Investigating the complex relationship between *in situ* Southern Ocean pCO_2 and its ocean physics and biogeochemical drivers using a nonparametric regression approach

Wesley Pretorius[1,2] and Sonali Das[2]

[1] Department of Statistics and Actuarial Sciences, Stellenbosch University,

Stellenbosch, 7600, South Africa

[2] Decision Support and Systems Analysis Research Group,

CSIR Built Environment, P.O.Box 395, Pretoria, 0001, South Africa

Pedro M.S. Monteiro

Ocean Systems and Climate Group, CSIR-CHPC , Rosebank, 7700,

Cape Town, South Africa

September 27, 2012

Abstract

The objective in this paper is to investigate the use of a non-parametric model approach to model the relationship between oceanic carbon dioxide (pCO_2) and a range of biogeochemical *in situ* variables in the Southern Ocean, which influence its *in situ* variability. The need for this stems from the need to obtain reliable estimates of carbon dioxide concentrations in the Southern Ocean which plays an important role in the global carbon flux cycle. The main challenge involved in this objective is the spatial sparseness and seasonal bias of the *in situ* data. Moreover, studies have also reported that the relationship between pCO_2 and its drivers is complex.

As such, in this paper, we use the nonparametric kernel regression approach since it is able to accurately represent the complex relationships between the response and predictor variables using the *in situ* data obtained from the SANAE49 return leg journey between Antarctic to Cape Town. To the best of our knowledge, this is the first time this data set has been subjected to such analysis. The model variants were developed on a training data subset, and the 'goodness' of the models were assessed on an "unseen" testing subset. Results indicate that the nonparametric approach consistently captures the relationship more accurately in terms of MSE, RMSE and MAE, than a standard parametric approach (multiple linear regression). These results provide a platform for using the developed nonparametric regression model based on *in situ* measurements to predict pCO_2 for a larger spatial region in the Southern Ocean based on satellite biogeochemical measurements of predictor variables, given that satellite measurements do not measure pCO_2.

Keywords: Southern Ocean; Carbon Flux; Nonparametric Regression; SANAE; Carbon Dioxide; Prediction

1 Introduction

Motivated by the need to quantify the changing role of the Southern Ocean in terms the global carbon budget, in this study we use a nonparametric kernel regression approach to model the relationship between Southern Ocean *in situ* partial pressure of carbon dioxide (pCO_2) using other *in situ* drivers such as sea surface temperature, mixed layer depth, salinity, chlorophyll concentration and altimetry. Variants of the model are compared. Given that ocean pCO_2 cannot be measured by satellites sensors, this investigation is a step towards developing a model that captures the *in situ* relationship between pCO_2 and its drivers, as a first step in predicting pCO_2 based on satellite-derived observations of the same proxy variables for a larger region in the Southern Ocean. CO_2 gas in the atmosphere is considered to be one of the leading causes of global warming as due to the increasing emissions of anthropogenic CO_2 and associated trapping of outgoing long-wave radiation produced by the Earth's surface (Sarmiento & Gruber 2002). However, the build-up of CO_2

in the atmosphere, is less than half the rate at which CO_2 is being produced by humans (8.5Gt C y^{-1}) (Le Quéré et al. 2007). The main reason for this are the oceanic and terrestrial sinks of CO_2, which presently take up about 50% of CO_2 emissions approximately equally: ocean: 25% (2.2Gt C y^{-1}); terrestrial (2Gt C y^{-1}) (Sarmiento & Gruber 2002).

The Southern Ocean is both a major sink of anthropogenic CO_2 (1Gt C y^{-1} or half the ocean sink) as well as a major influence in the much larger, but until now, balanced exchange of natural CO_2 between the ocean and the atmosphere (). Modeling data has suggested that the magnitude of this CO_2 sink may be changing as a result of a number of factors, which include increased upwelling of Circumpolar Deep Water, increased acidification (Revelle Factor) and reduced primary productivity (Takahashi et al., 2012). Being able to quantify the annual change in this flux is a potentially critical contribution to the attribution of long term trends in atmospheric CO_2. The scientific challenge of the global ocean CO_2 community is to reduce the uncertainty of observations based annual mean CO_2 flux from the present 40% to close to 10% which is required to resolve interannual changes in the magnitude of that sink (Monteiro et al., 2010). The relatively sparse density of observations in the Southern Ocean as well as their strong seasonal bias towards summer season means that the Southern Ocean is a major contributor to the uncertainty in global mean annual ocean CO_2 fluxes (Lenton et al., 2012). Presently, the only way to address this is through the use of empirical models capable of linking remote sensing proxy variables to pCO_2 (). Recent global efforts in this area have focused on MLRs and SOMs approaches which have had some success in the data rich North Atlantic Ocean (). This study investigates a non-parametric approach to developing a low uncertainty relationship between pCO_2 and its main physical and biogeochemical proxy variables that can also be derived from remote sensing observations.

This area accounts for $\pm$10% of the entire global ocean, while contributes more than 20% of the annual uptake of CO_2. In addition, the estimated air-sea flux of CO_2 obtained indicated a large CO_2 sink occurring in the Southern Ocean between the latitudes 40^oS and 60^oS and is considered of high importance due to its ability to regulate a large portion of the flux of CO_2, and hence is considered a major car-

bon sink. However, relative to the northern hemisphere, empirical understanding in the Southern Ocean is new, as *in situ* measurements have been limited, depicted in Figure 1, and also fairly recent. Another problem is the seasonal bias of the measurements obtained in the Southern Ocean since *in situ* measurements are generally restricted in this region to summer months (Schlitzer 2002).

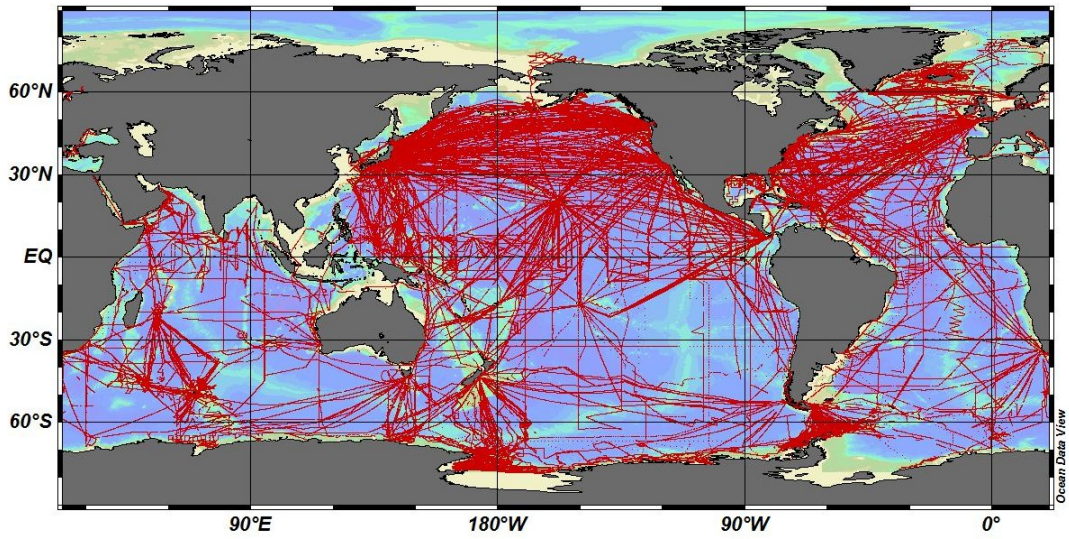

Figure 1: Location of LDEO V2009 master database of sea surface pCO_2 observations
(Takahashi et al. 2009)

The rest of the paper is organized as follows: In Section 2, we discuss in detail the *in situ* data used in this investigation; in Section 3 we present the nonparametric kernel regression approach and compare it with the parametric approach; in Section 4 we present the results and discuss the findings; while Section 5 concludes the paper.

2 Data

The data used in this analysis was taken from the 2009-2010 journey of the South African National Antarctic Expedition (SANAE)49 ship traveling on its return leg from Antarctica to Cape Town. The data collected over this stretch has, to the

best of our knowledge, never been subjected to statistical techniques and modelling as is done in this paper. The novelty of this paper also lies in the application of an understandable nonparametric method applied to this data, moving away from the black box self organising maps (SOMs) of Telszewski *et al.* (2009) (Telszewski et al. 2009). *In situ* measurements of the properties of the SO, including pCO_2,[1] sea surface temperature (SST), mixed layer depth (MLD), salinity, chlorophyll-a concentration and latitude were collected. The data used in this investigation will be referred to henceforth as SANAE49-L6. Inconsistencies in some of variables of interest existed due to possible faulty measurements around the 60^oS, 50^oS and 40^oS latitude lines, which were removed. Further, observations north of 37^oS and south of 70^oS were disregarded to eliminate terrestrial effects and to match the available range of MLD respectively. The final part of the SANAE49-L6 data set used consisted of 6103 observations in 6 variables, that spanned 13 February 2010 to 21 February 2010, and were within the GPS co-ordinates of (69.5998^oS, $5{,}9036^oW$) and (37.0004^oS, 12.918^oE).

Table 1: Descriptive Statistics for the final SANAE49-L6 data set

Variable	Means	Standard Deviation	Coefficient of Variation
pCO_2	360.19	37.72	0.105
Salinity	34.16	0.55	0.016
Chlorophyll Conc.	1.16	1.23	1.066
Intake Temp.	6.29	5.68	0.903
MLD	61.58	24.92	0.405

Figure 2 graphically displays the observed pCO_2 plotted against latitude (negative latitude values indicate degrees of latitude below the equator). This indicates the large spatial variability of pCO_2 between Cape Town and Antarctica. The models developed in this study attempt to accurately capture this variability by using other variable measured *in situ* on the SANAE49 ship.

[1] All pCO_2 values referred to are the partial pressure of CO_2 measured in the ocean surface. All models are fit with the pCO_2 values as response. This is done since the atmospheric pCO_2 is known to remain relatively constant over seasons and geographical space when compared to the variability of surface water (sea water) pCO_2 and the flux of pCO_2 in the ocean can therefore be identified as being driven by the sea water pCO_2 (Sarmiento & Gruber 2002, Takahashi et al. 2002, Jamet et al. 2007, Telszewski et al. 2009).

Table 2: Five Number summary for SANAE49-L6 data set

Variable	Minimum	Q1	Median	Q3	Maximum
pCO_2	251.19	351.20	368.62	380.18	435.98
Salinity	33.36	33.82	33.98	34.18	35.69
Chlorophyll Conc.	0.12	0.46	0.62	1.44	5.14
Intake Temp.	-0.28	2.65	3.61	8.37	21.30
MLD	13.15	42.08	55.85	82.45	127.93

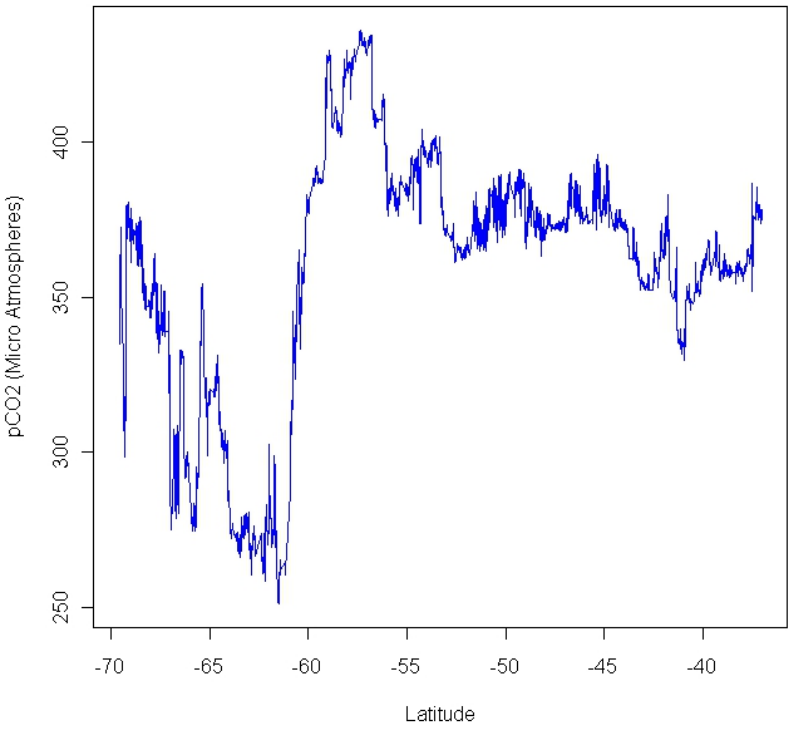

Figure 2: Line plot of pCO_2 versus latitude

3 Methodology

The relationship between the response variable in this study, namely pCO_2, and the predictor variables is known to be complex. Not only do changes in the fluxes of pCO_2 occur over time with regards to seasonality, but changes can be observed to occur spatially as well (Sarmiento & Gruber 2002, Tréguer & Jacques 1992). Parametric models have been proposed, mainly in the north Atlantic region, that include the spatial position of the measurements and SST in order to define a model for estimating pCO_2 values for a given set of input variables (Rangama 2005). Other models attempted to include other predictor variables such as MLD in order to account for vertical mixing in the ocean (Lüger et al. 2004). The issue with each of these approaches, however, is the fact that they are confined to small regions of oceanic activity, defined usually, by bio-geochemical provinces.

In this paper, we attempt to move away from parametric modeling to the nonparametric modeling framework to predict pCO_2 in terms of predictor variables in the SO. Nonparametric modeling allows for flexibility specifically with respect to accommodating nonlinear relationships that are considered to be complex in nature. An advantage of nonparametric regression over parametric regression is that no prior form of the regression equation needs to be specified. The data rather determines the relationships in the model. This unspecified regression function allows the nonparametric estimation methods to identify certain structures in the data which would not otherwise be identified by traditional parametric methods due to its strict assumptions (Racine & Li 2004). This makes nonparametric regression methods well suited to estimate nonlinear functions which may not follow a known parametric distribution (Fox 2005). A second benefit of the use of nonparametric estimation methods for the modeling is the range of options available such as local polynomial regression, regression splines and nearest neighbourhood methods. In this paper we focus on the kernel regression approach, however, other approaches such as local polynomial models and splines could also be applied in order to obtain a nonparametric model for the data (Racine & Li 2004, Li & Racine 2004).

The primary disadvantage of nonparametric models is its curse of dimensionality constraint since local methods, such as kernel regression, require large data sets

in order to obtain consistently accurate estimates. This is especially true in multivariate analyses, where the size of the data required in order to obtain the estimates of the same accuracy increases exponentially as the number of predictor variables in the model increases (Fox 2005), and can be attributed to the decreasing number of observations falling within a fixed local region around each possible input vector. In our analysis, however, the size of SANAE49-L6 is well suited for the nonparametric regression application.

3.1 Nonparametric Kernel Regression

In this section we present the details of nonparametric kernel regression modelling. In general, a regression function describes the average value (or conditional average) of a real valued response variable y, as a function of one or more predictor variables $x_1, x_2, ..., x_p$. This implies that the focus is to determine a function $g(x_1, x_2, ..., x_p)$ which estimates the conditional mean of the response y, i.e. $\mu_y|x_1, x_2, ..., x_p$ as given below:

$$\mu_y|x_1, x_2, ..., x_p = g(x_1, x_2, ..., x_p). \tag{1}$$

In simple linear regression $g(.)$ is a linear function of the input variables of the form:

$$\hat{g}(x_1, x_2, ..., x_p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{2}$$

In this case, it is usually assumed that the conditional distribution of y given the input variables is a Gaussian distribution with expectation $\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p$ and constant variance σ^2. These assumptions are restrictive and may be inaccurate in certain real world application (Fox 2005).

In the nonparametric regression approach, the only assumption made is that the function describing the relationship between the conditional mean response and the input variables is a smooth function and that it exists. However, as mentioned earlier, relaxing the stringent assumption of linearity comes at a price, in the form longer computational time as well as a loss in the simplicity of results. The advantages, however, are potentially more accurate models for estimating the response. Critics of the method point to the lack of a pre-defined regression function as a

disadvantage, which in the light of adequate data allows the data to define its own model, which summarises the information in the data effectively (Fox 2005).

An intuitive description of nonparametric kernel regression is that it defines a function $g(x_1, x_2, ..., x_p)$ as an empirically weighted average of responses corresponding to observed sets of predictor variables within a close neighborhood of a target input vector. This neighborhood is defined by the window widths, also known as the bandwidths, $\mathbf{h} = (h_1, h_2, ..., h_p)$. These bandwidths can also be considered to be smoothing parameters since larger values of individual result in a smoother function $g(x_1, x_2, ..., x_p)$, which can increase the bias resulting in an underfit model with a high test error rate. Similarly smaller values of the bandwidth values of h_i can result in a more variable function which may have a smaller bias resulting in an overfit model that may not be able to generalise well to unseen data sets.

We consider the nonparametric model for the univariate response variable Y and the p-dimensional input vector $\mathbf{X} = (X_1, X_2, ..., X_p)$ as follows:

$$Y_i = g(\mathbf{X}_i) + \epsilon_i, \qquad i = 1, 2, ..., n \tag{3}$$

where the error term ϵ_i is assumed to have a Gaussian distribution with mean 0 and constant variance σ^2 as is the case in regular linear regression. No explicit form of $g(.)$ is defined. We require estimates of the joint density function $f(\mathbf{x})$ of the input variables as well as the joint density function of the response y and the input vectors, i.e. $m(y, \mathbf{x})$. These we obtain using product kernel estimates of the density functions (Racine & Li 2004). The estimated joint density of the input variables is given by the average of the n kernel functions for each of the observed $\mathbf{X}$ vectors as follows:

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}(\mathbf{x}), \tag{4}$$

where the function $K_{h,i}$ is the product of p univariate kernel functions given by:

$$K_{h,i}(\mathbf{x}) = \frac{1}{h_1 h_2 ... h_p} \prod_{j=1}^{p} k\left(\frac{X_{ij} - x_j}{h_j}\right), \tag{5}$$

where the function $k(.)$ is a univariate, symmetric kernel which is a decreasing

function of the distance from the target input value.

The estimated joint density function of y and the input variables is similar to this and is given by:

$$\hat{m}(y, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_y} k\left(\frac{Y_i - y}{h_y}\right) K_{h,i}(\mathbf{x}). \tag{6}$$

The estimate of the function $g(\mathbf{x}) = \mathrm{E}[\mathrm{Y}_i|\mathbf{X}_i = \mathbf{x}]$ is then:

$$\hat{g}(\mathbf{x}) = \frac{\int y\ \hat{m}(y, \mathbf{x}) dx}{\hat{f}(\mathbf{x})}. \tag{7}$$

By estimating the above integral using observed information and calculating the empirical average of the Y_i multiplied by the estimated density $\hat{m}$(y,$\mathbf{x}$) and noting that $\int k(v)dv = 1$ and $\int vk(v)dv = 0$ the estimated function $\hat{g}(\mathbf{x})$ can be written as:

$$\hat{g}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_{h,i}(\mathbf{x})}{\sum_{i=1}^{n} K_{h,i}(\mathbf{x})}. \tag{8}$$

Equation 8 is a weighted average of the responses corresponding to the input vectors surrounding the target input vector $\mathbf{x}$. The weights are defined by the product kernel $K_{h,i}(\mathbf{x})$ and are therefore symmetric and decreasing with respect to the distance between the target input vector and the observed $\mathbf{X}_i$ (Racine & Li 2004).

To define the choice of kernel function and the bandwidths for the respective variables, it must be noted that as long as the kernel function attributes higher weights to those observations closer to the target vector $\mathbf{x}$, and that the weights decrease symmetrically as the distance between the target and the weighted value increases, the specific choice of kernel function in the model is not critical. The choice of optimal bandwidths, however, is important. This is done using the *R* package *np*, and in particular the function *npregbw* which uses leave-one-out cross-validation to determine the optimal bandwidths. The function *npreg* is used to obtain the fitted regression function (Racine & Li 2004, R Development Core Team 2011). We first define the leave-one-out kernel estimator of the joint density function of the input

vectors as:

$$\hat{f}_{-i}(\mathbf{X}_i) = \frac{1}{n} \sum_{j \neq i} K_{h,j}(\mathbf{X}_i). \tag{9}$$

$\hat{f}_{-i}$ defines the estimated joint density function of the input variables for the training data set omitting observation i. This is done for each of $i = 1, 2, ..., n$ where n is the number of observations in the training data set. Using this, we can obtain an estimate of the response for the i^{th} input vector using the nonparametric kernel regression estimate based on the data set which omits the i^{th} observation. This estimate is denoted $\hat{g}_{-i}(\mathbf{X}_i)$ and given by the formula:

$$\hat{g}_{-i}(\mathbf{X}_i) = \frac{\frac{1}{n} \sum_{j \neq i} Y_j K_{h,j}(\mathbf{X}_i)}{\hat{f}_{-i}(\mathbf{X}_i)}. \tag{10}$$

The leave-one-out cross-validation technique for choosing the optimal values of the bandwidths is chosen to solve the following minimization problem with regards to the mean square error:

$$\min_{h_1, h_2, ..., h_p} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(\mathbf{X}_i))^2 \right). \tag{11}$$

The *npregbw* function performs an iterative procedure of minimizing this function of the bandwidth values in order to determine the combination of values which provide us with the minimum cross-validation mean square error. This is a computationally intensive process, especially with large data sets since the model has to be fit and assessed for each observation in the data set that is being omitted. This, as mentioned before, is the cost of the nonparametric modelling, which however will be shown to be outweighed by the improvement in the model's predictive ability. (Racine & Li 2004, Li & Racine 2004, R Development Core Team 2011).

3.2 Details of the nonparametric models compared

This section provides some insight into the models used in this analysis, as well as variations in each of the models, in order to determine the best set of input variables to describe the response of interest. In order to test the generalising ability of the

developed model, the SANAE49-L6 data set is divided randomly into 2 subsets. The first subset is used as a training data set to estimate the regression function (in nonparametric kernel regression framework) or to estimate the regression parameters (in the MLR framework). The second subset is used to assess the ability of the models to predict the responses in this "unseen" part of the data.

Recall that the nonparametric models apply a local constant regression fit with fixed bandwidths which are estimated using leave-one-out cross-validation as discussed. The number of predictor variables used in the models vary and the reasons for their respective uses are discussed later. The kernel function applied is the Epanechnikov kernel which has the functional form:

$$k(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1) \quad ; \quad I(|u| \leq 1) = \begin{cases} 1, & \text{if } |u| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

A global multiple linear regression (MLR) model is also fit and assessed in order to compare results with the nonparametric regression approach. This provides a point of reference for the performance of the model in terms of predicting "unseen" data. The MLR models consist of the same response and predictor variables as the nonparametric models and include only the main effects of the variables as was performed in papers by Lefévre et al. (2005) and Jamet et al. (2007).

In total 4 variations of nonparametric and MLR models are fit on the training subset and are assessed on the test data. These models differ in the independent variables used and the ration of the data splits, as described in Table 3.

Model M1 includes all the drivers in SANAE49-L6. The salinity of the ocean was identified by domain specialists as not being (as of yet) globally and reliably available via satellite, even though it has been identified as an important determinant in sea water carbon levels (Goyet & Davis 1997, McNeil et al. 2007, Takahashi et al. 1981). Model M1 is also applied with varying divisions of the training and test subsets in order to determine the effect of the amount of data in the training subset on the performance of the nonparametric model. These models are referred to as model M1.1–M1.4 and represent percentage divisions of 70-30, 60-40, 50-50 and 40-60 respectively where the first figure describes the percentage of the data allocated

Table 3: Model description and division of data sets

Model	**Variables Included**	**Training Test Split**
M1.1	Salinity, Ch.conc, Intake Temp, MLD	70-30
M1.2	Salinity, Ch.conc, Intake Temp, MLD	60-40
M1.3	Salinity, Ch.conc, Intake Temp, MLD	50-50
M1.4	Salinity, Ch.conc, Intake Temp, MLD	40-60
M2	Salinity, Ch.conc, Intake Temp, MLD, Latitude	70-30
M3	Ch.conc, Intake Temp, MLD, Latitude	70-30
M4	Ch.conc, Intake Temp, MLD	70-30

to the training subset and the latter, the percentage in the test subset. Model M2 includes the latitude co-ordinate of the measurements as a predictor variable in an attempt to provide further information to the model of the positional correlation of the measurements. Models M3 and M4 omit salinity from the models, since models including salinity will not be useful when applied to satellite data as reliable estimates of salinity are not yet globally available. Model M4 excludes latitude as an independent variable in the model in order to avoid the curse of dimensionality and since it may cause the model to act as an interpolation method on the ships course rather than describing the true relationship between the pCO_2 and other independent variables.

4 Results

The *in situ* underway pCO_2 observation data set used in this study shows a large scale spatial variability between Cape Town and Antarctica that is typical of summer

(Fig 2). Its main features are:

1. The relatively low pCO_2 of the Sub-Antarctic Zone (40 - 45^{o}S) sustained by elevated primary productivity.

2. The strong outgassing of CO_2 (pCO_2 >385) between the Polar Front and the southern Boundary of the Antarctic Circumpolar Current (50 - 58^{o}S) sustained by upwelling of upper Circumpolar Deep Water (uCDW).

3. The strong undersaturation (ingassing) in the eastern Weddell Gyre sustained by summer primary productivity stimulated by stratification sustained by melting of sea-ice.

4. Upwelling of lower Circumpolar Deep Water (lCDW) in the southern half of the eastern Weddell Gyre.

Collectively these regimes define not just sharp transitions but also strong non-linear characteristics of variability that pose rigorous tests for linear and non-linear empirical models.

The nonparametric models were fit to a randomised training subset to obtain the optimal bandwidths for the nonparametric models as defined in Table 3. Table 4 presents the results obtained from fitting the models and predicting the responses in the test subset.

These optimal bandwidth values are determined using cross validation as described earlier, and they allow the model to determine the "local neighbourhood" for each input variable over which the weighted average of the responses is taken in order to provide a predicted response. Larger bandwidth values imply a larger neighbourhood necessary for that variable for the local averaging method to accurately estimate the response.

Table 5 indicates the model results of the multiple linear regression (MLR) model fit. The table provides the regression parameter estimates (least squares estimates) with their respective p-values for the null hypothesis that the parameter values is equal to 0 versus the alternative that it is significantly different from 0.

Table 4: Cross-validated bandwidths for nonparametric kernel regression models

Model	**Bandwidths**				
	Salinity	**Chlorophyll-a**	**SST**	**MLD**	**Latitude**
M1.1	0.0366	0.5559	0.6084	2.4668	
M1.2	0.1175	0.0921	0.1882	7.7211	
M1.3	0.1017	0.4293	0.1737	2.9488	
M1.4	0.1017	0.4311	0.1682	2.9488	
M2	0.0697	0.4293	0.1737	5.2792	0.3626
M3		0.9955	0.2501	3.3325	0.0329
M4		0.1368	0.2079	2.1233	

The models in Table 5 are all fitted with an intercept term. Positive parameter values indicate variables that have a positive (direct) relationship to the response variable, while negative parameter values indicate an inverse relationship. A significance level of 0.05 (5%) is used to determine whether variables are considered significant or not.

The models in Table 5 were used to predict the responses for the input variables in the "unseen" test data sub-sets. This gives an indication of how well the models are able to predict new data. Figures 3 to 9 plot the observed response values (blue dots) along with the predicted responses from the parametric (red) and nonparametric models (purple), versus latitude, for each of the models in Table 3 on the left-hand-side. On the right-hand-side, the deviations of the model predictions from the observed values are plotted for the parametric (red) and nonparametric (purple) models.

Table 5: Multiple linear regression model summary

	Regression Parameter (p-value)					
Model	**Intercept**	**Salinity**	**Chlorophyll-a**	**SST**	**MLD**	**Latitude**
M1.1	-11.37 (0.773)	12.32 (< 0.0001)	-29.26 (< 0.0001)	-3.39 (< 0.0001)	0.103 (< 0.0001)	
M1.2	-23.08 (0.587)	12.66 (< 0.0001)	-29.27 (< 0.0001)	-3.44 (< 0.0001)	0.106 (< 0.0001)	
M1.3	1.21 (0.979)	11.97 (< 0.0001)	-29.69 (< 0.0001)	-3.42 (< 0.0001)	0.097 (< 0.0001)	
M1.4	-27.14 (0.603)	12.82 (< 0.0001)	-29.54 (< 0.0001)	-3.53 (< 0.0001)	0.100 (< 0.0001)	
M2	-793.99 (< 0.0001)	40.76 (< 0.0001)	-24.83 (< 0.0001)	-9.48 (< 0.0001)	0.013 (0.352)	2.82 (< 0.0001)
M3	432.41(< 0.0001)		-28.80 (< 0.0001)	-2.81 (< 0.0001)	0.017 (0.257)	0.41 (< 0.0001)
M4	405.84 (< 0.0001)		-29.41 (< 0.0001)	-2.29 (< 0.0001)	0.049 (0.0004)	

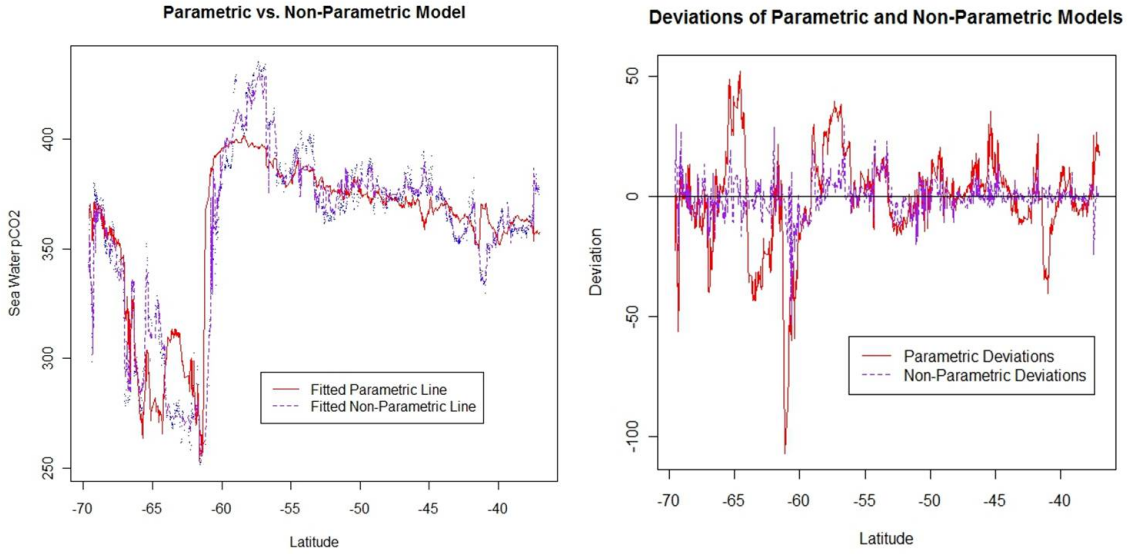

Figure 3: Predicted responses and deviations from true response of the test data set for parametric and nonparametric M1.1

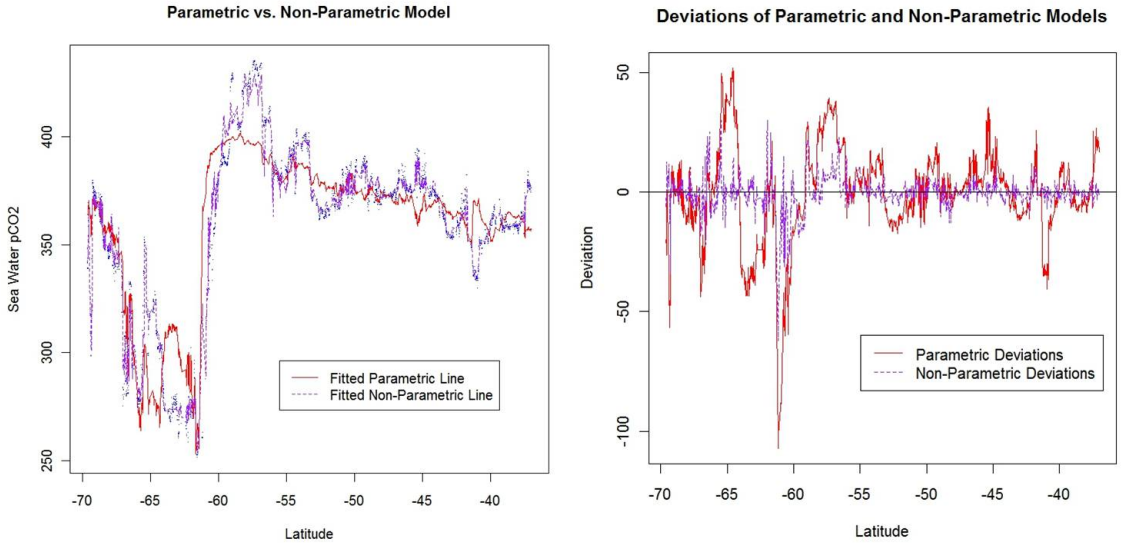

Figure 4: Predicted responses and deviations from true response of the test data set for parametric and nonparametric M1.2

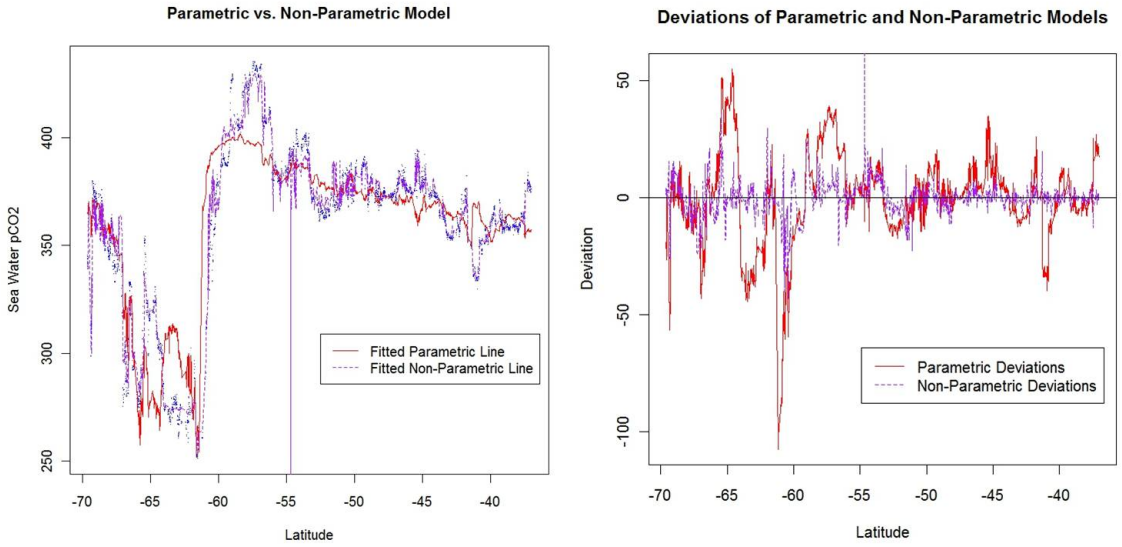

Figure 5: Predicted responses and deviations from true response of the test data set for parametric and nonparametric M1.3

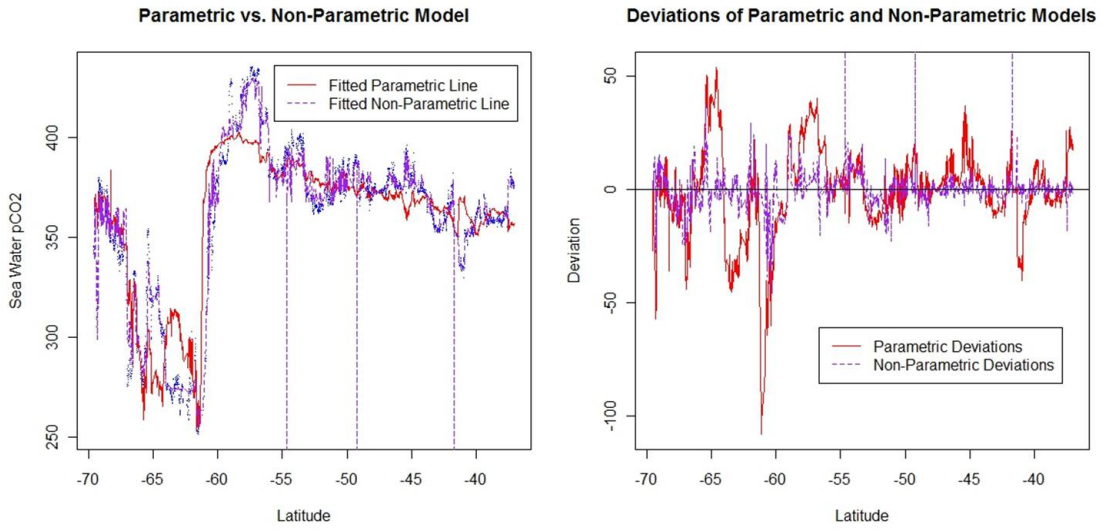

Figure 6: Predicted responses and deviations from true response of the test data set for parametric and nonparametric M1.4

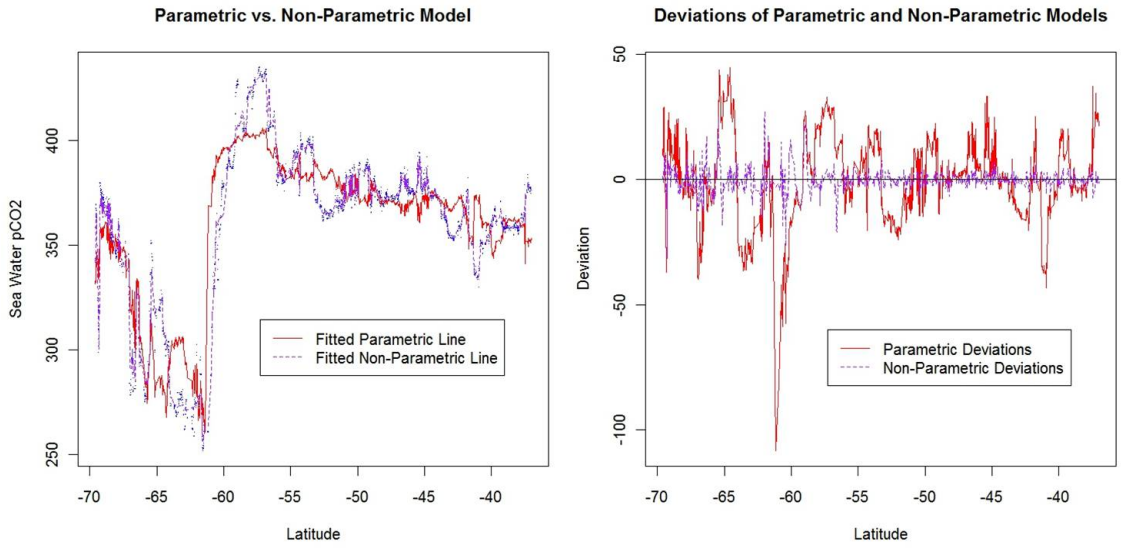

Figure 7: Predicted responses and deviations from true response of the test data set for parametric and nonparametric M2

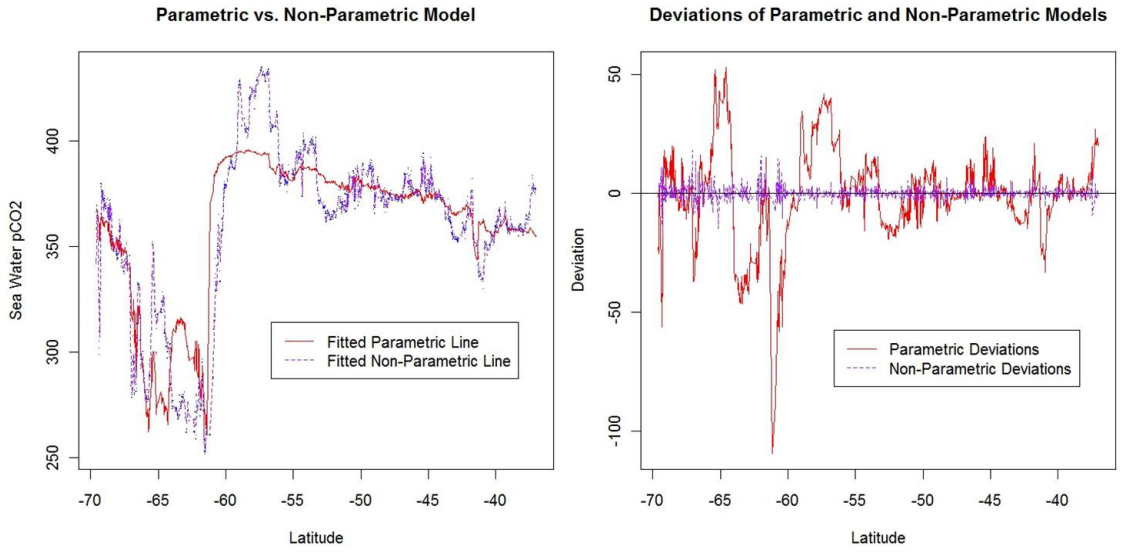

Figure 8: Predicted responses and deviations from true response of the test data set for parametric and nonparametric M3

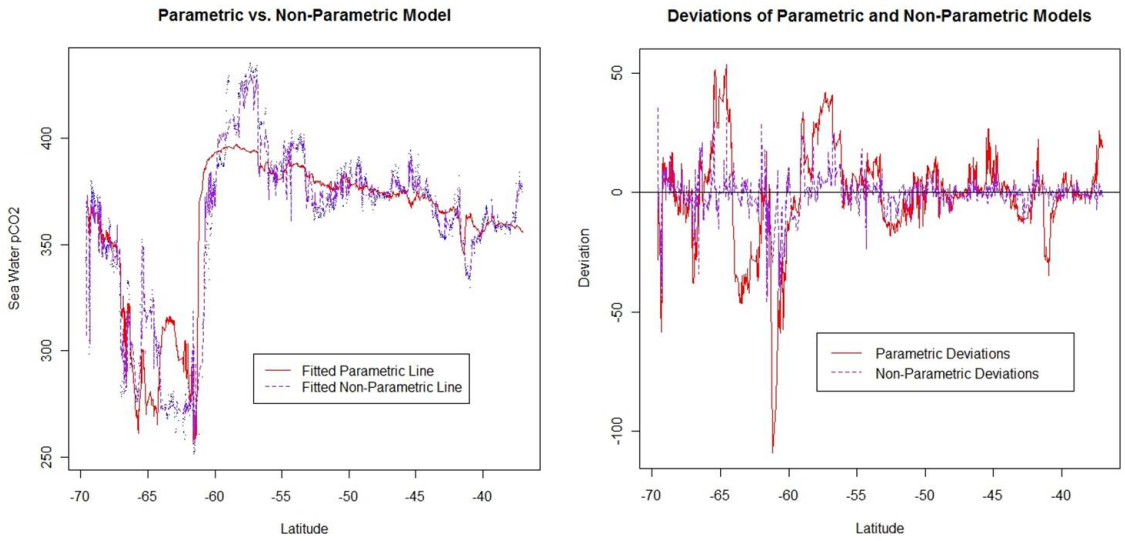

Figure 9: Predicted responses and deviations from true response of the test data set for parametric and nonparametric M4.

While Figures 3 to 9 provide a graphical comparison of the prediction ability of the respective models, the quantifiable comparison of the prediction ability of the two models is provided in Table 6 which compare the maximum underestimation, maximum overestimation, mean square error (MSE), root mean square error (RMSE) and the mean absolute error (MAE) for each of the models described in Table 3. [2]

[2]In table 6, MLR refers to the Multiple Linear Regression models, while NP refers to the Nonparametric Kernel Regression models.

4.1 Discussion

We now discuss the implications of the results and compare the parametric and nonparametric models. The optimised bandwidth values for the nonparametric kernel regression models shown in Table 4 are not directly comparable with one another since they are not independent of the units of measurement for the variables. What can be said about these values, however, is that in order to obtain useable, optimal bandwidths, it is necessary that the training data sets obtain sufficient data to overcome the curse of dimensionality.

The parameter estimates in Table 5 for the multiple linear regression models describe the proposed relationships between the independent variables and the response. The negative parameter estimates for SST and Chlorophyll-a concentration indicate a negative relationship between these predictor variables and the response while the remaining variables, salinity, MLD and latitude, indicate a positive relationship with the response.

Figures 3 to 9 show an improved accuracy in the prediction of pCO_2 in an "unseen" test data set by the nonparametric kernel regression model over the parametric MLR model in all areas of the ocean in terms of latitude. The MLR predictions deviate from the observed values of pCO_2 in the latitude areas slightly north of 65^oS, between 60^oS and 55^oS and as well as slightly south of 40^oS in each of the models. These anomalies highlight the limitations of the MLR and are critical sources of error because they exclude the main seasonal sources and sinks from this type of empirical treatment. The nonparametric predictions, however, do not present this deviation from the observed values, which indicates that these model are more likely to generalise well. The deviation plots in these figures indicate that the deviations in each of the MLR models M1.1 to M4 seem to be larger than the corresponding nonparametric models confirming that the predicted values using the MLR models tend to be further away from the observed values than the predictions from the nonparametric models.

Moreover, from Figures 5 and 6 it can be seen that as the amount of data in the training data set decreases, it creates a situation where there is less data which fall into the neighbourhoods of the input variables in the test data set. The consequences

Table 6: Generalisation measures for parametric and nonparametric models

	Maximum Underestimation		**Maximum Overestimation**		**MSE**		**RMSE**		**MAE**	
Models	**MLR**	**NP**	**MLR**	**NP**	**MLR**	**NP**	**MLR**	**NP**	**MLR**	**NP**
M1.1	52.08	29.76	-107.36	-53.40	327.51	52.14	18.10	7.22	12.81	4.87
M1.2	52.05	30.13	-107.33	-62.06	338.93	41.81	18.41	6.47	13.00	3.98
M1.3	54.93	377.93	-107.65	-43.18	343.71	96.47	18.54	9.82	13.04	4.70
M1.4	53.94	377.93	-108.04	-45.65	343.31	166.06	18.53	12.89	13.06	4.96
M2	44.78	27.27	-108.37	-31.58	300.79	21.77	17.34	4.67	13.15	2.94
M3	52.92	18.12	-109.63	-18.64	329.52	7.05	18.15	2.66	12.48	1.67
M4	53.50	35.21	-108.95	-46.39	331.81	64.18	18.22	8.01	12.44	4.84

of this are "bad" or unreliable estimates as can be seen near 55^oS (figures 5 and 6), 51^oS and 47^oS (figure 6 only). These predictions indicate a deteriorating prediction ability of the models as the size of the training data sets decrease allowing for larger areas in input space where no training data points are observed.

Table 6 indicates a much smaller bandwidth for the difference between the maximum under and over estimations for those nonparametric models which do not produce unreliable estimates discussed above. Ignoring models M1.3 and M1.4, the maximum range of errors for the nonparametric regression models was 92.190 micro-atmospheres, while the maximum range of errors for the MLR models was 162.547 micro-atmospheres. This represents a decrease in the range of the errors of more than 40%. The MSE and RMSE values in Table 6 are all smaller for the nonparametric models than for MLR models. The MSE and MAE values for the nonparametric models indicate a decrease of between 50%–98% and 60%–87% respectively from the MSE and MAE values for the MLR models. This further indicates that the nonparametric models are able to generalise to "unseen" data more effectively than the MLR models. The correlation coefficients (R^2) of the nonparametric models based on the training data set (not provided here) all were above 96% and were observed to be higher than the R^2 for the MLR models which were around 75%. This further supports the better fit of the nonparametric models than the MLR (parametric) models although the R^2 values alone may indicate an overfit model which would not generalise well to the "unseen" data.

The MLR approach taken can be compared to the application of MLR models in the North Atlantic performed by Jamet *et al.* (2007). The models developed in the North Atlantic were based on pCO_2 observations from the northern summer over the years 1994–1995. Three models were identified and used for predictions, the first of which used only SST as an independent variable. The second model includes the spatial co-ordinates of the measurements as independent variables while the final model replaces the co-ordinates with chlorophyll-a concentration and MLD. The respective mean square error rates for these models were 203.633μatm, 178.49μatm and 130.874μatm respectively. The MSE's obtained from the MLR models in the SO are much larger, indicating that the MLR is not able to capture the variability of the pCO_2 in the SO to the same extent as in the North Atlantic.

The NPKR approach in this paper can be compared to the SOM neural networks used by Telszewski *et al.* (2009) to predict pCO_2 along the lines of the volunteer observing ship (VOS) lines as well as where there are gaps in the remote sensing data. The SOM's were developed using the same independent variables as models M1.1–M1.4. The RMSE's obtained for the years 2004, 2005 and 2006 were 8.1μatm, 12.6μatm and 12.5μatm respectively for the models predicting remotely sensed data along the same lines as the VOS. The RMSE's obtained from the NPKR approach are very close to these values, however these errors are based on *in situ* test data. The maximum RMSE obtained from the NPKR models is 12.89μatm. The advantage of the NPKR models is the simplicity of its application. The SOM approach is a black box method which is complex and difficult to explain in a simple manner. The NPKR method, on the other hand, has a simple and logical methodology and can therefore be understood by those wanting to implement it in any domain.

5 Conclusion

In this paper, the aim was to investigate the relationship between *in situ* pCO_2 and corresponding biogeochemical predictor variables such as sea surface temperature, mixed layer depth, salinity, chlorophyll concentration and latitude from the Southern Ocean. Since such relationships have been reported to be complex, we used the nonparametric kernel regression approach, and compared the results to a parametric approach. The goodness of the model variants was tested by dividing the dataset randomly into a training subset and a testing subset, with the model being developed on training subset, and tested on the remaining subset. Results indicate that for the dataset used, the nonparametric kernel regression method for predicting *in situ* pCO_2 using *in situ* predictor variables provides consistently more accurate results than the parametric multiple linear regression counterparts. These results were expected due to initial exploratory data analysis which indicated that the distribution of pCO_2 was typically non-Gaussian for the dataset used, and therefore the assumptions of normality of errors would not hold.

Further research would involve taking the developed model to predict oceanic

pCO_2 for a larger spatial region in the Southern ocean based on satellite measurements of biogeochemical predictor variables. Note, satellite measurements do not measure pCO_2, and hence, predicting pCO_2 from satellite predictor variables based on a model developed based on complete *in situ* data, will provide a measure of oceanic pCO_2. The challenge however, is to measure the accuracy of such predictions, one being to compare the model based predicted pCO_2 using satellite predictors for the same region and time as the available *in situ* measurements. The major challenge in this approach is to resolve the scale issues, both temporal and spatial, when comparing satellite based predictions and *in situ* based predictions for pCO_2.

Acknowledgements

This work was supported by the CSIR Parliamentary and the Strategic Research Panel grants [TA_2010_035] in the ambit of the Southern Ocean Carbon Climate Observatory (SOCCO) programme. We would also like to thank Dr Nicolas Fauchereau for data facilitation. The authors declare that they have no conflict of interest.

References

Fox, J. (2005), Introduction to nonparametric regression. Economic and Social Research Council, Available at: http://socserv.mcmaster.ca/jfox/Courses/Oxford-2005/slides-handout.pdf, Accessed: 14 November 2011.

Goyet, C. & Davis, D. (1997), 'Estimation of total CO_2 concentration throughout the water column', *Deep-Sea Research I* **44**(5), 859–877.

Jamet, C., Moulin, C. & Lefévre, N. (2007), 'Estimation of the oceanic pCO_2 in the North Atlantic from VOS lines *in situ* measurements: Parameters needed to generate seasonally mean maps', *Annales Geophysicae* **25**, 2247–2257.

Le Quéré, C., Rödenbeck, C., E.T. Buitenhuis, T. C., Langenfelds, R., Gomez, A., Labuschagne, C., M. Ramonet, T. N., Metzl, N., Gillett, N. & Heimann, M. (2007), 'Saturation of the Southern Ocean CO_2 sink due to recent climate change.', *Science (New York, N.Y.)* **316**(5832), 1735–1738.

Lefévre, N., Watson, A. & Watson, A. (2005), 'A comparison of Multiple Regression and Neural Network techniques for mapping *in situ* pCO_2 data', *Tellus* **57B**, 375–384.

Li, Q. & Racine, J. (2004), 'Cross-validated local linear nonparametric regression', *Statistica Sinica* **14**, 485–512.

Lüger, H., Wallace, D. & Körtzinger, A. (2004), 'The pCO_2 variability in the midlatitude North Atlantic Ocean during a full annual cycle', *Global Biogeochemical Cycles* **18**(GB3023), doi:10.1029/2003GB002200.

McNeil, B., Metzl, N., Key, R., Matear, R. & Corbiere, A. (2007), 'An empirical estimate of the Southern Ocean air-sea CO_2 flux', *Global Biogeochemical Cycles* **21**(GB3011), doi:10.1029/2007GB002991.

R Development Core Team (2011), 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Racine, J. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', *Journal of Econometrics* **119**(1), 99–130.

Rangama, Y. (2005), 'Variability of the net air-sea CO_2 flux inferred from shipboard and satellite measurements in the Southern Ocean south of Tasmania and New Zealand', *Journal of Geophysical Research* **110**(C9), 1–17.

Sarmiento, J. & Gruber, N. (2002), 'Sinks for anthropogenic carbon', *Physics Today* **55**(8), 30.

Schlitzer, R. (2002), 'Carbon export fluxes in the Southern Ocean: results from inverse modeling and comparison with satellite-based estimates', *Deep-Sea Research II* **49**, 1623–1644.

Takahashi, T., Broecker, W. & Bainbridge, A. (1981), *The Alkalinity and Total Carbon Dioxide Concentration in the World Oceans*, New York, Wiley, chapter 16, pp. 271–286.

Takahashi, T., Sutherland, S. & Kozyr, A. (2009), 'Global ocean surface water partial pressure of CO_2 database: Measurements performed during 1957–2009 (version 2009)'. ORNL/CDIAC-152, NDP-088(V2009), Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, doi: 10.3334/CDIAC/otg.ndp088(V2009).

Takahashi, T., Sutherland, S., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R., Feely, R. & Sabine, C. (2002), 'Global sea-air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects', *Deep Sea Research Part II: Topical Studies in Oceanography* **49**(9–10), 1601–1622.

Telszewski, M., Chazottes, A., Schuster, U., Watson, A., Moulin, C., Bakker, D., González-Dávila, M., Johannessen, T., Körtzinger, A., Lüger, H., Olsen, A., Omar, A., Padin, X., Ríos, A., Steinhoff, T., Santana-Casiano, M., Wallace, D. & Wanninkhof, R. (2009), 'Estimating the monthly pCO_2 distribution in the North Atlantic using a self-organizing neural network', *Biogeosciences* **6**, 1405–1421.

Tréguer, P. & Jacques, G. (1992), 'Dynamics of nutrients and phytoplankton, and fluxes of carbon, nitrogen and silicon in the Antarctic Ocean', *Polar Biology* **12**(2), 149–162.