Manuscript prepared for J. Name
with version 2014/05/30 6.91 Copernicus papers of the LaTeX class copernicus.cls.
Date: 10 May 2015

Remote sensing the sea surface CO_2 of the Baltic Sea using the SOMLO methodology

Gaëlle Parard[1], Anastase A. Charantonis[2,3], and Anna Rutgerson[1]

[1]Department of Earth Sciences, Uppsala University, Uppsala, Sweden
[2]Centre d'études et de recherche en informatique, Conservatoire des Arts et Métiers, Paris, France
[3]Laboratoire d'océanographie et du climat : expérimentations et approches numériques, Université Pierre et Marie Curie, Paris, France

Correspondence to: gaelle.parard@geo.uu.se

Abstract. Studies of coastal seas in Europe have noted the high variability of the CO_2 system. This high variability, generated by the complex mechanisms driving the CO_2 fluxes, complicates the accurate estimation of these mechanisms. This is particularly pronounced in the Baltic Sea, where the mechanisms driving the fluxes have not been characterized in as much detail as in the open oceans. In addition, the joint availability of in situ measurements of CO_2 and of sea-surface satellite data is limited in the area. In this paper, we used the SOMLO (Sasse et al., 2013) methodology, which combines two existing methods (i.e., Self-Organizing-Maps and multiple linear regression) to estimate the ocean surface partial pressure of CO_2 (pCO_2) in the Baltic Sea from the remotely sensed sea surface temperature, chlorophyll, colored dissolved organic matter, net primary production, and mixed-layer depth. The outputs of this research have a horizontal resolution of 4 km and cover the 1998–2011 period. These outputs give a monthly map of the Baltic Sea on a very fine spatial resolution. The reconstructed pCO_2 values over the validation dataset have a correlation of 0.93 with the in situ measurements and a root mean square error of 36 μatm. Removing any of the satellite parameters degraded this reconstructed CO_2 flux, so we chose to supply any missing data using statistical imputation. The pCO_2 maps produced using this method also provide a confidence level of the reconstruction at each grid point. The results obtained are encouraging given the sparsity of available data, and we expect to be able to produce even more accurate reconstructions in coming years, given the predicted acquisition of new data.

1 Introduction

The ocean plays an important role in the global carbon budget. It acts as a major carbon sink for anthropogenic carbon dioxide (CO_2) emitted to the atmosphere from fossil fuel burning, cement production, biomass burning, deforestation, and various land use changes. The ocean is currently slowing the rate of climate change, having absorbed approximately 30% of human emissions of CO_2 to the atmosphere since the industrial revolution (Stocker et al., 2013). The exchange of CO_2 between coastal environments and the atmosphere is a significant part of the global carbon budget (e.g. Borges et al., 2005; Chen and Borges, 2009; Laruelle et al., 2010). While continental shelves represent only 7% of the oceanic surface area and less than 0.5 % of the ocean volume, the estimated overall sink of CO_2 in the continental shelf sea is –0.22 PgC yr^{-1} (Laruelle et al., 2010), corresponding to 16% of the open oceanic sink (Takahashi et al., 2009). These estimates are subject to great uncertainty related to sparse data coverage in time and space. Monitoring the oceanic partial pressure of CO_2 (pCO_2) at monthly and seasonal time scales is essential for estimating the regional and global air–sea CO_2 fluxes and reducing this uncertainty. For technical and budgetary reasons, in situ measurements of marine pCO_2 are sparsely distributed in time and space. However, over the last decade, technical improvements and cooperation with the shipping industry have allowed for the installation of several autonomous monitoring systems aboard commercial vessels routinely crossing the ocean basins. These instruments make quasi-continuous measurements, allowing regional analysis of the highly variable spatial and temporal distributions of pCO_2 (e.g. Lefèvre et al., 2004; Lüger et al., 2004; Corbière et al., 2007; Schneider et al., 2003). The Baltic Sea, a semi-enclosed sea in Northern Europe, is relatively well monitored and has been studied for several decades (Meier et al., 2014). Despite the increased number of measurements made in the Baltic Sea, assessing the carbon fluxes in the Baltic Sea remains particularly challenging due to the nonlinearity of the emission and absorption system. This nonlinearity is complicated by a combination of varying salinity, varying river input of DOC, and large general variability due to the strong seasonal cycle in the region. Using new methodologies could generate additional information from the relatively limited number of existing measurement data. Neural network techniques are empirical statistical tools that somewhat resolve the nonlinear and often discontinuous relationships among proxy parameters without any a priori assumptions. In the past decade, several authors have reported the application of a neural network technique to basin-scale pCO_2 sea analysis (Lefèvre et al., 2005; Jamet et al., 2007; Friedrich and Oschlies, 2009; Telszewski et al., 2009; Landschützer et al., 2013; Nakaoka et al., 2013; Schuster et al., 2013), concentrating mainly on the North Atlantic Ocean. Most recently, Telszewski et al. (2009) successfully applied a neural network technique based on a Self-Organizing Map (SOM) to reconstruct the seawater pCO_2 distribution in the North Atlantic (10.5° to 75.5°N, 9.5°E to 75.5°W) for three years (i.e., 2004 to 2006) by examining the nonlinear/discontinuous relationship between pCO_2 sea and the ocean parameters of sea surface temperature (SST), mixed-layer depth (MLD), and chlorophyll a concentration (Chl). In this paper, we applied the SOMLO methodology (self-organizing multiple linear output), which creates an SOM classification of the available explicative oceanic parameters in the Baltic Sea and then calculates multiple linear regression (MLR) parameters for estimating the pCO_2 from the elements belonging to each class separately. The major benefit of this methodology is that it allows the use of a linear model (i.e., the MLR) despite the nonlinear relationship between pCO_2 and its explicative parameters, by using the SOM classifier. The SOM classifier can determine the region of the multidimensional data space in which to perform the linear regression. If the classification is fine enough, each region will represent a single type of relationship between the pCO_2 and the explicative data, a type that can be reproduced by an MLR. Due to the temporal and spatial limitations of the in situ pCO_2 data, the satellite data can help estimate the air–sea CO_2 fluxes over the entire Baltic Sea. The satellite data have a higher spatial coverage of the Baltic Sea, allowing estimation of the pCO_2 from in situ data using the SOMLO method. Chierici et al. (2009) demonstrate that in the North Atlantic Ocean, SST, Chl, and MLD contributed significantly to the estimation of pCO_2 from a linear relationship. Based on this idea, we applied these parameters to the Baltic Sea, adding two other parameters, i.e., Net Primary Production (NPP) and Colored Dissolved Organic Matter (CDOM), that provide information about the biological activity occurring in summer. From this, we develop pCO_2 algorithms applicable to the Baltic Sea using in situ pCO_2 values; remotely sensed SST, Chl, and CDOM; modeled MLD and NPP; and time.

The manuscript is structured in four parts. First, we present a synopsis of the problem studied, including existing studies of pCO_2 reconstruction in other maritime regions. Next we present the available data and briefly describe the methodology used. In the third part of the article we present our results, namely, the topological maps obtained and the reconstructions performed with them. We conclude the article by discussing the results obtained and future possible improvements of the method used.

2 Materials and methods

2.1 Study area

The Baltic Sea is a semi-enclosed sea with limited exchange with the North Atlantic through the North Sea–Skagerrak system. Previous investigations of the Baltic Proper found large temporal and spatial variability of pCO_2. The amplitude of the annual pCO_2 cycle varies significantly depending on the region, ranging from 400 μatm in the northeastern Baltic Proper to 120 μatm in the transition areas to the North Sea (Schneider and Kaitala, 2006). The Baltic Sea receives significant river runoff from surrounding land (a total of approximately 15,000 m^3 s^{-1} (Bergstrom, 1994) and net precipitation of approximately 1500 m^3 s^{-1} (Omstedt et al., 2004). This large freshwater addition brings large amounts of nutrients and inorganic and organic carbon to the Baltic Sea basin (Omstedt et al., 2004; Hjalmarsson et al., 2008). The biogeochemical processes in the Baltic Sea marine environment are controlled mainly by the biological production and decomposition of organic matter occurring in

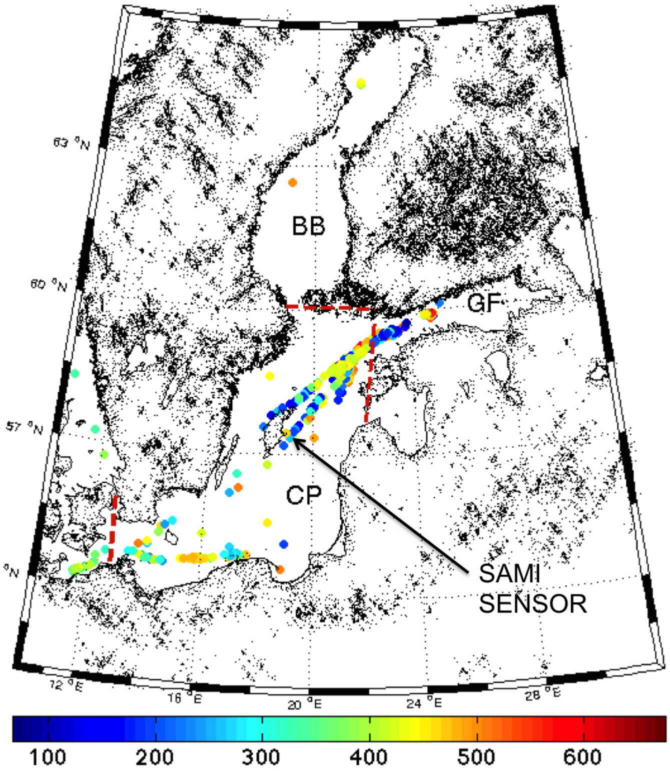

Figure 1. The dashed red lines divide the Baltic Sea into three basins: CP : Central part, BB: Gulf of Bothnia, GF: Gulf of Finland. Monthly data are available from 1998 to 2011 in the Baltic Sea. The colorbar shows the pCO_2 values in μatm. The black arrow show the position of the SAMI sensor.

the context of the region's hydrography (Siegel and Gerth, 2012). Physical forcing controls the water transport, stratification, temperature, and salinity in the Baltic Sea; these factors then influence the nutrient and carbon distribution, thereby affecting biogeochemical processes. We divide the Baltic Sea into three basins, i.e., the central part (CP), the Gulf of Bothnia (GB), and Gulf of Finland (GF), as shown in Figure 1. The Baltic Sea has an average depth of 55 m and a maximum depth of 460 m at the Landsort Deep (Wesslander, 2011).

2.2 pCO_2 observations

To compile the pCO_2 maps, we use measured data from three sources.

1. The Östergarnsholm site: This site is located next to the small island of Östergarnsholm in the central Baltic Sea and is further described by (Rutgersson et al., 2008; Norman et al., 2013a). The island is situated 4 km from the east coast of the larger island of Gotland. SST and pCO_2 are measured semi-continuously 4 m below the sea surface using a submersible autonomous moored instrument (SAMI) CO_2 sensor moored at a buoy 1 km southeast of the tower situate on the island. In addition, SST is also measured using a wave rider buoy (operated by the

Finnish Meteorological Institute) at 0.5 m depth situated approximately 4 km southeast of the tower.

2. Cargo ship: This dataset derives from continuous measurements of the surface water pCO_2 made in the Baltic Sea using a fully automated measurement system deployed on a cargo ship. The Leibniz Institute for Baltic Sea Research, Warnemünde Germany (IOW—Institut für Ostseeforschung Warnemünde) has made continuous measurements of pCO_2 at 5 m depth aboard the cargo vessel Finnpartner. This ship crosses between Lübeck and Helsinki at a two-day interval, alternately crossing the eastern and western Gotland Sea (Schneider and Kaitala, 2006; Schneider et al., 2009). Data from Finnpartner were acquired between July 2003 and December 2005.

3. Swedish Meteorological and Hydrological Institute (SMHI) database Svenskt Havsarkiv (SHARK): pH, (measured using the method of Grasshoff et al. (1999)) and Total Alkalinity (TA) (measured using potentiometric titration as described by Grasshoff et al. (1999)) are measured continuously at a monthly or semi-monthly resolution in the Baltic Sea at various stations. All measurements are made at a depth of 5 m depth. The uncertainty of the pH is ± 0.03 pH units and of the TA is $\pm 5\%$ (Wesslander et al., 2009). pCO_2 is estimated from the pH, TA, salinity, and temperature measurements using the standard CO2SYS program (Lewis and Wallace, 1998) with the equilibrium constant from (Weiss, 1974) and (Merbach et al., 1973) as refitted by Dickson and Millero (1987) as in Wesslander et al. (2009). The pCO_2 estimation were compared with the other source of pCO_2 data, all the data were compared in time (less than 1 hour) and position (arround less than 0.2°), the correlation coefficient (R) give 0.98 and the standart deviation (STD) is 9 μatm. The high standard deviation is explain by the presence of upwelling with high pCO_2 present near the coast on the SAMI sensor measurement, when the we remove the uppwelling event the STD is 1.5 μatm.

2.3 Remote sensing data

The satellite data used in this study are from various sources. We use a monthly temporal resolution and a spatial resolution based on the lowest spatial resolution of our datasets, i.e., based on the lower spatial resolution CDOM dataset.

We obtain values for five parameters from various sources:

SST Sea Surface Temperature: Several datasets are used for SST, and we combine two types of datasets for 2007 and 2011. For 2005–2011, we use data from the Federal Maritime and Hydrographic Agency (BSH), which processed data from AVHRR-NOAA, and data from the Group for High-Resolution Sea Surface Temperature (GRHSST) dataset for the Baltic Sea, 2007–2011. The spatial resolution is 0.03 ° at a daily temporal resolution (http://podaac.jpl.nasa.gov/dataset/DMI-L4UHfnd-NSEABALTIC-DMI_OI). For 1998–2004, the data come from a reanalysis of the NOAA/NASA Advanced Very High Resolution Radiometer (AVHRR) data stream conducted by the University of Miami's Rosenstiel School of Marine and Atmospheric Science (RSMAS) and the NOAA National Oceanographic Data Center (NODC). This dataset consists of the monthly average SST (in °C) over the zone, with a spatial resolution of 4 km, extracted from version 5.2 of the AVHRR Pathfinder project (Casey et al., 2010)(http://www.nodc.noaa.gov/SatelliteData/pathfinder4km/). The various SST datasets were compared with measured SMHI temperature data at a monthly resolution from 1998 to 2011, giving a correlation coefficient (R) of 0.99 and a mean difference (MD) of $0.05°C$ between the two datasets. The difference observed between the measured and satellite data between 1998 and 2004 give a value (R = 0.99 and MD = $0.09°C$) near the difference between 2005 and 2011 (R = 0.99 and MD = 0.14). The two SST datasets used between 2007 and 2011 were also compared with the SAMI sensor data at a daily resolution, giving a good correlation for the BSH data (R = 0.95 and MD = $0.06°C$) and the GRHSST data (R = 0.95 and MD = $0.08°C$).

Chl Chlorophyll a: This dataset consists of monthly averages from the following sensors: Sea-viewing Wide Field-of-view Sensor (SeaWiFS) (Sept. 1998–Dec. 2002) with 4-km spatial and monthly temporal resolutions and Moderate Resolution Imaging Spectroradiometer (MODIS-Aqua) (Jul. 2002–Jun. 2011) with 4-km spatial and monthly temporal resolutions (Casey et al., 2010). A lognormal distribution was assumed for the Chl data. Comparison with SMHI measurement data and in situ data (personal communication from Dr. Tiit Kutser) give R = 0.67 and MD = 7 mg m^{-3}. The chlorophyll levels from the satellite data seem to be overestimated compared with the in situ data. Since we use the same dataset over the whole study period, this bias was learned during the classification process and and the MLR parameters are calculated considering that any further input will include that bias.

CDOM Colored Dissolved Organic Matter values come from MODIS-Aqua 4-km monthly average data. The CDOM index quantifies the deviation in the relationship between the CDOM and Chl concentrations, where 1.0 represents the mean relationship for Morel and Gentili (2009) case 1 waters, and values above or below 1.0 indicate an excess or deficit, respectively, in CDOM relative to the mean relationship. The algorithm and its application are fully described by Morel and Gentili (2009). In situ CDOM data (personal communication from Dr. Tiit Kutser) give a lower correlation coefficient and a low average difference (R = 0.48, MD = 2.3). As for the chlorophyll, the bias is applied to all years so it does not affect the estimation of pCO_2.

NPP : Net Primary Production values come from two data sources. The first dataset comes from the Environmental Marine Information System (EMIS): The EMIS model is depth integrated but allows for depth-dependent variability in the diffuse attenuation coefficient, which is calculated from a multiple-component semi-analytical inversion algorithm (Lee et al., 2005). The primary production calculation is based on the formulation obtained through dimensional analysis by Platt and Sathyendranath (1993). The photosynthetic parameters are assigned by the combined use of a temperature-dependent relationship for the maximum growth rate Eppley (1972) and a variable formulation to retrieve the C:Chl ratio following the empirical relationship of Cloern et al. (1995). The EMIS dataset comprises monthly average values from October 1997 to September 2008. The second dataset, for 2009–2011, uses the Vertically Generalized Production Model (VGPM) of Behrenfeld and Boss (2006) as the standard algorithm. VGPM is a "chlorophyll-based" model that estimates net primary production from chlorophyll using a temperature-dependent description of chlorophyll-specific photosynthetic efficiency. For VGPM, net primary production is a function of chlorophyll, available light, and photosynthetic efficiency. VGPM uses MODIS-Aqua chlorophyll and temperature data, SeaWiFS photosynthetically active radiation (PAR) data, and estimates of the euphotic zone depth from a model developed by Morel and Berthon (1989) and based on chlorophyll concentrations. For the NPP for 2009–2011, the observed maximum value was limited to 10 to be comparable to the data for 1998–2008. Validation of NPP was difficult due to the number of data available in the area. Comparison of the two datasets gives similar values and seasonal cycles. We compared the seasonal cycles between 1998–2008 and 2009–2011, obtaining values on the same order of magnitude.

MLD Mixed- Layer Depth: There are also two sources for the MLD data. Monthly averages from 1998 to 2007 come from a 3D hydrodynamic model currently used at the the Joint Research Centre/Institute for Environment and Sustainability (JRC/IES), i.e., the public-domain General Estuarine Transport Model (GETM; www.getm.eu), which has its roots in developments at the JRC/IES Burchard and Bolding (2002). GETM simulates the most important hydrodynamic and thermodynamic processes in coastal and marine waters and includes flexible vertical and horizontal coordinate systems. Different turbulence schemes are incorporated from the general ocean turbulence model (GOTM; www.gotm.net). For 2008–2011, we use data from the carbon-based production model at a monthly resolution (Behrenfeld et al., 2005). The MLD was estimated from SMHI temperature and salinity profile measurements using the density criterion of (Boyer Montégut et al., 2004); the comparison between the model estimation and estimated SMHI MLD is good (R = 0.63 and MD = 17 m). Between 1998 and 2007, the correlation coefficient is higher (R = 0.8) than between 2008 and 2011 (R = 0.5). Nevertheless, the MD is lower for the second data source, i.e., 20 m versus 11 m. This can be explained by the available data coverage: 63% of the in situ data cover the 1998–2004 period. From 2008 to 2011, the maximal value is below 80 m between 1998 and 2007, so for the in situ data estimation, we replace every value above 80 m with "not a number."

In the Baltic Sea, there are many gaps in the satellite data; this is due to the high proportion of coastal waters where satellite data are less reliable, and to the frequent large-scale cloud coverage. To increase the total number of our data, we used a monthly temporal resolution and a method to improve the spatial distribution of the data. For statistical analysis, the irregularly spaced density of the measurements were first uniformly resampled. To this end, Gaussian grinding was used, as described by Greengard and Lee (2004); Dutt and Rokhlin (1995). The data points of the original series are convolved using a Gaussian kernel. As a result, the data points are smeared over their neighboring equi-spaced points, which are more densely distributed. This method produces more realistic values than does simple interpolation, particularly when there are many data gaps (Schomberg and Timmer, 1995). There is no discontinuity between the different datasets , but NPP, CDOM, and Chl data are missing for January and December, so no reconstruction can be performed for these months.

2.4 Data available

The positions and values of all the in situ pCO_2 data are shown in Figure 1. We use the spatial resolution of the parameter with the lowest resolution for the final dataset chosen (i.e., CDOM). A monthly temporal resolution is used for this study. Rutgersson et al. (2008) demonstrate that the agreement between SAMI sensor data and the ship data from Finnpartner (near the mooring maximum of 23 km) is quite good. This good agreement is confirmed by the comparison between pCO_2 data from the SAMI sensor, the data surrounding the mooring (0.2°), and the other datasets. These analyses give a good correlation factor of 0.98. The in situ data are available mainly for the central basin, but the number of data for the Gulf of Bothnia is very low, coming from two SMHI stations. The in situ pCO_2 data are well distributed over the twelve months (Figure 2). January is the month for which the number of data is lower (i.e., below 80), but the other months have 110–155 data points each. In our case, each in situ data point is characterized by SST, Chl, CDOM, NPP, and MLD as well as information on the date the measurements were made. This temporal information was normalized by sine and cosine, as follows:

$$time(cosine) = cos(\frac{Days * 2\pi}{365}) \tag{1}$$

$$time(sine) = sin(\frac{Days * 2\pi}{365}) \tag{2}$$

where Days represents the Julian day.

This definition of time is used to render the values continuous over the course of the year, sidestepping the artificial numerical transition from the last day of one year to the first day of the next, to be able to situate the process in relation to its seasonality.

Although time itself is not affecting the pCO_2, the inclusion of time (time(cosine) and time(sine)) as a parameter is important since, in the database, some situations have similar values for SST, Chl, CDOM, NPP, and MLD, but different pCO_2 values. For example, in May we had a situation (SST = 9.3 °C, Chl = 0.1 mg m^{-3}, CDOM = 4.3, NNP = 1.7 mg C m^{-2} d^{-1}, and MLD = 10.8 m) whose parameters were very close to those of a situation in November (SST = 9.1°C, Chl = 0.1 mg m^{-3}, CDOM = 4.3, NNP = 1.6 g C m^{-2} d^{-1}, and MLD = 10.3 m); however, the pCO_2 values differed greatly, being pCO_2 = 214 and 444 μatm, respectively, during the two situations. This dissimilarity

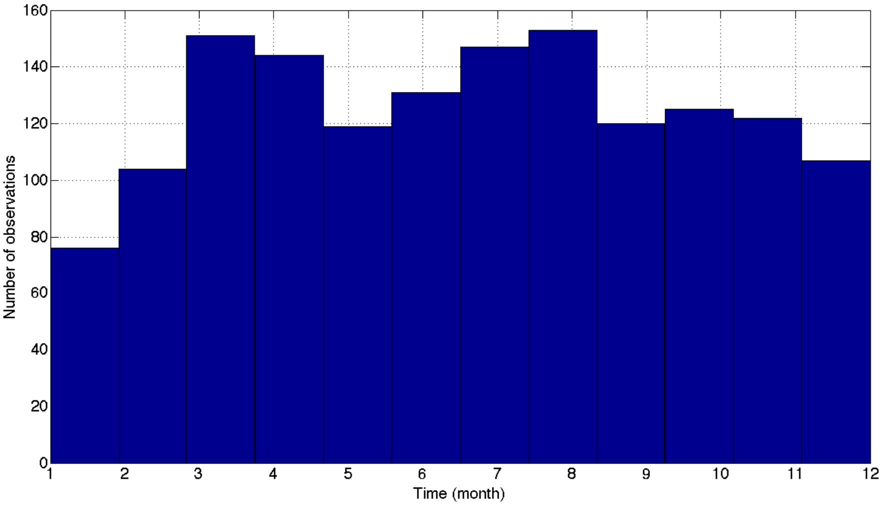

Figure 2. Histogram showing the number of observations in function of the month between 1998 to 2011.

in the pCO_2 values informs us that these situations are generated by other drivers than SST, Chl, CDOM, NPP, and MLD. Since these drivers can be related to seasonal patterns, we included an information on the time of the year,as a proxy that allows us to fine tune our classification. The inclusion of two temporal values instead of only one, even though those are highly correlated, was unavoidable in order to preserve continuity in the values obtained when changing years.

A Principal Component Analysis (PCA) was conducted to highlight the importance of the parameters in the pCO_2 variability (Figure 3). PCA is a method of analysis which involves finding the linear combination of a set of variables that has maximum variance and removing its effect, repeating this successively (Jolliffe, 2002). The percent of variance explained by each axis of the PCA is shown in Table 1. The results of this PCA indicate that the percentages of variance explained by all axes beyond the first do not differ greatly, with the notable exception of the last two axes, indicating that most parameters can be discriminant in the definition of the SOM states. The need to maintain the totality of these parameters is further demonstrated by the projection of the explicative parameters in the correlation circle, where all values are close to the boundaries of the correlation circle. This informs us that they are all tied to the phenomena explaining 62% of the total variation of our data. While we could reduce the dimension of the problem by using only the projection of these variables on the first few axes, we chose to maintain all the explicative parameters presented when applying SOMLO to estimate the pCO_2 in order to be able to not loose any information when performing the SOM classifications.

In total, 1445 pCO_2 data are used in this study, after having removed the outliers from our dataset. These 170 oultiers were beyond three standard deviations away of at least one of the explicative parameters. All parameters (i.e., SST, Chl, CDOM, NPP, and MLD) are located around each pCO_2

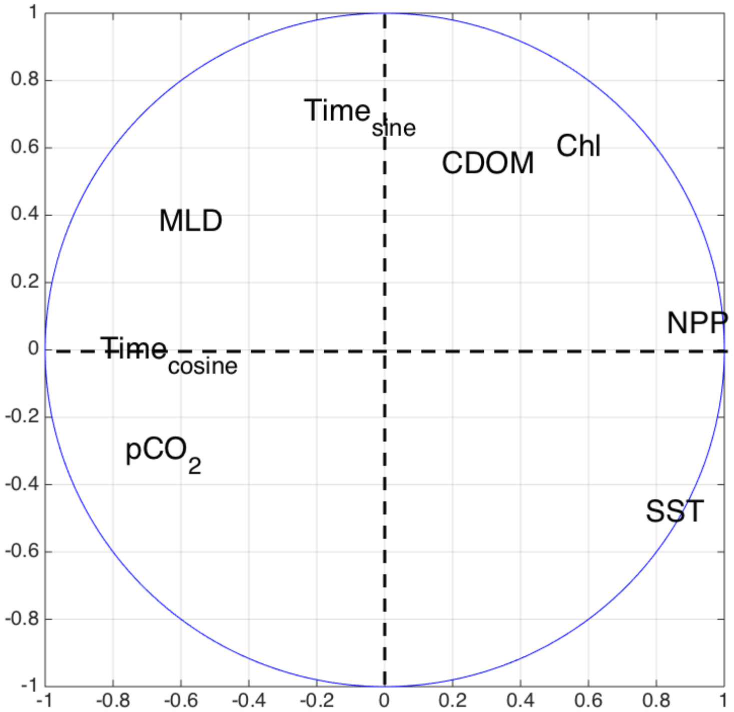

Figure 3. Correlation scatterplot with the representation of parameters on axes 1 and 2. The first component plane contains 62% of the total variance of the system under study, and all parameters are close to the exterior of the circle, indicating that they are all important in order to invert the pCO2.

Table 1. Percent of variance explained by each axis of the PCA.

Axis number	1	2	3	4	5	6	7	8
Percent of variance	42	20	14	9	8	4	2.5	0.5

datum. In winter (i.e., October to March), more data are missing (Table 2, column 1), particularly for Chl, CDOM, and NPP, winter being the period when it is more difficult to measure or estimate these parameters. Between April and September, the number of missing SST, Chl, CDOM, and MLD data is relatively low compared with the total number of data (Table 2, column 2), i.e., fewer than 3% of 275 the total. To increase the number of data available, we completed the data by training the topological map. Further details on this are presented in section 2.5.

2.5 Methodology

The relationship between pCO_2 and the environmental parameters is highly nonlinear: a slight variation in some of the environmental parameters could correspond to significant variations in pCO_2. We 280 chose to use the SOMLO methodology, which combines two statistical approaches: *self-organizing maps* (SOMs) (Kohonen, 1990) and ***linear regression***. SOMs are a subfamily of neural network

Table 2. Number of missing values for each parameter for the satellite data for the October–March and April–September periods. The numbers in parentheses indicate the total data points in each period.

Parameter	October–March (685)	April–September (814)
SST	28	0
Chl	202	24
CDOM	320	5
NPP	468	571
MLD	6	2

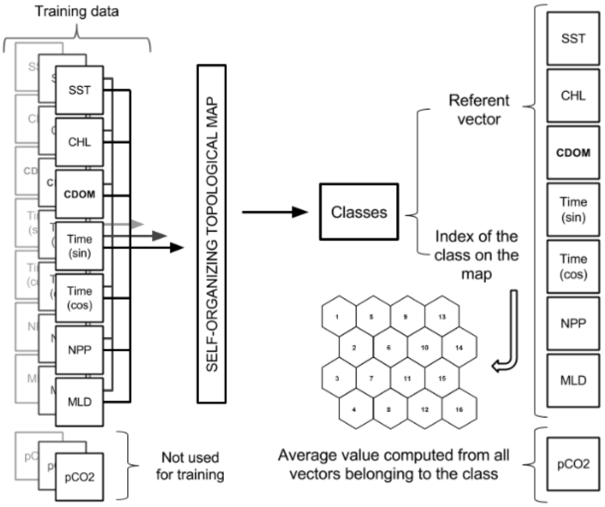

Figure 4. The various elements used in training a self-organizing map. On the left we have the dataset used to train the SOM, which discretizes it into classes. Each class contains a referent vector containing, for each parameter, the average value of the elements comprising it, and an index of the class that informs us of its location on the topological map.

algorithms used to perform multidimensional classification. A defining characteristic of SOMs is that their classes can represent a Gaussian distribution centered around the typical profile of environmental parameters, if there is high discretization of the training dataset (Dreyfus, 2005). We use this hypothesis to classify the environmental parameter dataset, and then estimate the parameters of a linear regression for each class. In the following section, we present a brief overview of the two statistical algorithms and their application to our datasets.

2.5.1 Self-organizing maps

Self-organizing topological maps (SOMs) are a clustering method based on neural networks. They cluster a learning dataset into a reduced number of subsets, called classes, with common statistical characteristics.

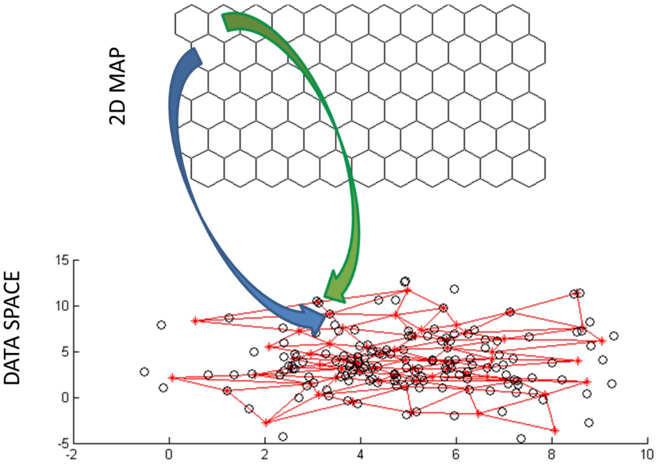

Figure 5. On top we have a representation of the topological map as a lattice, while on the bottom we have a projection of data in the data space (black circles), as well as the average values of each class (red circles). Adjacent classes on the SOM lattice correspond to adjacent areas in the multidimensional data space. The red lines indicate the connections between neighboring classes.

Generating a SOM requires the creation of a training database that contains homogenous vectors. After a training phase, we obtain a SOM. The term "map" corresponds to a 2D matrix that stores, for each class, its referent vector, r_i, which approximates the mean value of the elements belonging to it, and its index, which positions it in the map matrix used to situate it in relation to the other classes (Figure 4).

SOMs are also called self-organizing topological maps, “topological” indicating that the SOM training algorithm forces a topological ordering of the classes in the matrix, meaning that any two neighboring classes C_i and C_j on the map matrix have referent vectors r_i and r_j that are close in the Euclidean sense in the data space.

Let us consider a vector x that is of the same dimensions and nature as the data used to generate the topological map; we can find the index of the class to which it is classified by choosing: index = $\arg\max_i$ (||x – r_i||), (where $argmax$ is the argument maximal) therefore assigning it to the class whose referent is closest to it in the Euclidean sense (Figure 5). A classified vector x will be represented by its class index, C_{index}. If we are trying to classify a vector that has some missing values, the comparison is performed between the existing values of x and the corresponding values of each r_i.

As a version of the expectation–maximization algorithm, the SOM algorithm performs an iterative training. During the early phases of this training, the referent vectors of each class are strongly affected by the changes imparted on their neighbors’ referent vectors in order to capture the shape of the data cloud. Depending on the training parameters of the SOM, in the latter phases of the training, the effect of the neighboring vectors on the determination of the referent vector can be considered null. In these cases, each referent vector approximates, locally, the mean value of the multidimensional Gaussian random distribution that generated the training data assigned to that class (Dreyfus, 2005).

2.5.2 Multiple linear regression

Multiple linear regression (MLR) is a modeling method that expresses the value of one response variable, V (in our study pCO_2), as a linear function of other explicative variables, i.e., X = X_1, X_2, ... , X_i (in our study SST, Chl, CDOM, NPP, MLD, $time_{sin}$, and $time_{cos}$). An MLR is generally performed either to interpret the relationship between the variable y and each of the other predictive variables X_i, or to predict, from a dataset of vectors containing the values of X, the corresponding value of y. In this paper, we used both aspects of MLR.

However, to perform MLRs in the present case, we had to take into account their limitations and the nature of the problem. Specifically, to perform an MLR we are obliged to assume that the relationship between the predictor variables and the response variable is *linear*. However, this is not the case in our datasets: pCO_2 is not linearly related to the variables presented when considering the entirety of the problem. However, as noted above in subsection 2.5.1, if we consider the classes created by the SOM, they are very localized regions of the combined explicative and response data space that can be considered to approximate, locally, the mean value of a multidimensional Gaussian random distribution. We therefore assume that, if performed in the reduced neighborhood of a SOM class, the relationships between pCO_2 and the explicative variables are linear.

3 Application and results

3.1 Statistical imputation

As described in section 2.4, both the satellite and measured data available for the application present missing values. To complete these datasets, we chose to use imputation methods similar to those described by Schafer and Graham (2002) and Malek et al. (2008). The main idea of these methods is to use the classifying abilities of the SOMs to regroup the data in typical situations and replace the missing explicative data values with the corresponding values of the referent vector of the class to which it belongs.

We first selected the database containing SST, Chl, CDOM, NPP, MLD, $time_{sine}$, and $time_{cosine}$. The vectors were sorted according to the number of values missing from each vector and noting the locations of these missing values. We chose all complete data vectors and the first 5% of the sorted vectors containing missing data and trained a SOM. We proceeded by replacing the missing values of these first 5% of vectors with the corresponding values of the referent vector of the class to

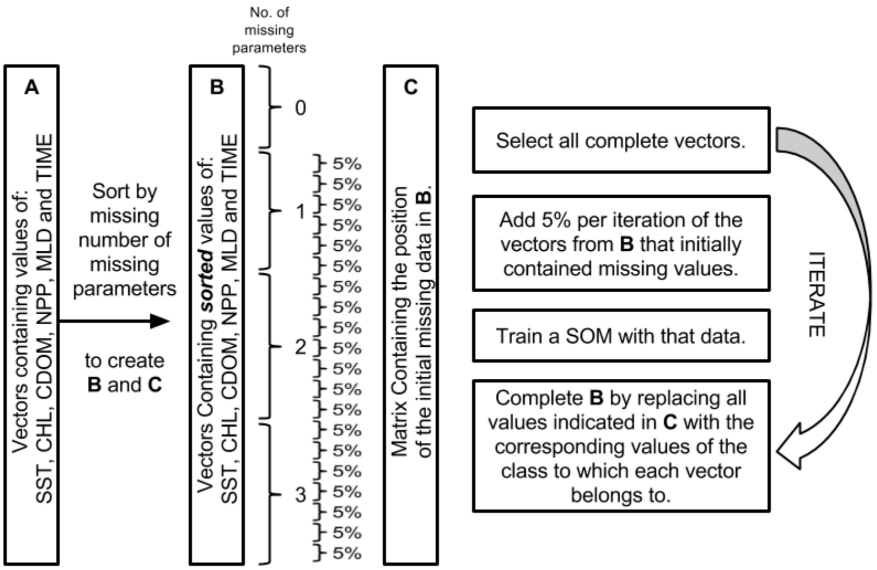
No. of missing parameters
A
Vectors containing values of: SST, CHL, CDOM, NPP, MLD and TIME
Sort by missing number of missing parameters
to create B and C
B
Vectors Containing sorted values of: SST, CHL, CDOM, NPP, MLD and TIME
0
1
2
3
5%
C
Matrix Containing the position of the initial missing data in B.
Select all complete vectors.
Add 5% per iteration of the vectors from B that initially contained missing values.
Train a SOM with that data.
Complete B by replacing all values indicated in C with the corresponding values of the class to which each vector belongs to.
ITERATE

Figure 6. A schematic of the imputation method used. We initially sort the data depending on the amount of missing values present, then progressively train SOMs on the dataset, steadily including more vectors for the training and completing and updating the missing data during each iteration.

which they each belong. We then included the next 5% of vectors with missing data in a new training dataset and created a new SOM. Based on this new SOM, we again filled the new missing values with the corresponding values of the referent vector of the class of the new SOM to which they each belong. In addition, we deleted the values we added to the first 5% of vectors and replaced them with the values of the referent vector of the class of the new SOM to which they each belong. The training parameters of this method, such as the number of classes of the topological map and the number of iterations were selected by parameter tuning. We then continued iterating this process, updating the previously filled missing values with the values of their corresponding referents belonging to the most recently trained SOM, until all missing values were filled. The updating of the previously filled values allows the method to progressively incorporate information and to rectify A schematic of the imputation method used can be seen in Figure 6.

After this imputation of the missing data through iterative training, the reconstructed data represent the original data well. Figures 7 and 8 show data for six variables before and after the reconstruction, respectively. The main difference is observed for the values of Chl, where the peak of over 200 individuals occurs at 0 mg m^{-3} because, at the initialization of the imputation process, we decided to replace all null values. The repartition of pCO$_2$ (Figure 8a) is very representative of the data variability with a large range of values. Some very high values occur during local events, such as coastal upwelling. Most of the data range between 180 μatm (value observed in summer) and 550 μatm (observed in winter). The SST (Figure 8b) is very representative of the variability in the Baltic

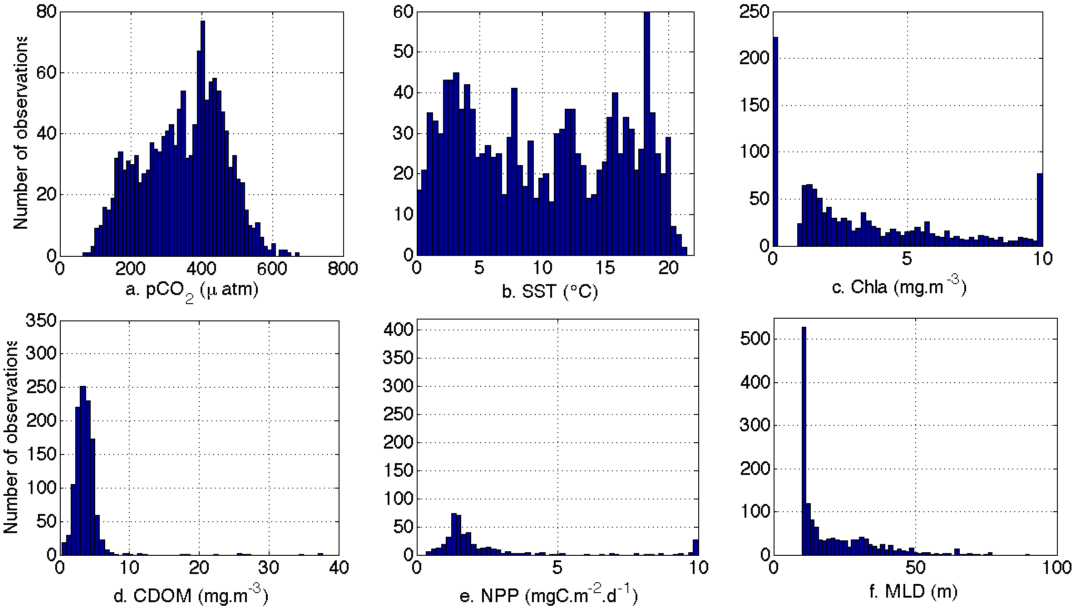

Figure 7. Histogram of a. pCO$_2$ and satellite data, b. SST, c. Chl, d. NPP, e. CDOM, and f. MLD available for the SOM before reconstruction. Y-axis represent the number of data and the X-axis the value of the parameters.

Sea with a maximum occurring between July and September in all basins around 18°C (Siegel and Gerth, 2012). The NPP variability is fairly homogenous, except for the peak at 10 mg C m^{-2} d^{-1}. This peak occurs because the first model providing us the NPP values has a set maximum of NPP at 10 mg C m^{-2} d^{-1}; therefore, the correction of the NPP satellite data takes this maximum into account.

The variability of chlorophyll results in one subset of the data with a low value and another subset with a value higher than 6 mg m^{-3}, which can be explained by the fact that the Baltic Sea is a narrow sea, with important coastal regions, and that two blooms take place, in spring and in summer. The chlorophyll value can be very high during these periods, and the reconstruction gives a mean value for this characteristic. A peak at 10 mg m^{-3} is observed in the chlorophyll data, not due to the reconstruction but to the maximum value in the satellite data file. The low MLD occurs in summer, and in the model the minimum is 10 m deep, which appears to be around the minimum value observed in Figure 8f. Absorption by CDOM decreases with increased distance from the riverine sources, reaching a relatively stable absorption background in the open sea. Most of our CDOM data capture open sea conditions, so the values are quite low.

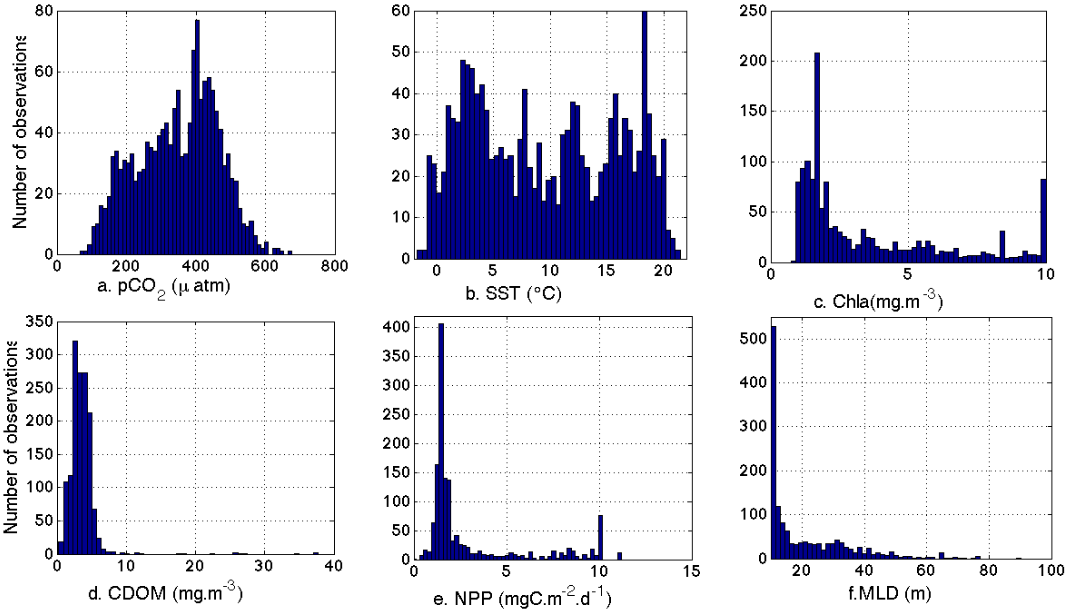

Figure 8. Histogram of a. pCO$_2$ and satellite data, b. SST, c. Chl, d. NPP, e. CDOM, and f. MLD available for the SOM after reconstruction. Y-axis represent the number of data and the X-axis the value of the parameters.

3.2 pCO$_2$ estimation

3.2.1 Topological map

We classified the explicative variables (i.e., SST, Chl, CDOM, NPP, MLD, time$_{sin}$, and time$_{cos}$) into classes that share similar characteristics. In order to optimize the SOM map size for the method and to calculate the method's performances we randomly sampled 90% of our completed dataset (1300 vectors) to be used for the training phase, keeping 10% (144 vectors) to be used for computing the performances of the method. We iterated this process, selecting many different random samplings for each map size, and selected the map size with the best average reconstruction when applying the SOMLO methodology.

At the end of our optimization, we selected a SOM consisting of 77 classes. The number of observations captured by each class ranges from 0 to 38 (Figure 9). The order of magnitude of the number of observations is constant throughout the SOM, and we can regard the classes as having spread in multidimensional space in order to accurately represent the data space of the explanatory parameters. The presence of classes that did not capture any elements can be justified as preventive: they preserve the topological aspect of the SOM by preventing classes that are not similar enough from becoming neighbors.

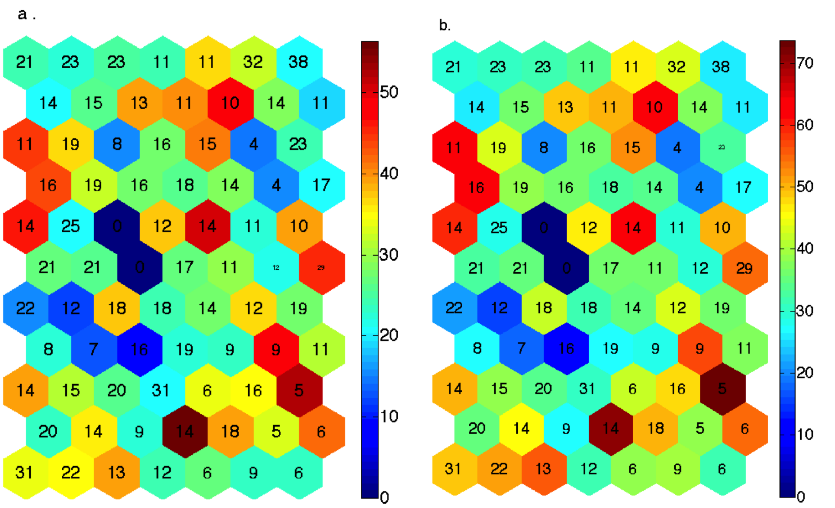

Figure 9. a. Distribution of the neural map of a.the average difference in pCO_2 for each neuron represents by the colorbar. b. the standard deviation (STD) in pCO_2 for each neuron represents by the colorbar. For a. and b., the numbers inside the neurons correspond to the number of data ton construct the characteristics of the neuron.

To estimate the average concentration of pCO_2 in each class, the measurements of pCO_2 associated with vectors consisting of SST, CDOM, NPP, MLD, and CHL components were presented to the already trained SOM as input data (Figure 10). The average value computed for the vectors belonging to each class corresponds to the average value of pCO_2 for that class.

In the final map, the distribution of pCO_2 is strongly dependent on the SST distribution, with low values of pCO_2 correlating with high values of SST (Figure 10). This is in agreement with the seasonal pCO_2 cycle, which is characterized by a large amplitude, ranging from a high value in winter ($\approx$500 μatm) and a low value in summer ($\approx$150 μatm), described, for example, by Wesslander (2011). According to Schneider and Kaitala (2006), the high winter value of pCO_2 is a consequence of mixing with a deeper water layer enriched in CO_2, which is in agreement with the distribution of the MLD (Figure 10,h), with the higher value in winter and autumn correlating with the high value of pCO_2. High values can also be explained by the mineralization, which exceeds production in winter (Wesslander, 2011). Biological production starts in spring when sunlight and nutrients are sufficient. The chlorophyll begins to increase in March–April due to the spring phytoplankton bloom, which reduces the pCO_2 level during this period. The more intensive decrease occurs in April and May, which is consistent with the higher value of NPP (Figure 10). Studies in the central Baltic Sea identify two summer minima, the first in April/May and the second in July/August, resulting from a second production period. Higher variability is observed during this period, with a standard deviation between 39 μatm and 50 μatm for different regions (Wesslander, 2011; Schneider and Kaitala, 2006).

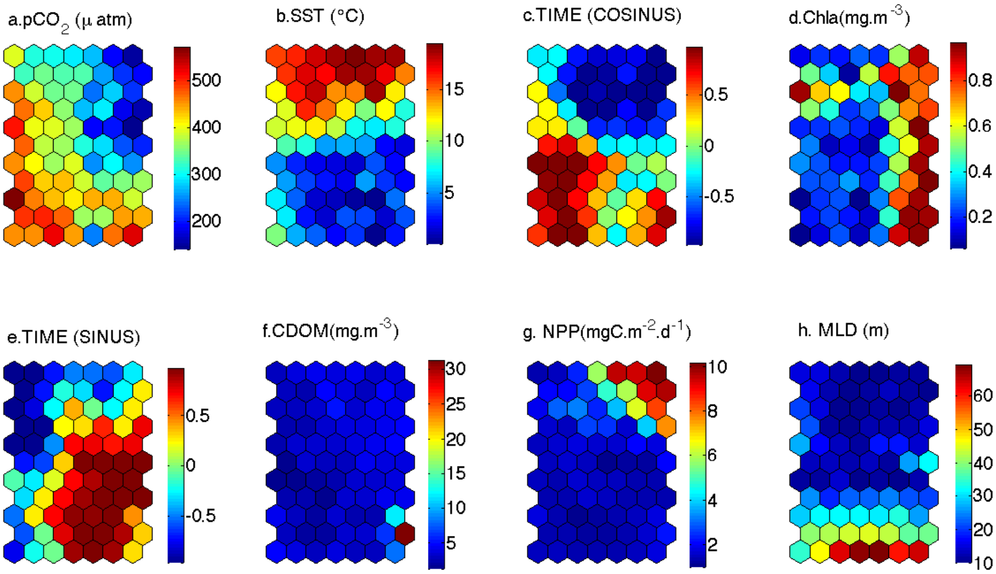

Figure 10. Distribution of each parameter in the neural map: a. pCO_2 in μatm, b. SST in °C, c. and e. time cosine and sine, respectively, d. Chl in mg m^{-3}, g. NPP in mg C m^{-2}, and h. MLD in m.

3.2.2 Linear regression in the neurons

To perform an MLR, we must assume that the relationship between the predictor variables and the response variable is *linear*. We could take this to be a valid hypothesis only when performing the MLR in the reduced neighborhood of a SOM class, where the relationship between pCO_2 and the explicative variables can be assumed to be linear.

For each class j a separate training dataset was created containing all the vectors assigned to that class and to all its adjacent classes. Based on that dataset, we computed the linear regression coefficient parameters for every explicative parameter and for a constant value.

The calculated linear regression coefficient parameter values for each class are shown in Figure 11. Note that all parameters are important in specific regions of the SOM, having both positive and negative correlations in different classes.

More importantly, the fact that each parameter has a significantly varying range of values over the different classes demonstrates that each parameter is important in reconstructing the pCO_2 in the Baltic Sea, even though a parameter may be highly significant in some classes and relatively stable in other regions of the topological map.

The addition of vectors belonging to adjacent classes did not generally perturb the estimation of the coefficient parameters because, as seen in Figure 10, the values of all parameters are generally organized coherently on the map. The assumption that they are close in the data space is not as

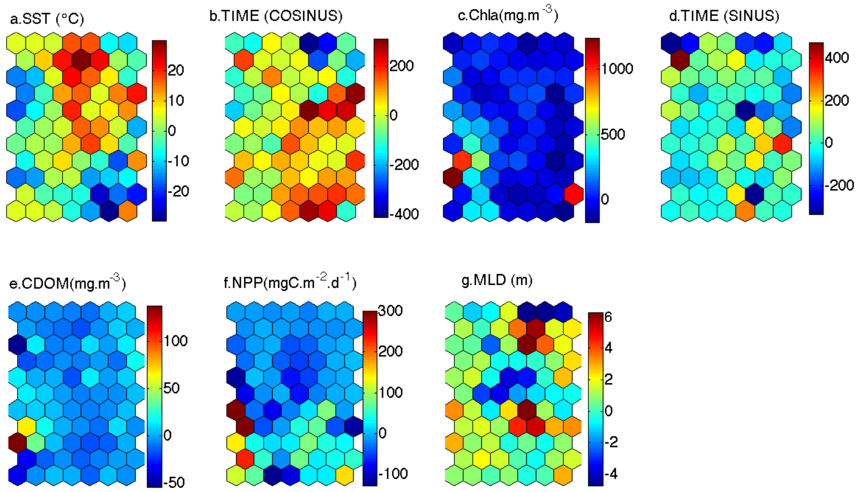

Figure 11. Distribution of the neural map of the coefficient from linear regression for each parameter: a. SST in °C, b. and d. time cosine and sine, respectively, c. Chl in mg m^{-3}, f. NPP in mg C m^{-2}, and g. MLD in m.

robust as it would have been had we solely considered the vectors belonging to each class but, given the limited number of data available for modeling this highly nonlinear and complex system, we would not have sufficient elements to correctly estimate the linear regression coefficients. Given the projected increase in available data in coming years, further applications of this approach will limit themselves to the elements belonging to each class.

3.2.3 Validation of reconstruction

To validate our results, we calculated the difference and standard deviation (std) between the value of pCO_2 reconstructed in each neuron and the observation defining that neuron (Figure 9). On average, the std is approximately 38 μatm and the difference observed is 25–30 μatm. By cross-validating a dataset (divided by 10 sequences), we obtained the following result with a mean std of 48 μatm with a variation of 32–57 μatm and R equal to 0.9 with a variation of 0.86–0.96. Nevertheless some points indicating higher values can be identified (shown in red in Figure 9). These values are explained by the positions of these points, which are at the edges of the cloud and therefore more likely to include outliers that disturb the estimation of the MLR coefficients. For the reconstruction of pCO_2, with this identifiable point, it is quite easy to organize a flag system. The flag can give information about the quality of the reconstructions based on the RMS errors of the neurons used for the reconstruction. The difference obtained for pCO_2 in each neuron ranges from 0 to 56 μatm (Table 3), but 58% of the values observed are under 30 μatm. The difference can be quite high for a parameter such as SST, with a maximum value of 1.9 °C, but most of the values are lower than 1°C and CDOM ranges from 0 to 5.15 (Table 3). The other parameters have quite low variability, such as MLD, which ranges

Table 3. Maximum and mean value observed in the difference between the data used for the trainee and the value in the neurons.

Parameter	Maximum	Average
pCO_2 (μatm)	56,4	29.15
SST (°C)	1.9	0.98
time(cos)	0.33	0.07
Chl ($mg.m^{-3}$)	0.14	0.06
time(sin)	0.4	0.06
CDOM	5.15	0.06
NPP ($mg.m^{-3}$)	1.19	0.25
MLD (m)	9.7	3.2

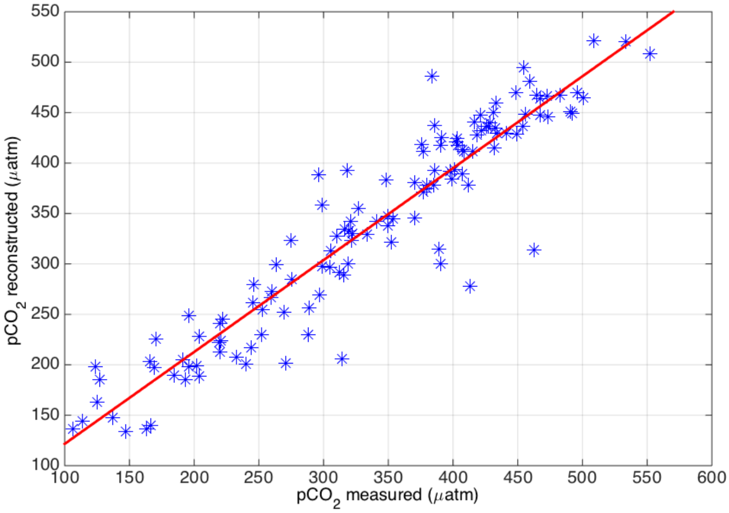

Figure 12. pCO_2 reconstructed in function of the measured. The red line represents the linear relationship. The data represents 10% of the total of data used for the validation dataset

from 0 to 9.7 m. The average is two to three times lower than the maximum value observed, which gives low value for all the satellite parameters.

455 The pCO_2 validation dataset gives a quite good correlation (R = 0.93) with the results of the reconstruction method (Figure 12), the root mean square (RMS) being 36.7 μatm: 12% of the validation data have a value higher than 20 μatm and 45% have a value between 20 μatm and 30 μatm (Figure 13). The characteristics of time, SST, MLD, CDOM, Chl, and NPP do not explain the difference observed in the reconstruction.

460 A reconstruction has been done using the satellite data from 1998 to 2011. The seasonal cycle of pCO_2 is well reproduced and in agreement with the results of other studies. The maximum is observed in winter with a pCO_2 of 437μatm on average, while the level is 274μatm in summer. These

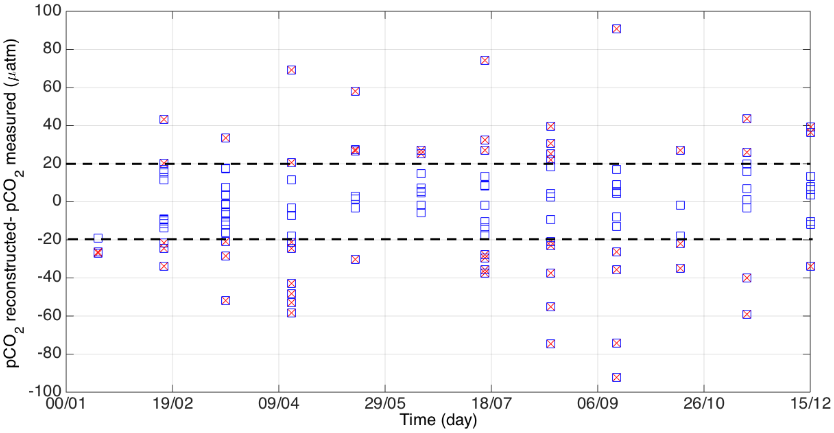

Figure 13. Difference between pCO$_2$ reconstructed and measured in function of time. The red crosses and black dotted lines represent differences greater than 20 μatm.The data represents 10% of the total of data used for the validation dataset

values are comparable to the averages estimated in the central Baltic Sea of 500 μatm in summer and 150 μatm in winter (Wesslander, 2011). The pCO$_2$ decreases in April due to the biological activity and increases slowly in September (Figure 14).

We also evaluate these results by comparing them with modeling results. The model output used in the present study is from a process-oriented biogeochemical ocean model in which the Baltic Sea is divided into 13 natural sub-basins (e.g. Omstedt et al., 2009; Norman et al., 2013b). The properties of each sub-basin are horizontally averaged and vertically resolved, and the various sub-basins are horizontally coupled to each other using strait flow models. The model is forced by meteorological gridded data with a 3-h temporal resolution and by river runoff and net precipitation data with a monthly resolution (Omstedt et al., 2005). To compare the output of the model with our results, we couple the 13 basins in optic to reduce the number to three basins, corresponding to the three basins defined in Figure 1. The modeled and estimated pCO$_2$ are compared for the entire Baltic Sea and for the three basins from 1998 to 2009 (Figure 14). The seasonal cycle for the entire Baltic Sea is well reproduced with a quite good correlation (R = 0.7) between the modeled and estimated pCO$_2$ values (Table 4), whose standard deviations differ by 74 μatm. The modeled and estimated pCO$_2$ values for the gulfs of Finland and Bothnia are not as correlated (R = 0.6), while the order of magnitude of the variability of pCO$_2$ in the Gulf of Finland as calculated with SOMLO is closer to the model estimate. (122 μatm for the modeled and 142 μatm for the estimated pCO$_2$). This lower correlation could be due to the lower number of data in this region available for these basins. The central basin is well reproduced but the amplitude of the seasonal pCO$_2$ cycle is lower in the simulation. In the southwest part of the Baltic Sea, SOMLO underestimates the pCO$_2$ concentration by 60 μatm compared with the model. In the eastern and western parts of the basin, SOMLO produces good estimates of pCO$_2$ compared with the model with an average difference of 20 μatm. In Omstedt et al. (2009), the

Table 4. Coefficient of correlation (R) between the modeled pCO_2 and the pCO_2 data estimated and the std for the model and the data.

Basin	R	STD model (μatm)	STD data (μatm)
Baltic Sea	0.7	48	96
Central Basin	0.7	69	92
Bothnian Basin	0.6	44	101
Gulf of Finland	0.6	122	144

simulated pCO_2 agrees quite well with the calculated values based on observations in the Eastern Gotland Basin.

A simple flag was constructed to monitor the reconstruction quality and give an idea of the confidence in the estimated pCO_2. The difference between the estimated and neural values is computed. The flag equals 1 for classes in which the average difference is less than 20 μatm, equals 2 for an average difference of 20–30 μatm, and equals 3 for higher average differences. In the example shown here, the flag values are high (i.e., 3), so the confidence in the reconstruction is low, but some points have flag values of 1 or 2 (Figure15d, e, and f) so the reconstruction is more reliable. On the geographic map (Figure15d, e, and f), the values of 4 correspond to the presence of ice, which is estimated using the satellite data of the National Snow and Ice Data Center based on NOAA level 3 data Njoku (2007). The flag gives confidence in our reconstruction, for example, in March 2010 (Figure 15a), the southern portion of the map (i.e., the Bornholm and Arkona basins) shows lower pCO_2 values than does the northern portion and than in February (not show here). In March 2010, this region corresponds to a flag value of 2, which was attributed medium confidence. In July 2010, the flag value is quite good and the variability of pCO_2 seems to be in line with the monthly variability (Figure 15b and e). In September 2010, the value of pCO_2 has a good order of magnitude when the flag is 2 but seems slightly too high when there is a poor confidence (i.e., a flag value of 3) (Figure 15c and f).

In conclusion, the reconstruction of pCO_2 needs to be improved to increase the confidence in the reconstruction data, particularly in the gulfs.

3.2.4 Sensitivity analysis

In order to estimate the sensitivity of the reconstruction of the pCO_2 to noisy data, a white noise was added on all incomplete data before the reconstruction. We performed three tests by adding a white noise to each parameter which was related to the STD of that parameter. The tested configurations were: 1*sigma, 2*sigma and 0.5*sigma.

The results as seen in 5) implicate that the method is sensitive to noisy data, since when increasing the noise we progressively get worse reconstructions. It is important to note that the values obtained

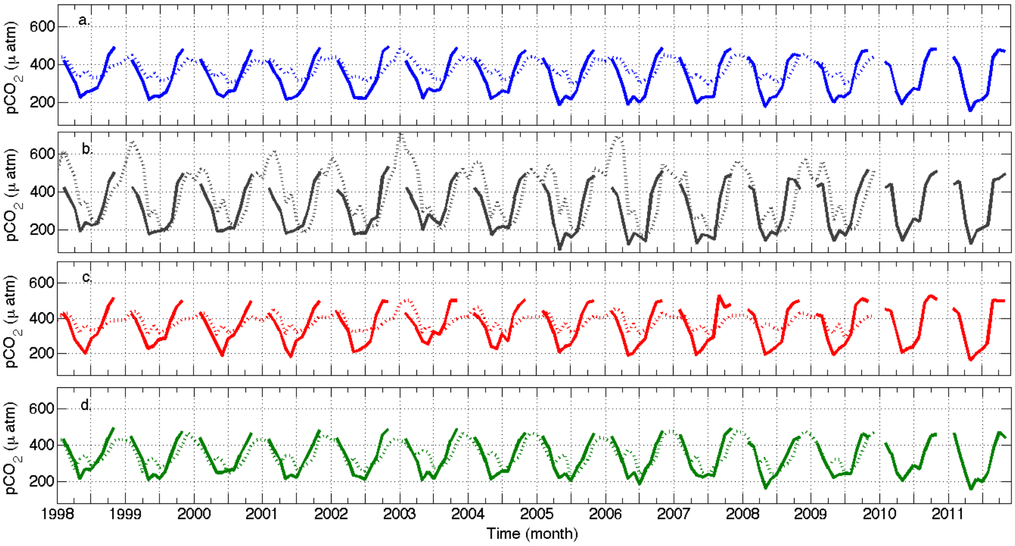

Figure 14. Comparison between modeled pCO_2 (dotted lines) and pCO_2 estimated using the SOM linear method (solid lines) for a. the Baltic Sea (BS, blue), b. the Gulf of Finland (GF, grey), c. the Gulf of Bothnia (BG, red), and d. the Central Basin (C, green).

Table 5. Coefficient of correlation (R) between the measured pCO_2 and the pCO_2 data estimated with SOMLO and the RMS in function of the sigma apply for the noise 0.5, 1 and 2 times the sigma.

X SIGMA	R	RMS (μatm)	Numbers of tests
0.5	0.85	61	701
1	0.80	77	733
2	0.59	134	709

when using 1*sigma give an average RMS of 77, and an average R coefficient of 0.80 which indicate that the method does not degrade too much.

Another area where errors might appear is the completion of missing data. In the dataset there were vectors with missing parameters. We kept only those missing up to 3 parameters. The difference observed between the pCO_2 measured and reconstructed in function of month and of number of missing parameters can be seen in Figure 16. We chose to present the data this way to also give a better understanding of the seasonal variation of the reconstruction. As seen in Table 6) the imputation method does introduce errors. When reconstructing the pCO_2 with data that have no missing data we obtain a correlation of 0.96 and an RMS of 25.7 μatm, while reconstructing the pCO_2 with data missing 3 parameters, the results are of less reliable with a correlation of 0.81 and an RMS of 51.4 μatm. The RMS we obtain in the case of the reconstruction using data that contain 1 missing value is higher than the one obtained when using data containing 2 missing values, while retaining

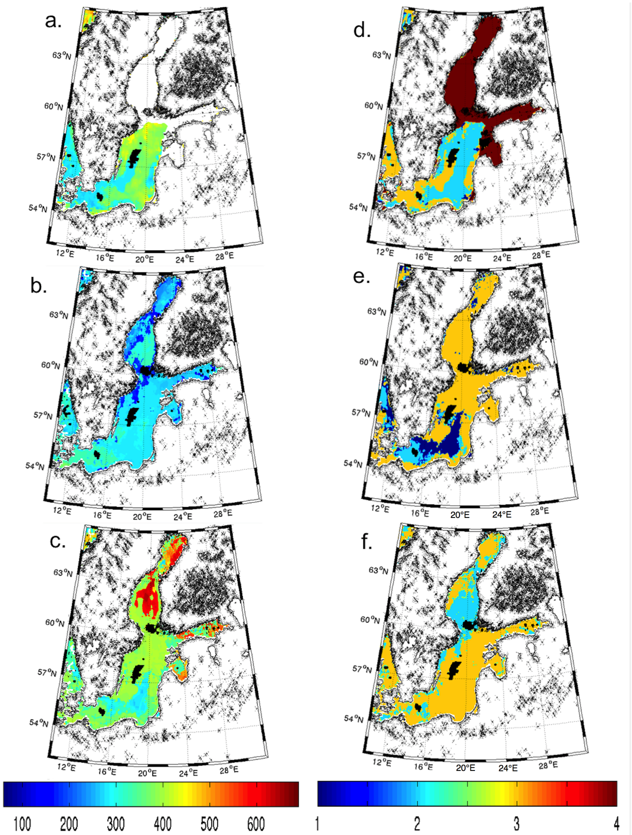

Figure 15. a, b, and c: reconstruction of the pCO_2 map and d, e, and f: the flag for each map. a.,d. March 2010,b.,e. July 2010 c.,f. September 2010. The flag values correspond to: 1 = high confidence, 2 = medium confidence, and 3 = low confidence.

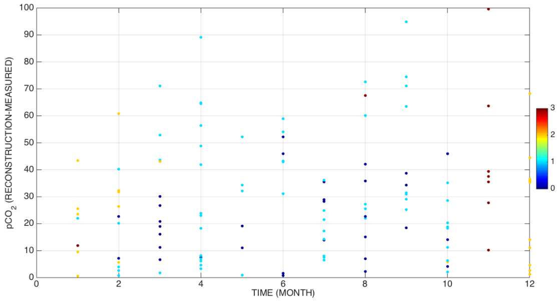

Figure 16. Absolute difference between pCO_2 reconstructed and pCO_2 measured in function of the month. The colorbar correspond to the number of empty parameter before the reconstruction (Blue : 0 missing data, light blue: 1 missing data, Yellow: 2 missing data, Red : 3 missing data

525 a higher correlation. This could be due to the higher number of vectors containing only one missing value.

Table 6. Coefficient of correlation (R) between the mesured pCO_2 and the pCO_2 data estimated and the std for the model and the data in function of the number of missing parameters.

Number of parameter missing	R	RMS (μatm)
0	0.96	25.7
1	0.93	39.5
2	0.86	31.9
3	0.81	51.4

4 Discussion and conclusions

In this paper, we used the SOMLO methodology to reconstruct the pCO_2 from satellite data for the Baltic Sea. SOMLO was used to accommodate the nonlinearity of the mechanics driving the pCO_2. It uses artificial neural networks to classify data into situations, and then performs a reconstruction by using an MLR in each class. The process involves classifying the explicative parameters (i.e., SST, CDOM, Chl, time, NPP, and MLD) and then using the linear regression coefficients corresponding to that class in order to reconstruct the pCO_2. The satellite data used was also completed using an iterative process of SOM training.

We also performed a statistical analysis of the reconstructions obtained, which allowed us to add a flag to each class, informing us of the quality of the reconstruction obtained. This could both influence further numerical modeling of other phenomena depending on the pCO_2 and allow for an informed interpretation of the reconstructions obtained.

The current results obtained using this method, based on 1445 vectors, gave a high correlation coefficient of 0.93% and an RMS of 36 μatm. These results are promising given the conditions under which we obtained them since, in addition to having a limited number of in situ pCO_2 measurements, the co-localized satellite data were frequently incomplete.

In comparison, existing studies performed over the North Atlantic and North Pacific, based on a minimum of 10,000 data points (which take into account all the data from SOCAT) to a maximum of 800,000 data points (e.g., Friedrich and Oschlies, 2009; Hales et al., 2012; Landschützer et al., 2013). Friedrich and Oschlies (2009) obtained an RMS error (RMSE) of 19 μatm. A similar study over the totality of the Atlantic Ocean obtained an RMSE of 17μatm for independent time series (Landschützer et al., 2013). Hales et al. (2012) obtained an RMSE of 20 μatm with a correlation coefficient of 0.81. The RMSE we obtained here was higher than that obtained in a previous study of the Atlantic Ocean but, taking into account the much smaller number of data available, and a stronger spatial pCO_2 variability might in the Baltic Sea compared to open-ocean. the results presented are promising.

The organization of the values of various MLR coefficients over each class indicated that all the satellite data parameters are important to reconstructing the pCO_2 in the Baltic Sea, even if only in certain cases. Improved satellite data availability could therefore also improve the performance of our reconstruction.

This study could be further developed so as to reconstruct the spatial fields of pCO_2. Specifically, one could imagine a Bayesian approach that would select which class to use for the MLR by also taking into account the potential classes attributed to the neighboring grid points of a geographic study area. This, however, remains dependent on the acquisition of additional in situ measurements to allow for the robust estimation of such Bayesian probabilities.

Many programs exist for the acquisition of new data. Data from the Östergarnsholm site are still being acquired; 2012 did not yield much data from this site, and the data from 2013 and 2014 still need to be validated. In time, the SMHI station could also supply additional data. The cargo ship transect data are not yet available for 2012–2014, but these measurements will continue, and some data will soon be available. Data are also being gathered from ferries sailing the Gothenburg–Kemi–Oulu–Lübeck–Gothenburg route. This Gothenburg transect is weekly (see http://www.hzg.de/imperia/md/content/ferryboxusergroup/presentations/fb-ws2011_karlson.pdf). The first tests of these data were conducted in 2010 and 2011, so some data should soon be available. In addition, new measurements of pCO_2 began in 2012 at the Utö Atmospheric and Marine Research Station (see http://en.ilmatieteenlaitos.fi/GHG-measurement-sites#Uto).

Given the amount of new data soon to be available, we remain optimistic that comprehension and statistical modeling of pCO_2 in the Baltic Sea will continue to improve in coming years.

Acknowledgements. We wish to thank the two reviewers for their useful comments. The authors would like to thank Dr. Tiit Kutser and Dr. Melissa Chierici for their valuable help with data collection. We would like to thank Prof. Sylvie Thiria for her advice on the SOM classifications and Marion Leduc-Leballeur for her help with the method for filling the gaps in satellite data. We wish to thank Maria Norman to provide us the model output. We would also like to thank Bernd Schneider and the BONUS/Baltic-C programmes for Cargo Ship Observations. We would like to thank the Swedish National Space Board (nr 120/11:3) for financing the project.

References

Behrenfeld, M. J. and Boss, E.: Beam attenuation and chlorophyll concentration as alternative optical indices of phytoplankton biomass, Journal of Marine Research, 64, 431–451, 2006.

Behrenfeld, M. J., Boss, E., Siegel, D. A., and Shea, D. M.: Carbon-based ocean productivity and phytoplankton physiology from space, Global biogeochemical cycles, 19, 1–25, 2005.

Bergstrom, S.: River runoff to the Baltic Sea: 1950-1990, Ambio, 23, 280–287, 1994.

Borges, A. V., Delille, B., and Frankignoulle, M.: Budgeting sinks and sources of CO_2 in the coastal ocean: Diversity of ecosystems counts, Geophysical Research Letters, 32, 2005.

Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., and Ludicone, D.: Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, J. Geophys. Res, 109, C12 003, 2004.

Burchard, H. and Bolding, K.: GETM: A General Estuarine Transport Model; Scientific Documentation, European Commission, Joint Research Centre, Institute for Environment and Sustainability, 2002.

Casey, K. S., Brandon, T. B., Cornillon, P., and Evans, R.: The past, present, and future of the AVHRR Pathfinder SST program, in: Oceanography from Space, pp. 273–287, Springer, 2010.

Chen, C.-T. A. and Borges, A. V.: Reconciling opposing views on carbon cycling in the coastal ocean: Continental shelves as sinks and near-shore ecosystems as sources of atmospheric CO_2, Deep Sea Research Part II: Topical Studies in Oceanography, 56, 578–590, doi:10.1016/j.dsr2.2009.01.001, 2009.

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 Research Part II: Topical Studies in Oceanography, 56, 630–639, 2009.

Cloern, J. E., Grenz, C., and Vidergar-Lucas, L.: An empirical model of the phytoplankton chlorophyll: carbon ratio-the conversion factor between productivity and growth rate, Limnology Oceanography, 40, 1313–1321, 1995.

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

Dickson, A. and Millero, F.: A comparison of the equilibrium constants for the dissociation of carbonic acid in seawater media, Deep Sea Research Part A. Oceanographic Research Papers, 34, 1733–1743, 1987.

Dreyfus, G.: Neural networks : methodology and applications, Springer, Berlin ; New York, 2005.

Dutt, A. and Rokhlin, V.: Fast Fourier transforms for nonequispaced data, II, Applied and Computational Harmonic Analysis, 2, 85–100, 1995.

Eppley, R. W.: Temperature and phytoplankton growth in the sea, Fish. Bull, 70, 1063–1085, 1972.

Friedrich, T. and Oschlies, A.: Neural network-based estimates of North Atlantic surface pCO_2 from satellite data: A methodological study, Journal of Geophysical Research: Oceans (1978–2012), 114, 1–12, 2009.

Grasshoff, K., Kremling, K., and Ehrhardt, M.: Methods of seawater analysis third edition, VCH Publishers, 1999.

Greengard, L. and Lee, J.-Y. . Y.: Accelerating the nonuniform fast Fourier transform, SIAM review, 46, 443–454, 2004.

Hales, B., Strutton, P. G., Saraceno, M., Letelier, R., Takahashi, T., Feely, R., Sabine, C., and Chavez, F.: Satellite-based prediction of pCO_2 in coastal waters of the eastern North Pacific, Progress in Oceanography, 103, 1–15, doi:10.1016/j.pocean.2012.03.001, 2012.

Hjalmarsson, S., Wesslander, K., Anderson, L. G., Omstedt, A., Perttilä, M., and Mintrop, L.: Distribution, long-term development and mass balance calculation of total alkalinity in the Baltic Sea, Continental Shelf Research, 28, 593–601, 2008.

Jamet, C., Moulin, C., and Lefèvre, N.: Estimation of the oceanic pCO_2 in the North Atlantic from VOS lines in-situ measurements: parameters needed to generate seasonally mean maps, in: Annales Geophysicae, vol. 25, pp. 2247–2257, Copernicus EGU, 2007.

Jolliffe, I. T.: Principal component analysis, Springer, New York, 2002.

Kohonen, T.: The self-organizing map, Proceedings of the IEEE, 78, 1464–1480, 1990.

Landschützer, P., Gruber, N., Bakker, D. C. E., Schuster, U., Nakaoka, S., Payne, M. R., Sasse, T. P., and Zeng, J.: A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink, Biogeosciences, 10, 7793–7815, 2013.

Laruelle, G. G., Dürr, H. H., Slomp, C. P., and Borges, A. V.: Evaluation of sinks and sources of CO_2 in the global coastal ocean using a spatially-explicit typology of estuaries and continental shelves, Geophysical Research Letters, 37, doi:10.1029/2010GL043691, 2010.

Lee, Z.-P., Darecki, M., Carder, K. L., Davis, C. O., Stramski, D., and Rhea, W. J.: Diffuse attenuation coefficient of downwelling irradiance: An evaluation of remote sensing methods, Journal of Geophysical Research: Oceans (1978–2012), 110, 1–9, 2005.

Lefèvre, N., Watson, A. J., Olsen, A., Ríos, A. F., Pérez, F. F., and Johannessen, T.: A decrease in the sink for atmospheric CO_2 in the North Atlantic, Geophysical Research Letters, 31, 1–4, doi:10.1029/2003GL018957, 2004.

Lefèvre, N., Watson, A. J., and Watson, A. R.: A comparison of multiple regression and neural network techniques for mapping in situ pCO_2 data, Tellus B, 57, 375–384, doi:10.1111/j.1600-0889.2005.00164.x, 2005.

Lewis, E. and Wallace, D. W. R.: Program developed for CO_2 system calculations, Tech. rep., ORNL/CDIAC, US Department of Energy, Oak Ridge, Tennessee, 1998.

Lüger, H., Wallace, D. W., Körtzinger, A., and Nojiri, Y.: The pCO_2 variability in the midlatitude North Atlantic Ocean during a full annual cycle, Global biogeochemical cycles, 18, 1–16, doi:10.1029/2003GB002200, 2004.

Malek, M. A., Harun, S., Shamsuddin, S. M., and Mohamad, I.: Imputation of time series data via Kohonen self organizing maps in the presence of missing data, Engineering and Technology, 41, 501–506, 2008.

Merbach, C., Culberson, C. H., and Hawley, J. E.: Measurements of the apparent dissociation constants of carbonic acid in seawater of atmospheric pressure, Lymnology and oceanography, 18, 897–907, 1973.

Morel, A. and Berthon, J.-F. . F.: Surface pigments, algal biomass profiles, and potential production of the euphotic layer: Relationships reinvestigated in view of remote-sensing applications, Limnol. Oceanogr, 34, 1545–1562, 1989.

Morel, A. and Gentili, B.: A simple band ratio technique to quantify the colored dissolved and detrital organic material from ocean color remotely sensed data, Remote Sensing of Environment, 113, 998–1011, 2009.

Nakaoka, S., Telszewski, M., Nojiri, Y., Yasunaka, S., Miyazaki, C., Mukai, H., and Usui, N.: Estimating temporal and spatial variation of ocean surface pCO_2 in the North Pacific using a self-organizing map neural network technique, Biogeosciences, 10, 6093–6106, 2013.

Njoku, E.: AMSR-E/Aqua L2B Surface Soil Moisture, Ancillary Parms, & QC EASE-Grids v002, Información digital. Actualizado diariamente. Boulder, Colorado EEUU: National Snow and Ice Data Center, 2007.

Norman, M., Raj Parampil, S., Rutgersson, A., and Sahlée, E.: Influence of coastal upwelling on the air-sea gas exchange of CO_2 in a Baltic Sea Basin, Tellus B, 65, 2013a.

Norman, M., Rutgersson, A., and Sahlée, E.: Impact of improved air-sea gas transfer velocity on fluxes and water chemistry in a Baltic Sea model, Journal of Marine Systems, 111, 175–188, doi:10.1016/j.jmarsys.2012.10.013, 2013b.

Omstedt, A., Elken, J., Lehmann, A., and Piechura, J.: Knowledge of the Baltic Sea physics gained during the BALTEX and related programmes, Progress in Oceanography, 63, 1–28, 2004.

Omstedt, A., Chen, Y., and Wesslander, K.: A comparison between the ERA40 and the SMHI gridded meteorological databases as applied to Baltic Sea modelling., Nordic hydrology, 36, 369–380, 2005.

Omstedt, A., Gustafsson, E., and Wesslander, K.: Modelling the uptake and release of carbon dioxide in the Baltic Sea surface water, Continental Shelf Research, 29, 870–885, 2009.

Platt, T. and Sathyendranath, S.: Estimators of primary production for interpretation of remotely sensed data on ocean color, Journal of Geophysical Research: Oceans (1978–2012), 98, 14 561–14 576, 1993.

Rutgersson, A., Norman, M., Schneider, B., Pettersson, H., and Sahlée, E.: The annual cycle of carbon dioxide and parameters influencing the air–sea carbon exchange in the Baltic Proper, Journal of Marine Systems, 74, 381–394, 2008.

Sasse, T. P., McNeil, B. I., and Abramowitz, G.: A novel method for diagnosing seasonal to inter-annual surface ocean carbon dynamics from bottle data using neural networks, Biogeosciences, 10, 4319–4340, 2013.

Schafer, J. L. and Graham, J. W.: Missing data: our view of the state of the art., Psychol Methods, 7, 147–77, 2002.

Schneider, B. and Kaitala, S.: Identification and quantification of plankton bloom events in the Baltic Sea by continuous pCO_2 and chlorophyll a measurements on a cargo ship, Journal of Marine systems, 59, 238–248, 2006.

Schneider, B., Nausch, G., Nagel, K., and Wasmund, N.: The surface water CO_2 budget for the Baltic Proper: a new way to determine nitrogen fixation, Journal of Marine Systems, 42, 53–64, 2003.

Schneider, B., Kaitala, S., Raateoja, M., and Sadkowiak, B.: A nitrogen fixation estimate for the Baltic Sea based on continuous pCO_2 measurements on a cargo ship and total nitrogen data, Continental Shelf Research, 29, 1535–1540, 2009.

Schomberg, H. and Timmer, J.: The gridding method for image reconstruction by Fourier transformation., IEEE Trans Med Imaging, 14, 596–607, doi:10.1109/42.414625, 1995.

Schuster, U., McKinley, G. A., Bates, N., Chevallier, F., Doney, S. C., Fay, A. R., González-Dávila, M., Gruber, N., Jones, S., and Krijnen, J.: An assessment of the Atlantic and Arctic sea-air CO_2 fluxes, 1990-2009, Biogeosciences, 10, 607–627, 2013.

Siegel, H. and Gerth, M.: Baltic Sea environment fact sheet Sea Surface Temperature in the Baltic Sea in 2011, HELCOM Baltic Sea Environment Fact Sheets [online], http://www.helcom.fi/baltic-sea-trends/environment-fact-sheets/, 2012.

Stocker, T. F., Qin, D., Plattner, G.-K. . K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M.: Climate Change 2013. The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change-Abstract for decision-makers, Tech. rep., Groupe d'experts intergouvernemental sur l'evolution du climat/Intergovernmental Panel on Climate Change-IPCC, C/O World Meteorological Organization, 7bis Avenue de la Paix, CP 2300 CH-1211 Geneva 2 (Switzerland), 2013.

Takahashi, T., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A., Bakker, D. C., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T. S., Tilbrook, B., Johannessen, T., Olsen, A. O., Tilbrook, B., Bellerby, R., Wong, C. S., Delille, B., Bates, N. R., and De Baar, H. J. W.: Climatological mean and decadal change in surface ocean pCO_2, and net sea–air CO_2 flux over the global oceans, Deep-Sea Research Part II, 56, 554–577, doi:10.1016/j.dsr2.2008.12.009, 2009.

Telszewski, M., Padín, X. A., and Ríos, A. F.: Estimating the monthly pCO_2 distribution in the North Atlantic using a self-organizing neural network, Biogeosciences, 6, 1405–1421, 2009.

Weiss, R. F.: CO_2 in water and seawater: The solubility of a non ideal gas, Marine Chemistry., 2, 203–215, 1974.

Wesslander, K.: The carbon dioxide system in the Baltic Sea surface waters, Ph.D. thesis, University of Gothenburg, 2011.

Wesslander, K., Omstedt, A., and Schneider, B.: Inter-annual variation of the air-sea CO_2 balance in the southern Baltic Sea and the Kattegat, in: EGU General Assembly Conference Abstracts, vol. 11, p. 8629, 2009.

(R1)