Remote sensing algorithm for sea surface CO₂ in the Baltic Sea

G. Parard¹, A. A. Charantonis²,³, and A. Rutgerson¹

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

Received: 30 June 2014 – Accepted: 8 July 2014 – Published: 12 August 2014

Correspondence to: G. Parard (gaelle.parard@geo.uu.se)

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

Studies of coastal seas in Europe have brought forth the high variability in the CO$_2$ system. This high variability, generated by the complex mechanisms driving the CO$_2$ fluxes makes their accurate estimation an arduous task. This is more pronounced in the Baltic Sea, where the mechanisms driving the fluxes have not been as highly detailed 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, a combination of two existing methods (Self-Organizing-Maps and Multiple Linear regression) is used to estimate ocean surface $p$CO$_2$ in the Baltic Sea from remotely sensed surface temperature, chlorophyll, coloured dissolved organic matter, net primary production and mixed layer depth. The outputs of this research have an horizontal resolution of 4 km, and cover the period from 1998 to 2011. The reconstructed $p$CO$_2$ values over the validation data set have a correlation of 0.93 with the in-situ measurements, and a root mean square error is of 38 µatm. The removal of any of the satellite parameters degraded this reconstruction of the CO$_2$ flux, and we chose therefore to complete any missing data through statistical imputation. The CO$_2$ maps produced by 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 the coming years, in view of the predicted acquisitions of new data.

1 Introduction

The ocean plays an important role as a major carbon reservoir for carbon dioxide (CO$_2$) emitted to the atmosphere from fossil fuel burning, cement production, biomass burning, deforestation and other land use change. The ocean is at present acting to slow the rate of climate change by absorbing about 30% of human emissions of CO$_2$ emitted to the atmosphere since the industrial revolution (Stocker et al., 2013). The exchange
of CO₂ between coastal environments and the atmosphere is significant for the global carbon budget (e.g. Borges et al., 2005; Chen and Borges, 2009; Laruelle et al., 2010). The estimation of the overall sink of CO₂ in continental shelf sea is −0.22 Pg C yr⁻¹ (Laruelle et al., 2010), corresponding to 16% of the open oceanic sink (Takahashi et al., 2009), while continental shelves only represent 7% of the oceanic surface area and less than 0.5% of the ocean volume. These estimates are prone to a large uncertainty related to sparse data coverage in time and space. The monitoring of the oceanic partial pressure of CO₂ (pCO₂) at monthly and seasonal time scale is essential to estimate the regional and global air-sea CO₂ flux and improve this uncertainty. Due to technical as well as financial restrictions, in situ measurements of marine pCO₂ are sparse in spatial and temporal distribution. However, over the last decade, technical improvements and cooperation with the shipping industry have allowed for the installation of several autonomous underway systems on board commercial vessels routinely crossing the ocean basin. Those instruments perform quasi-continuous measurements, offering temporal and spatial coverage which allows for regional analysis of the highly variable spatial and temporal distribution of pCO₂ (e.g. Lefèvre et al., 2004; Lüger et al., 2004; Corbière et al., 2007; Schneider et al., 2003). In spite of the increased number of measurements in the Baltic Sea, the assessment of the carbon fluxes in the Baltic Sea remains particularly challenging due to the non-linearity of the emission and absorption system. Neural network techniques can be generally described as empirical statistical tools that resolve, to a certain degree, the nonlinear and often discontinuous relationships among proxy parameters without any a priori assumptions. In the past decade a handful of authors have reported the application of a neuronal network technique to basin-scale pCO₂ sea analysis (Lefévre et al., 2005; Jamet et al., 2007; Friedrich and Oschlies, 2009; Telszewski et al., 2009), concentrating mainly on the North Atlantic Ocean. Most recently, (Telszewski et al., 2009) successfully applied a self-organizing-map (SOM) based neuronal network technique to reconstruct pCO₂ sea distribution in the North Atlantic (10.5° to 75.5° N, 9.5° E to 75.5° W) for three years (2004 to 2006) by examining non-linear/discontinuous relationship between pCO₂ sea
and ocean parameters of sea surface temperature (SST), mixed layer depth (MLD), and chlorophyll a concentration (CHL). One of the main benefits of this approach over the more traditional techniques, such as multiple linear regression (MLR), is that there are numerous empirical relationships established (e.g., 2220 in Telszewski et al., 2009) between examined parameters, allowing for more accurate representation of the highly variable system of interconnected water properties. In this paper, we departed from a similar, but rougher SOM classification of the available explicative oceanic parameters in the Baltic sea. Making the assumption that the explicative parameters of each of the obtained classes could locally be considered to be linearly related to the $pCO_2$, we performed a MLR of the $pCO_2$ from these explicative parameters in each of the retrieved classes. The purpose of the method developed is to use the classes and MLR parameters calculated for each class, to estimate the $pCO_2$ in the Baltic Sea and to improve the estimation of air-sea CO$_2$ fluxes in the future. The manuscript has been structured in four parts. In the first part we present a synopsis of the problem studied, including the existing studies performed on the reconstruction of the $pCO_2$ in other maritime regions. We followed that with a presentation of the available data and a brief description of the methodologies used. In the third part of the article we presented our results, namely the topological maps obtained as well as the reconstructions performed with them. We concluded the article with a discussion on the results obtained and future possible improvements.

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 from the Baltic Proper showed large variability of $pCO_2$ in time and in space. The amplitude of the annual cycle of $pCO_2$ varies significantly depending on the region (Schneider and Kaitala, 2009).
2006) between 400 µatm and 120 µatm respectively in North eastern Baltic Proper and in the transition areas to the North Sea. It receives a significant river runoff from surrounding land (a total of about 15,000 m$^3$ s$^{-1}$ (Bergstrom, 1994) and a net precipitation rate of about 1500 m$^3$ s$^{-1}$ (Omsted et al., 2004). This large freshwater addition bring large amounts of nutrients and inorganic and organic carbon (Omsted et al., 2004; Hjalmarsson et al., 2008). The biogeochemical processes in the marine environment of the Baltic Sea are mainly controlled by the biological production and decomposition of organic matter taking place in the context of the hydrography of the region. Physical forcing controls water transport, stratification, temperature and salinity in the Baltic Sea; these then influence the distribution of nutrients and carbon and thus have an impact on biogeochemical processes. The Baltic Sea system can be divided into several sub-bassins (Baltic Proper, Gulf of Riga, Gulf of Finland, Bothnian Sea and Bothnian Bay) seen in Fig. 1 from Omstedt et al. (2009). The average depth of the Baltic Sea is 55 m with a maximum depth of 460 m at the Landsort Deep (Wesslander, 2011).

2.2 $p$CO$_2$ observations

We used the compilation of $p$CO$_2$ data from three different sources.

1. The Östergarnsholm site: this site is located next to the small island Östergarnsholm in the central Baltic Sea described in (Rutgersson et al., 2008; Norman et al., 2013). The measurements of $p$CO$_2$ has been running since 2005, 4 km from the East coast of Gotland. Sea surface temperature (SST) and $p$CO$_2$ are measured by a SAMI-CO2 sensor (Submersible Autonomous Moored Instrument). The sensor is located at 4 m depth, 1 km southeast from the tower, and in use semi-continuously. In addition, SST is also measured by a wave rider buoy (operated by the Finnish Meteorological Institute) at 0.5 m depth about 4 km southeast of the tower.

2. Cargo ship: this data set come from continuous measurements of the surface water $p$CO$_2$. They take place 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 \( p\text{CO}_2 \) at a depth of 5 m on board 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 taken between July 2003 to December 2005.

3. Swedish Meteorological and Hydrological Institute (SMHI) data set: pH (measured with the method from Grasshoff et al., 1999) and alkalinity (TA) (measured by potentiometric titration described in Grasshoff et al., 1999) are measured continuously at a monthly or semi-monthly resolution in the Baltic Sea at different stations. All data used are from 5 m depth below the surface. The uncertainty for the pH is \( \pm 0.03 \) pH units and for the TA is \( \pm 5\% \) (Wesslander et al., 2009). \( p\text{CO}_2 \) is estimated from the pH, TA, salinity and temperature measurements and used the standard CO2SYS program (Lewis and Wallace, 1998) with the same equilibrium constant from Weiss (1974) and Merbach et al. (1973) as refitted by Dickson and Millero (1987) like in Wesslander et al. (2009).

To apply a neural network method to reconstruct the \( p\text{CO}_2 \) from 1998 to 2011, we used several satellite data set.

2.3 Remote sensing data

The satellite data used for this study have different sources. We used a monthly time resolution and a spatial resolution based from the lower spatial resolution of our products. The lower spatial resolution is the CDOM (Coloured Dissolved Organic Matter) product.

We used five parameters from different sources:

**SST** Sea Surface Temperature: for the SST several products are used, we combine two types of products for 2007 and 2011. For 2005–2011: we used the
data from Federal Maritime and Hydrographic Agency (BSH) processing the
data from AVHRR-NOAA for 2005–2011 and the data from GRHSST (Group for
The spatial resolution is 0.03° at daily scale (http://podaac.jpl.nasa.gov/dataset/
DMI-L4UHfind-NSEABALTIC-DMI_OI). From 1998–2004, the data is a re-analysis
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). It consists of 4 km monthly SST (in °C) extracted from version
5.2 of the AVHRR Pathfinder project (Casey et al., 2010, http://www.nodc.noaa.
gov/SatelliteData/pathfinder4km/).

**Chl** Chlorophyll a: The dataset consists of monthly averages from the following sen-
sors: SeaWiFS (September 1998–December 2002) 4 km monthly and MODIS-
Aqua (July 2002–June 2011) 4 km monthly (Casey et al., 2010). A lognormal dis-
tribution was assumed for the Chl.

**CDOM** Coloured Dissolved Organic Matter come from MODIS data, at 4 km monthly
average. The CDOM index quantifies the deviation in the relationship between
CDOM and chl concentration, where 1.0 represents the mean relationship for
Morel and Gentili (2009) case 1 waters, and values above or below 1.0 indicate
excess or deficit in CDOM relative to that mean relationship, respectively. The
algorithm and its application is fully described in Morel and Gentili (2009)

**NPP** Primary Production come from two products. The first one come from EMIS: This
model is depth-integrated but allows for depth-dependent variability in the dif-
fuse attenuation coefficient, which is calculated from a multiple-component semi-
analytical inversion algorithm (Lee et al., 2005). The primary production calcula-
tion is based on the formulation obtained through dimensional analysis by Platt
and Sathyendranath (1993).The assignment of the photosynthetic parameters is
achieved 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 following the empirical relation of Cloern et al. (1995). The dataset in EMIS consists of monthly average values from October 1997 to September 2008. The second for 2009–2011 is from the Vertically Generalized Production Model (VGPM) (Behrenfeld and Boss, 2006) as the standard algorithm. The VGPM is a “chlorophyll-based” model that estimate net primary production from chlorophyll using a temperature-dependent description of chlorophyll-specific photosynthetic efficiency. For the VGPM, net primary production is a function of chlorophyll, available light, and the photosynthetic efficiency. Standard product are based on MODIS chlorophyll and temperature data, SeaWiFS PAR, and estimates of euphotic zone depth from a model developed by Morel and Berthon (1989) and based on chlorophyll concentration. A correction between this two product has be done for the maximum value.

MLD  Mixed Layer Depth: There is also two sources for the MLD. At monthly averages from 1998 to 2007, come from one3D hydrodynamic model currently used at the JRC/IES is the public domain GETM model (General Estuarine Transport Model – http://www.getm.eu), which has its roots within 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 GOTM (General Ocean Turbulence – http://www.gotm.net). Between 2008 and 2011, we used the data from Carbon-based Production Model at monthly scale (Behrenfeld et al., 2005).

Some corrections were applied for each parameter of the data to render the different products coherent between themselves.

In the Baltic Sea the satellite data have a lot of gaps, due to the high proportion coastal waters, where the satellite products are less reliable, and the frequent large-scale cloud coverage. To improve the number of our data we used a monthly time scale
and we apply a method to improve the spatial distribution. To be used for statistical analysis, the irregularly-spaced density measurements were first uniformly re-sampled. To this end, a Gaussian grinding was used, as described in Greengard and Lee (2004); Dutt and Rokhlin (1995). The data points of the original series are convolved with a Gaussian kernel. As a result, the data points are smeared over their neighbouring equi-spaced points, which are more densely distributed. This type of method produces more realistic values than simple interpolation, particularly when there are many data gaps (Schomberg and Timmer, 1995).

### 2.4 Data available

All the \( p\text{CO}_2 \) data were put together (Fig. 2). We use the spatial resolution of the parameter with the lowest resolution for the final product choose (this is CDOM). The time resolution used for this study is monthly scale. Comparison between \( p\text{CO}_2 \) from SAMI sensor with the data around the mooring (0.2°) are done with the other data set and give a good correlation factor and (0.98). The data are mainly available in the Gotland Basin and Arkoria basin but the quantity of data in the Bothnian Sea are very low and correspond of two stations from SMHI. The \( p\text{CO}_2 \) data are well distributed over the twelve months (Fig. 3). January is the month where the number of data is lower (lower than 80), but for the other month is range between 110 and 155.

In our case, the each data point is characterised by SST, Chl, CDOM, NPP, MLD but also an information on the date the measurements were taken. This temporal information was normalised by sine and cosine, as follows:

\[
T(\text{cosine}) = \cos \left( \frac{\text{Days} \cdot 2\pi}{365} \right) \tag{1}
\]

\[
T(\text{sine}) = \sin \left( \frac{\text{Days} \cdot 2\pi}{365} \right) \tag{2}
\]

This definition of time is used to render the values continuous over a year, sidestepping the artificial numerical transition from the 365th day of the year to the 1rst day.
of the following year, and to therefore be able to situate the process in relation to its seasonality.

In total, 1445 data of \( pCO_2 \) were used for this study. Each parameter (SST, Chl, CDOM, NPP and MLD) was located around each data of \( pCO_2 \). During winter (October to March), more data were missing (Table 1, column 1), in particular for Chl, CDOM and NPP which is the period when it’s more difficult to measure or estimate these parameters. Between April and September, the number of data missing for SST, Chl, CDOM and MLD are relatively low compared to the total amount of data. Missing data represents less than 3 %. To increase the number of data available, we completed the data with a training of the topological map. Further detailed is given in Sect. 2.5.

2.5 Methodology

The relationship between \( pCO_2 \) and the environmental parameters available is highly non-linear. As mentioned in the introduction, we chose to combine two statistical approaches: self-organizing maps (Kohonen, 1990; Dreyfus, 2005) and linear regression. The SOM are a subfamily of the neuronal networks algorithms, used to perform multidimensional classification. One of the defining characteristics of the SOMs is their ability of their classes to locally represent a gaussian distribution centred around its typical profile of the environmental parameters. We used this hypothesis to classify the environmental parameters dataset, and then estimate for each class the parameters of a linear regression. In the following section we will present an overview of the two statistical algorithms and their application to our data sets.

2.5.1 Self organizing maps

Self-organizing topological maps (SOM) is a clustering method based on neural networks. They provide a clustering of a learning data set into a reduced number of subsets, called classes, with common statistical characteristics.
The generation of a SOM requires the creation of a training database that contains homogenous vectors. After a training phase, we obtain a Self-Organizing Map. Each class is represented by its referent vector $r_i$ which is an approximation of the mean value of the elements belonging to it and it index that positions it on the map in relation to the other classes (Fig. 4). The topological aspect of the maps can be justified by considering the map as an undirected graph on a two-dimensional lattice whose vertices are the $N$ classes. This graph structure permits the definition of a discrete distance $d(C_i, C_j)$ between two classes $C_i$ and $C_j$, defined as the length of the shortest path between $C_i$ and $C_j$ on the map. The nature of the SOM training algorithm forces a topological ordering upon the map and, therefore, any neighbouring classes $C_i$ and $C_j$ on the map ($d(C_i, C_j) = 1$) 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$_i$ ($||x - r_i||$), therefore assigning it to the class whose referent is closest to it in the Euclidean sense (Fig. 5). A classified vector $x$ will be represented by its class index, $C_{\text{index}}$. In the case 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 perform 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 neighbours’ 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 neighbouring vectors on the determination of the referent vector can be considered null. In these cases, each referent vector approximates, locally, the mean value of a multidimensional Gaussian random distribution that generated the training data assigned to that class (Dreyfus, 2005).
2.5.2 Multiple linear regression

A multiple linear regression is a modelling method that expresses the value of one response variable \( y \) (in our study \( p\text{CO}_2 \)) as a linear function of other explicative variables \( X = X_1, X_2, \ldots, X_i \) (in our study SST, Chl, CDOM, NPP, MLD, \( \text{time}_{\text{sin}}, \text{time}_{\text{cos}} \)). The purpose of performing a multiple linear regression generally is either to interpret the relationship of the variable \( y \) with 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, in order to perform our MLRs, we had to take into account their limitations and the nature of our problem. More specifically, in order to perform a MLR we are obliged to assume that the relationship between the predictor variables and the response variable is \textit{linear}. However, this is not the case in our data sets. The \( p\text{CO}_2 \) does not follow a linear relationship with the variables presented when considering the entirety of the problem presented. However as noted above in Sect. 2.5.1, if we consider the classes created by the SOM, they are very localised 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 make the assumption that, if performed in the reduced neighbourhood of a SOM class, the relationship between \( p\text{CO}_2 \) and the explicative variables is linear.

3 Application and results

3.1 Statistical imputation

As described in Sect. 2.4, the data available for the application were presenting many missing values. In order to complete them we chose to use an imputation method 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 in order to
regroup the data in typical situations and replace the missing values of the explicative data with the corresponding values of the referent vector of the class it belongs to.

We first selected the database containing, as mentioned before: SST, Chl, CDOM, NPP, MLD, time\textsubscript{sin}, time\textsubscript{cos}. We then sorted its vectors according to the number of missing values in them and noting the placement of missing values for each vector. We chose all complete data vectors and the first 5\% of the sorted vectors with missing data and trained a SOM. We proceeded by replacing the missing values of these initial 5\% with the corresponding values of the referent vector of the class they each belonged to. We performed the same process by iteratively increased by 5\% the amount of the vectors with missing values included in the training of a new SOM, and by replacing all missing values, even those completed in previous iterations, by the corresponding values of the referent vector of the class that each vector with missing values belonged to. A more detailed presentation of this method is the subject of another forthcoming article, but a schematic representation version of the imputation method used can be seen in Fig. 6.

After this imputation of the missing data through the iterative training give a good representation of the data as presented in Fig. 7. The repartition of $pCO_2$ (Fig. 7a) is well representative of the variability of the data with a large range of value. Some very high in particular local events like a coastal upwelling but most of the data are range between 180 $\mu$atm (value observed in summer) and 550 $\mu$atm (observed in winter). The SST (Fig. 7b) is well representative of the variability in the Baltic Sea with a maximum of observation between July and September in all the basin around 18°C (Siegel and Gerth, 2012). The NPP variability is quite homogenous except the pic at 10 mg C m\textsuperscript{-2} d\textsuperscript{-1} this is due to the first model which have is maximum of NPP at 10 mg C m\textsuperscript{-2} d\textsuperscript{-1}. So the correction on the satellite data of NPP take into account this maximum.

The variability of chlorophyll gives a part of data with a low value and an other part with a higher value than 6 mg m\textsuperscript{-3}, which can be explain by the fact that the Baltic Sea is a narrow sea so the coast are quite important and the two bloom take place during
spring/summer, during this period the value of chlorophyll can be very high, and the reconstruction give a mean value for this characteristic. A pic at 10 mg m$^{-3}$ is observed on the chlorophyll data, this pic is not due to the reconstruction but to the maximum value in the satellite data file. The low level of MLD occurs in summer and in the model the minimum is 10 m depth, which appear to be around the minimum value observed on the Fig. 7f. Absorption by CDOM decreased with increased distance from the riverine source and reached a relatively stable absorption background in the open sea. Most of our data are more in open sea conditions, the value are quite low.

Once the data was completed we could combine the two datasets again and train the SOM used for the reconstruction of $p$CO$_2$.

3.2 $p$CO$_2$ estimation

3.2.1 Topological map

In this study we classified the explicative variables (SST, Chl, CDOM, NPP, MLD, time$^\text{sin}$, time$^\text{cos}$) into classes that share similar characteristics. We separated our data set in two parts: 90% of the completed data (1300 vectors) were used for the training phase, with the remaining 10% split into 5% (72 vectors) for testing and 5% for the validation of our method. We iteratively tested classifications with varying number of classes, and selected the parameters of our SOM based on the performances on the test dataset.

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 (Fig. 8). The order of magnitude of the number of observations is constant throughout the SOM, and we can consider that it has been well deployed to represent the data space of the explicatory 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 neighbours.
In order to estimate the average concentration of $pCO_2$ in each class, its measurements associated with vectors consisting of SST, CDOM, NPP, MLD and CHL components were presented to the already trained SOM as input data (Fig. 9). 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 of the SST distribution with low values of $pCO_2$ correlating with high value of SST (Fig. 9). This is in agreement with the seasonal cycle of $pCO_2$, which is characterized by a large amplitude, ranging from high value in winter ($\approx 500\,\mu$atm) and low value in summer ($\approx 150\,\mu$atm) Wesslander (2011). From Schneider and Kaitala (2006), the high winter value of $pCO_2$ are a consequence of mixing with deeper water layer enriched in CO$_2$ which is in agreement with the distribution of the MLD (Fig. 9h) with the higher value during winter and autumn correlate with the high value of $pCO_2$. Wesslander (2011) explained also that it can be related to the mineralization, which exceed production in winter. Biological production starts in spring when sunlight and nutrients are sufficient. The chlorophyll begin to increase in March/April due to the spring phytoplankton bloom, which have for effect to reduce the $pCO_2$ during this period. The more intensive decrease take place in April and May which is consistent with the higher value of NPP (Fig. 9). The studies in the central Baltic Sea shows two summer minima, the first is in April/May and a second July/August which is resulting from a second production period. The higher variability are observed during this period with standard deviation between 39 $\mu$atm and 50 $\mu$atm from different region (Wesslander, 2011; Schneider and Kaitala, 2006).

### 3.2.2 Linear regression in the neurons

As mentioned in Sect. 2.5.2, in order to perform a MLR we are obliged to assume that the relationship between the predictor variables and the response variable is linear, which we could only take as a valid hypothesis when performing the MLR in the reduced neighbourhood of a SOM class, where the relationship between $pCO_2$ and the explicative variables can be assumed to be linear.
For each class \( j \) therefore we created a separate training data set containing all vectors that were assigned to that class and to all its adjacent classes. Based on that data-set we computed the linear regression coefficient parameters for every explicative parameter and a constant value.

The calculated linear regression coefficient parameter values for each class are shown in Fig. 10. We can notice 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 demonstrated that each and every parameter is important in the reconstruction of the \( p\text{CO}_2 \) in the Baltic Sea, even if it is highly significant in some particular classes and relatively stable in other regions of the topological map.

The addition of vectors belonging to adjacent classes in general did not perturb the estimation of the coefficient parameters since, as seen in Fig. 9, the values of all parameters are generally organized in a coherent fashion on the map. The assumption that they are close in the data space is not as robust as it would have been had we solely considered the vectors belonging to each class but, given the limited amount of data available for the modeling of this highly non-linear and complex system, we would not have sufficient elements to correctly estimate the linear regression coefficients. Given the projected increase in available data in the coming years, further applications of this approach will limit themselves to the elements belonging to each class.

### 3.2.3 Validation of reconstruction

In order to validate our results, we calculated the difference and the standard deviation (std) between the value of \( p\text{CO}_2 \) reconstructed in each neurons and the observations which define the neurons (Fig. 11). In average the std is around 38 \( \mu \text{atm} \) and the difference is observed between 30 to 25 \( \mu \text{atm} \). Nevertheless some points with higher value can be identified (in red in Fig. 11). These values are explained by the position of these points, which are at the edges of the cloud, therefore more likely to include outliers that disturbed the estimation of the MLR coefficients. For the reconstruction of \( p\text{CO}_2 \), with
this identifiable point, it is quite easy to organize a system of flag. This flag can give an information about the quality in function of the position of the neurons. The difference obtained for pCO$_2$ in each neuron range from 0 to 56 µatm (Table 2), but 58% of the values observed are under 30 µatm of difference. The difference for each parameters can be quite high for some parameter like SST with a maximum value of 1.9 °C, but most of the value are lower than 1 °C and CDOM range between 0 and 5.15 (Table 2). The other parameters have quite low variability like MLD range between 0 and 9.7 m. The average is between two or three time lower than the maximum value observed, which give low value for all the satellite parameters.

The validation data set give a quite good correlation ($R = 0.93$) with the reconstruction method (Fig. 12), the root mean square (RMS) is 36.7 µatm. 12% have a value higher than 20 µatm and 45% between 20 µatm and 30 µatm (Fig. 13). The characteristic in time, SST, MLD, CDOM, Chl and NPP, do not explain the difference observed in the reconstruction.

A reconstruction has been done with the satellite data from 1997 to 2011. The seasonal cycle is well reproduced and in agreement with other studies. The maximum is observed during winter with 437 µatm in average and 274 µatm in summer. This values are comparable to the average estimated in central Baltic with 500 µatm in summer and 150 µatm in winter (Wesslander, 2011). In April, the pCO$_2$ decrease due to the biology and increase slowly in September (Fig. 14).

A simple flag was constructed to monitor the reconstruction quality. The flag is equal to 1 for classes where the average difference is lower than 20 µatm, equal at 2 for a difference between 20 µatm and 30 µatm and equal at 3 for higher average differences. In the example shown here the values are high (3) so the confidence on the reconstruction is low but some point have flag of 1 or 2 (Fig. 14d–f) and the reconstruction is more reliable. On the geographic map (Fig. 14d–f) the values of 4 correspond at the presence of ice which is estimated with the satellite product of the National Snow and Icea and Ice Data Center based on NOAA Level 3 products (Njoku, 2007).
The flag gives a confidence in our reconstruction, like in March 2010 (Fig. 14a), where the south of the map (Bornholm Basin and Arkona Basin) show lower $p$CO$_2$ value than the North and than February (not show here). This region in March 2010 corresponds to a flag of 2, which was attributed a medium confidence. In July 2010, the flag is quite good and the variability of $p$CO$_2$ seems in agreement with the monthly variability (Fig. 14b and e). In September 2010, the value of $p$CO$_2$ has a good order of magnitude when the flag is 2 but seems little to high when there is a poor confidence (3) (Fig. 14c and f).

In conclusion, the reconstruction of $p$CO$_2$ need to be improve to increase the confidence in our product, but this version seems in agreement with the evolution of the data.

4 Discussion and conclusions

In this paper, two methods previously used in separate studies relating to the reconstruction of the $p$CO$_2$ from satellite data in maritime regions, were combined to estimates the $p$CO$_2$ in the Baltic Sea. These methods, the Self Organizing Maps (SOM) and Multiple Linear Regression (MLR) were used to palliate the non-linearity of the mechanics driving the $p$CO$_2$ through the use of artificial networks, but retain a more detailed reconstruction than an average of a classification by using MLR in each class. The process involves classifying the explicative parameters (SST, CDOM, Chl, time, NPP and MLD) and then using the linear regression coefficients corresponding to that class in order to reconstruct the $p$CO$_2$. In the reconstruction obtained, by making use of the statistics obtained on each of the classes, is it possible to add a flag to each class to inform us on the quality of the reconstruction obtained. This could be important for the numerical modeling of other phenomena depending on the $p$CO$_2$ as well as the informed interpretation of the reconstructions obtained.

The current results with this method based on 1445, gave a high coefficient correlation of 0.93 % and a RMS of 36 µatm. In addition to having a limited amount of in-situ
$p\text{CO}_2$ measurements, the colocalized satellite data were frequently incomplete. This lead us to having to complete our database using a novel imputation method based on SOMs.

In comparison, existing studies performed over the North Atlantic and North Pacific, based on at a minimum 10000 data points (which take into account all the data from SOCAT) to at maximum 800 000 data (e.g. Friedrich and Oschlies, 2009; Landschützer et al., 2013; Hales et al., 2012). Friedrich and Oschlies (2009) had a RMSE error of 19 µatm. A similar study over the totality of the Atlantic Ocean gave a RMSE of 17 µatm for the independent time series (Landschützer et al., 2013). Hales et al. (2012) had a RMS deviation of 20 µatm with a correlation coefficient of 0.81. The RMSE obtained in our study were higher than the previous study in Atlantic Ocean, but, taking into account the highly reduced amount of data available, the results presented are promising.

The organisation of the values different coefficients of the MLR over each class ascertained that all the parameters from satellite data are important to reconstructing the $p\text{CO}_2$ in the Baltic Sea, even if it is for some particular cases. Therefore the improvement of the satellite data availability could also improve the performance of our reconstruction.

The methodological approach presented could potentially be further developed when reconstructing spatial fields of $p\text{CO}_2$ by taking the information of the classes attributed to the neighbouring grid points of geographic study area, when selecting the class whose linear regression coefficients to use in the reconstruction of the $p\text{CO}_2$. This aspect is however dependent on the acquisition of additional in-situ measurements.

There exist many programs in place for the acquisition of new data. The data at Östergarnsholm site are still forthcoming, the year 2012 did not give too much data but 2013 and 2014 need to be validated. The SMHI station also continue in time could be also add to our data. The VOS transect are not yet available for 2012 to 2014 but this measurement will also in continue, and some data will soon be available. There also some data from ferry boat from Gothenburg (Gothenburg–Kemi–Oulu–Lübeck–Gothenburg). The Gothenburg transect is weekly http://www.hzg.de/
The first test was made in 2010 and 2011 so some data could be soon available. New measurement of \( p\text{CO}_2 \) began in 2012 at the Utö Atmospheric and Marine Research Station http://en.ilmatieteenlaitos.fi/GHG-measurement-sites#Uto.

Given the amount of new data soon to be available we remain optimistic that the comprehension and statistical modeling of the \( p\text{CO}_2 \) in the Baltic Sea will continue to improve in the coming years.

**Acknowledgements.** The authors would like to thank Tiit Kutser and Melissa Chierici for their valuable help with data collection. We would like to thank Sylvie Thiria for her advising over the SOM classifications and Marion Leduc-Leballeur for her help with the method to fill the gap on satellite data. We would like to thank Swedish National Space Board (nr 120/11:3) for financing the project.

**References**


Table 1. Number of missing values for each parameter of the satellite data for the October–March and April–September periods. The number in parenthesis correspond of the total of point in each period.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>October–March (685)</th>
<th>April–September (814)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Chl</td>
<td>202</td>
<td>24</td>
</tr>
<tr>
<td>CDOM</td>
<td>320</td>
<td>5</td>
</tr>
<tr>
<td>NPP</td>
<td>468</td>
<td>571</td>
</tr>
<tr>
<td>MLD</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 2. Maximum and mean value observed in the difference between the data used for the trainee and the value in the neurons.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Maximum</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$CO$_2$ (µatm)</td>
<td>56.4</td>
<td>29.15</td>
</tr>
<tr>
<td>SST (°C)</td>
<td>1.9</td>
<td>0.98</td>
</tr>
<tr>
<td>time(cos)</td>
<td>0.33</td>
<td>0.07</td>
</tr>
<tr>
<td>Chl (mg m$^{-3}$)</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>time(sin)</td>
<td>0.4</td>
<td>0.06</td>
</tr>
<tr>
<td>CDOM</td>
<td>5.15</td>
<td>0.06</td>
</tr>
<tr>
<td>NPP (mg m$^{-3}$)</td>
<td>1.19</td>
<td>0.25</td>
</tr>
<tr>
<td>MLD (m)</td>
<td>9.7</td>
<td>3.2</td>
</tr>
</tbody>
</table>
Figure 1. Map of the Baltic Sea. The major basins and sea regions are named.

2. Cargo ship: this data set comes from continuous measurements of the surface water pCO$_2$. They take place 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 a depth of 5 m on board 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 taken between July 2003 to December 2005.

3. Swedish Meteorological and Hydrological Institute (SMHI) data set: pH (measured with the method from Grasshoff et al. (1999)) and alkalinity (TA) (measured by potentiometric titration described in Grasshoff et al. (1999)) are measured continuously at a monthly or semi-monthly resolution in the Baltic Sea at different stations. All data used are from 5 m depth below the surface. The uncertainty for the pH is ±0.03 pH units and for the TA is ±5% (Wesslander et al., 2009). pCO$_2$ is estimated from the pH, TA, salinity and temperature measurements and...
The relationship between $pCO_2$ and the environmental parameters available is highly non-linear. As mentioned in the introduction, we chose to combine two statistical approaches: self-organizing maps (Kohonen, 1990; Dreyfus, 2005) and linear regression. The SOM are a subfamily of the neuronal networks algorithms, used to perform multidimensional classification. One of the defining characteristics of the SOMs is their ability of their classes to locally represent a gaussian distribution centred around its typical profile of the environmental parameters. We used this hypothesis to classify the environmental parameters dataset, and then estimate for each class the parameters of a linear regression. In the following section we will present an overview of the two statistical algorithms and their application to our data sets.

Figure 2. Monthly data available from 1998 to 2011 in Baltic Sea. The colorbar described the $pCO_2$ value in µatm.
2.5.1 Self Organizing Maps

Self-organizing topological maps (SOM) is a clustering method based on neural networks. They provide a clustering of a learning data set into a reduced number of subsets, called classes, with common statistical characteristics.

The generation of a SOM requires the creation of a training database that contains homogenous vectors. After a training phase, we obtain a Self-Organizing Map. Each class is represented by its referent vector \( r_i \) which is an approximation of the mean value of the elements belonging to it and it index that positions it on the map in relation to the other classes (Figure 4). The topological aspect of the maps can be justified by considering the map as an undirected graph on a two-dimensional lattice whose vertices are the \( N \) classes. This graph structure permits the definition of a discrete distance \( d(C_i, C_j) \) between two classes \( C_i \) and \( C_j \), defined as the length of the shortest path between \( C_i \) and \( C_j \) on the map. The nature of the SOM training algorithm forces a topological ordering upon the map and, therefore, any neighbouring classes \( C_i \) and \( C_j \) on the map (\( d(C_i, C_j) = 1 \)) 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: \( \text{index} = \arg \max_i (||x - r_i||) \), 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, \( C_{\text{index}} \). In the case 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 \( r_i \).

Figure 3. Histogram showing number of observations for each month of the year.
Figure 4. The different elements of the training of the self-organizing map.

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 neighbours' 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 neighbouring vectors on the determination of the referent vector can be considered null. In these cases, each referent vector approximates, locally, the mean value of a multidimensional Gaussian random distribution that generated the training data assigned to that class (Dreyfus, 2005).

2.5.2 Multiple Linear Regression

A multiple linear regression is a modelling method that expresses the value of one response variable $y$ (in our study $pCO_2$) as a linear function of other explicative variables $X = X_1, X_2, ..., X_i$ (in our study SST, Chl, CDOM, NPP, MLD, time $\sin$, time $\cos$). The purpose of performing a multiple linear regression generally is either to interpret the relationship of the variable $y$ with 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, in order to perform our MLRs, we had to take into account their limitations and the nature of our problem. More specifically, in order to perform a 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 data sets. The $pCO_2$ does not follow a linear relationship with the variables presented when considering the entirety of the problem presented. However as noted above in sub-section 2.5.1, if we consider the classes created by the SOM, they are very localised 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 make the assumption that, if performed in the reduced neighbourhood of a SOM class, the relationship between $pCO_2$ and the explicative variables is linear.

3 Application and Results

3.1 Statistical imputation

As described in section 2.4, the data available for the application were presenting many missing values. In order to complete them we chose to use an imputation method 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 in order to regroup the data in typical situations and replace...
We first selected the database containing, as mentioned before: SST, Chl, CDOM, NPP, MLD, time \( \sin \) and \( \cos \). We then sorted its vectors according to the number of missing values in them and noting the placement of missing values for each vector. We chose all complete data vectors and the first 5% of the sorted vectors with missing data and trained a SOM. We proceeded by replacing the missing values of these initial 5% with the corresponding values of the referent vector of the class they each belonged to. We performed the same process by iteratively increasing by 5% the amount of vectors with missing values included in the training of a new SOM, and by replacing all missing values, even those completed in previous iterations, by the corresponding values of the referent vector of the class that each vector with missing values belonged to. A more detailed presentation of this method is the subject of another forthcoming article, but a schematic representation version of the imputation method used can be seen in Figure 6.

After this imputation of the missing data through the iterative training give a good representation of the data as presented in Figure 7. The repartition of \( pCO_2 \) (Figure 7,a) is well representative of the variability of the data with a large range of value. Some very high in particular local events like a coastal upwelling but most of the data are range between 180 µatm (value observed in summer) and 550 µatm (observed in winter). The SST (Figure 7,b) is well representative of the variability in the Baltic Sea with a maximum of observation between July and September in all the basin around 18°C (Siegel and Gerth, 2012). The NPP variability is quite homogenous except the peak at around 10

**Figure 6.** A schematic representation of the imputation method used.
Figure 7. $pCO_2$ and satellite data (SST, Chl, NPP, CDOM and MLD) available for the SOM after reconstruction.
3.2 pCO$_2$ estimation

3.2.1 Topological map

In this study we classified the explicative variables (SST, Chl, CDOM, NPP, MLD, time$_{sin}$, time$_{cos}$) into classes that share similar characteristics. We separated our data set in two parts: 90% of the completed data (1300 vectors) were used for the training phase, with the remaining 10% split into 5% (72 vectors) for testing and 5% for the validation of our method. We iteratively tested classifications with varying number of classes, and selected the parameters of our SOM based on the performances on the test dataset.

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 8). The order of magnitude of the number of observations is constant throughout the SOM, and we can consider that it has been well deployed to represent the data space of the explicatory 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 neighbours.

In order to estimate the average concentration of pCO$_2$ in each class, its measurements associated with vectors consisting of SST, CDOM, NPP, MLD and CHL components were presented to the already trained SOM as input data (figure 9). 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 of the SST distribution with low values of pCO$_2$ correlating with high value of SST (figure 9). This is in agreement with the seasonal cycle of pCO$_2$, which is characterized by a large amplitude, ranging from high value in winter ($\approx$ 500 µatm) and low value in summer ($\approx$ 150 µatm) Wesslander (2011). From Schneider and Kaitala (2006), the high winter value of pCO$_2$ are a consequence of mixing with deeper water layer enriched...
in CO$_2$ which is in agreement with the distribution of the MLD (figure 9,h) with the higher value during winter and autumn correlate with the high value of $p$CO$_2$. Wesslander (2011) explained also that it can be related to the mineralization, which exceed production in winter. Biological production starts in spring when sunlight and nutrients are sufficient. The chlorophyll begin to increase in March/April due to the spring phytoplankton bloom, which have for effect to reduce the $p$CO$_2$ during this period. The more intensive decrease take place in April and May which is consistent with the higher value of NPP (figure 9). The studies in the central Baltic Sea shows two summer minima, the first is in April/May and a second July /August which is resulting from a second production period. The higher variability are observed during this period with standard deviation between 39 µatm and 50 µatm from different region (Wesslander, 2011; Schneider and Kaitala, 2006).

3.2.2 Linear regression in the neurons

As mentioned in section 2.5.2, in order to perform a MLR we are obliged to assume that the relationship between the predictor variables and the response variable is linear, which we could only take as a valid hypothesis when performing the MLR in the reduced neighbourhood of a SOM class, where the relationship between $p$CO$_2$ and the explicative variables can be assumed to be linear.

**Figure 9.** Distribution of each parameter in the neural map. (a) $p$CO$_2$ in µatm (b) SST in °C (c and e) time respectively cosine and sine, (d) Chl $a$ in mg m$^{-3}$, (g) NPP in mg C m$^{-2}$ and (h) MLD in m.
For each class \( j \) there therefore we created a separate training data set containing all vectors that were assigned to that class and to all its adjacent classes. Based on that data-set we computed the linear regression coefficient parameters for every explicative parameter and a constant value. The calculated linear regression coefficient parameter values for each class are shown in figure 10. We can notice 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 demonstrated that each and every parameter is important in the reconstruction of the \( pCO_2 \) in the Baltic Sea, even if it is highly significant in some particular classes and relatively stable in other regions of the topological map.

The addition of vectors belonging to adjacent classes in general did not perturb the estimation of the coefficient parameters since, as seen in figure 9, the values of all parameters are generally organized in a coherent fashion on the map. The assumption that they are close in the data space is not as robust as it would have been had we solely considered the vectors belonging to each class but, given the limited amount of data available for the modeling of this highly non-linear and complex system, we would not have sufficient elements to correctly estimate the linear regression coefficients. Given the projected increase in available data in the coming years, further applications of this approach will limit themselves to the elements belonging to each class.

**Figure 10.** Coefficient from linear regression for each parameters. (a) SST in °C (b and d) time respectively cosine and sine, (c) Chl \( a \) in mg m\(^{-3}\), (f) NPP in mg C m\(^{-2}\) and (g) MLD in m.
Figure 11. (a) The colorbar represents the average difference in each neuron for $p$CO$_2$. (b) The colorbar represents the standard deviation (std) in each neuron for $p$CO$_2$. For (a) and (b) the number inside the neurons correspond at the number of data.
Table 2. Maximum and mean value observed in the difference between the data used for the trainee and the
value in the neurons

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Maximum</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p\text{CO}_2$ ($\mu$atm)</td>
<td>56.4</td>
<td>29.15</td>
</tr>
<tr>
<td>SST ($^\circ$C)</td>
<td>1.9</td>
<td>0.98</td>
</tr>
<tr>
<td>time(cos)</td>
<td>0.33</td>
<td>0.07</td>
</tr>
<tr>
<td>Chl (mg.m$^{-3}$)</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>time(sin)</td>
<td>0.4</td>
<td>0.06</td>
</tr>
<tr>
<td>CDOM</td>
<td>5.15</td>
<td></td>
</tr>
<tr>
<td>NPP (mg.m$^{-3}$)</td>
<td>1.19</td>
<td>0.25</td>
</tr>
<tr>
<td>MLD (m)</td>
<td>9.7</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Figure 12. $p\text{CO}_2$ reconstructed and measured for the validation data set in central Baltic with 500 $\mu$atm in summer and 150 $\mu$atm in winter (Wesslander, 2011). In April, the $p\text{CO}_2$ decrease due to the biology and increase slowly in September (figure 14).

A simple flag was constructed to monitor the reconstruction quality. The flag is equal to 1 for classes where the average difference is lower than 20 $\mu$atm, equal at 2 for a difference between 20 $\mu$atm and 30 $\mu$atm and equal at 3 for higher average differences. In the example shown here the values are high (3) so the confidence on the reconstruction is low but some point have flag of 1 or 2 (figure 14d,e and f) and the reconstruction is more reliable. On the geographic map (figure 14d,e and f) the values of 4 correspond at the presence of ice which is estimated with the satellite product of the National Snow and Ice Data Center based on NOAA Level 3 products Njoku (2007). The flag gives a confidence in our reconstruction, like in March 2010 (figure 14,a), where the south of the map (Bornholm Basin and Arkona Basin) show lower $p\text{CO}_2$ value than the North and than February (not show here). This region in March 2010 corresponds to a flag of 2, which was attributed a medium confidence. In July 2010, the flag is quite good and the variability of $p\text{CO}_2$ seems in agreement with the monthly variability (figure 14,b. and e.). In September 2010, the value seems in agreement with the monthly variability (figure 14,b. and e.). In September 2010, the value 18

Figure 12. $p\text{CO}_2$ reconstructed and measured for the validation data set.
Figure 13. Difference between $pCO_2$ reconstructed and measured for the validation data set.

Figure 14. a.b.c. Reconstruction of the $pCO_2$ map and d.,e. and f. The flag for each map in a.,d. March 2010, b.,e. July 2010 c.,f. September 2010. The flag value correspond to: 1: high confidence, 2: medium confidence 3: poor confidence.

The $pCO_2$ has a good order of magnitude when the flag is 2 but seems little to high when there is a poor confidence (3) (figure 14,c. and f.).
**Figure 14.** (a–c) Reconstruction of the $p$CO$_2$ map and (d–f) the flag for each map in (a and d). March 2010 (b and e). July 2010 (c and f) September 2010. The flag value correspond to: 1: high confidence, 2: medium confidence, 3: poor confidence.