I., Metzl, N., Key, R. M., Matear, R. J., and Corbiere, A.: An empirical estimate of the Southern Ocean air-sea CO_2 flux, Global Biogeochem. Cy., 21, GB3011, doi:10.1029/2007gb002991, 2007.

Metzl, N., Brunet, C., Jabaud-Jan, A., Poisson, A., and Schauer, B.: Summer and winter air-sea CO_2 fluxes in the Southern Ocean, Deep-Sea Res. Pt. I, 53, 1548–1563, doi:10.1016/j.dsr.2006.07.006, 2006.

Mikaloff-Fletcher, S. E., Gruber, N., Jacobson, A. R., Doney, S. C., Dutkiewicz, S., Gerber, M., Follows, M., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Müller, S. A., and Sarmiento, J. L.: Inverse estimates of anthropogenic CO_2 uptake, transport, and storage by the ocean, Global Biogeochem. Cy., 20, GB2002, doi:10.1029/2005gb002530, 2006.

Millero, F. J., Lee, K., and Roche, M.: Distribution of alkalinity in the surface waters of the major oceans, Mar. Chem., 60, 111–130, doi:10.1016/s0304-4203(97)00084-4, 1998.

Park, G.-H., Wanninkhof, R. I. K., Doney, S. C., Takahashi, T., Lee, K., Feely, R. A., Sabine, C. L., Triñanes, J., and Lima, I. D.: Variability of global net sea–air CO_2 fluxes over the last three decades using empirical relationships, Tellus B, 62, 352–368, doi:10.1111/j.1600-0889.2010.00498.x, 2010.

Patra, P. K., Gurney, K. R., Denning, A. S., Maksyutov, S., Nakazawa, T., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fan, S., Fung, I., Gloor, M., Heimann, M., Higuchi, K., John, J., Law, R. M., Maki, T., Pak, B. C., Peylin, P., Prather, M., Rayner, P. J., Sarmiento, J. L., Taguchi, S., Takahashi, T., and Yuen, C.-W.: Sensitivity of inverse estimation of annual mean CO_2 sources and sinks to ocean-only sites versus all-sites observational networks, Geophys. Res. Lett., 33, L05814, doi:10.1029/2005gl025403, 2006.

Pöllä, M., Honkela, T., and Kohonen, T.: Bibliography of Self-Organizing Map (SOM) Papers, 2002–2005 Addendum, 236, 2009.

Rayner, P. J., Law, R. M., Allison, C. E., Francey, R. J., Trudinger, C. M., and Pickett-Heaps, C.: Interannual variability of the global carbon cycle (1992-2005) inferred by inversion of atmospheric CO_2 and $\delta^{13}CO_2$ measurements, Global Biogeochem. Cy., 22, GB3008, doi:10.1029/2007gb003068, 2008.

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., and Tilbrook, B.: The oceanic sink for anthropogenic CO_2, Science, 305, 367–371, doi:10.1126/science.1097403, 2004.

Sarmiento, J. L. and Gruber, N.: Ocean biogeochemical dynamics, Princeton University Press, 526 pp., 2006.

Sarmiento, J. L., Dunne, J., Gnanadesikan, A., Key, R. M., Matsumoto, K., and Slater, R.: A new estimate of the $CaCO_3$ to organic carbon export ratio, Global Biogeochem. Cy., 16, 1107, doi:10.1029/2002gb001919, 2002.

Sarmiento, J. L., Gloor, M., Gruber, N., Beaulieu, C., Jacobson, A. R., Mikaloff Fletcher, S. E., Pacala, S., and Rodgers, K.: Trends and regional distributions of land and ocean carbon sinks, Biogeosciences, 7, 2351–2367, doi:10.5194/bg-7-2351-2010, 2010.

Suzuki, T., Ishii, M., Aoyama, M., Christian, J. R., Enyo, K., Kawano, T., Key, R. M., Kosugi, N., Kozyr, A., Miller, L. A., Murata, A., Nakano, T., Ono, T., Saino, T., Sasaki, K., Sasano, D., Takatani, Y., Wakita, M., and Sabine, C.: PACIFICA Data Synthesis Project, ORNL/CDIAC-159, NDP-092. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, http://pacifica.pices.jp, doi:10.3334/CDIAC/OTG.PACIFICA_NDP092, 2013.

Takahashi, T., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G., Chavez, F., Sabine, C. L., Watson, A., Bakker, D. C. E., 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. G. J., 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, 554–577, doi:10.1016/j.dsr2.2008.12.009, 2009.

Takahashi, T., Sutherland, S. C., and Kozyr, A.: Global Ocean Surface Water Partial Pressure of CO_2 Database: Measurements Performed During 1957–2011 (Version 2011), Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, 2012.

Tanhua, T., van Heuven, S., Key, R. M., Velo, A., Olsen, A., and Schirnick, C.: Quality control procedures and methods of the CARINA database, Earth Syst. Sci. Data, 2, 35–49, doi:10.5194/essd-2-35-2010, 2010.

Telszewski, M., Chazottes, A., Schuster, U., Watson, A. J., Moulin, C., Bakker, D. C. E., González-Dávila, M., Johannessen, T., Körtzinger, A., Lüger, H., Olsen, A., Omar, A., Padin, X. A., Ríos, A. F., Steinhoff, T., Santana-Casiano, M., Wallace, D. W. R., and Wanninkhof, R.: Estimating the monthly pCO_2 distribution in the North Atlantic using a self-organizing neural network, Biogeosciences, 6, 1405–1421, doi:10.5194/bg-6-1405-2009, 2009.

Wallace, D. W. R.: Monitoring global ocean inventories, Dev. Panel Background Rep. 5, 54 pp, 1995.

Wehrens, R. and Buydens, L. M. C.: Self-and Super-organizing Maps in R: The kohonen Package, J. Stat. Softw., 21, 19 pp, 2007.

Weiss, R. F.: Carbon dioxide in water and seawater: the solubility of a non-ideal gas, Mar. Chem., 2, 203–215, doi:10.1016/0304-4203(74)90015-2, 1974.

Wickham, H.: ggplot2: elegant graphics for data analysis, Springer New York, 214 pp., 2009.