

Chapter 5

Methods and Results

5.1 Introduction

The main aim in this dissertation is to establish a relationship between oceanic properties that can be used to estimate $p\text{CO}_2$ in the ocean with a root mean square (RMS) error of less than 15% (less than approximately $60\mu\text{atm}$). The information about the physical processes that take place in the ocean are still insufficient to establish this relationship based on sound theoretical understanding. However, an empirical relationship between the ocean variables can be investigated without knowledge of the physical processes that take place in the ocean.

In order to investigate the empirical relationship, data from the SANAE49L6 cruise are used in this dissertation. This cruise was conducted, northward from Antarctica to Cape Town, in February 2010. The SANAE49L6 data set consists of 6103 measurements of oceanic properties for the part of the ocean that was covered by the cruise.

Least squares optimization is one of the methods that is used to establish a relationship between the ocean variables. Different linear equations are fitted to the data in order to find the minimal RMS error in the estimation of the $p\text{CO}_2$ in the ocean. Least squares optimization is carried out on a subset of sampled points from the complete data set and the RMS error is determined over the entire data set.

Another method that is used to investigate the relationship between the oceanic properties is radial basis function (RBF) interpolation. This is an interpolation method that makes use of a polynomial to fit the data and then enhances the polynomial function with a number of radial basis functions.

Local radial basis functions are used in this case, and several variations of the interpolation method is applied.

The sampling of the subset of points that are used to determine the equation is also explained in this chapter and the methods of sampling are explained and compared and the results are presented.

5.2 Sampling

The main aim of this project is to establish a relationship between the $p\text{CO}_2$ and other ocean variables using large data sets consisting of ocean data for the entire Southern Ocean. In this dissertation, the relationship for a particular part of the Southern Ocean is investigated. The *in situ* data sets contain oceanic $p\text{CO}_2$ data as well as data of other oceanic properties. The *in situ* measurements, however, are sparse in the Southern Ocean, and $p\text{CO}_2$ values for large regions in the Southern Ocean cannot be determined from *in situ* measurements. Satellite measurements are less sparse in the Southern Ocean, and measurements made by satellites are not disrupted by ice or clouds. Satellites, however, cannot measure the $p\text{CO}_2$ in the ocean. It is thus suggested that the relationship between $p\text{CO}_2$ and other satellite measurable properties be determined by using the *in situ* measurements. The relationship or equation that is determined from the *in situ* measurements can then be used to predict or estimate the $p\text{CO}_2$ values in the ocean using the more dense satellite data sets containing data of various other oceanic properties.

The CDIAC and SOCAT databases contain *in situ* data that is collected globally during boat cruises. It is a collection of the available data from the boat cruises that are conducted annually. These data sets are available on-line. These data sets have been validated and checked and the measurements have been proven to be accurate and can be used in ocean models. The part of the CDIAC data set that contains the Southern Ocean measurements contains approximately 1.6 million data points. The part of the SOCAT data set that contains Southern Ocean data contains approximately 8 million data points. To use all 1.6 million or 8 million points to establish the relationship between the $p\text{CO}_2$ and other variables for the Southern Ocean will require vast amounts of computer processing and high computational costs.

It is therefore advantageous to optimally select a subset of the data points that gives an accurate representation of the entire data set. This is not necessarily the data set with the best geographical distribution, but rather a selection of points that takes the variability of all the variables into account, and selects the optimal points from this. The main aim is to select the points in such a manner that the equation that is created from the selected points has as small an error as possible.

Furthermore, the cost of measuring oceanic properties at regular intervals in the ocean is high, and could be reduced if optimal points to measure the oceanic properties can be determined. These optimal sampling points can indicate positions for possible floats or buoys to take measurements of oceanic properties at regular intervals without the need of ships travelling in the area. The relationship between the oceanic properties can then be determined from these measurements.

There are several different ways in which this selection could be made. The one way is to randomly select points according to some performance criterion. Another method is to use a more structured method to select the points. The sampling is discussed in the following two sections.

5.2.1 Random sampling

The pCO_2 prediction error using the relationship established by a particular method, is influenced by the collection of points that is used to determine this relationship. If the subset of points selected from the complete data set is an accurate representation of the behaviour in the ocean as a whole, the prediction error will be small. If this is not the case, then the error of the estimation will be larger.

One way to select the points is by means of random sampling. In this case, a random selection of a number, k , of points from the complete data set is made. These randomly selected points are used to determine the relationship between the pCO_2 and other ocean variables. This method can yield a good selection of points, in the sense that the random points selected could be an accurate representation of the whole data set and could yield an equation that will yield small errors in predicting the pCO_2 . It could also yield a poor selection of random points, where the equation determined with these points will yield a larger error on the prediction.

Random sampling is an unreliable and unpredictable method of selecting the points, and there is no guarantee that the points selected in this manner will be adequate. Another more structured method is needed in this case. This is discussed in the following section. The random selection is used to quantify and compare the performance of the structured sampling method.

5.2.2 D-optimal sampling

The selection of an optimal set of points from the complete data set is crucial in determining a relationship that will make predictions of pCO_2 with as small as possible errors. D-optimal sampling selects the points in such a way that the best distribution in the search space of all possible points

will be sampled.

In D-optimal sampling, the objective is to maximise the determinant of the Fisher information matrix. The Fisher information matrix $\mathbf{M}$ is given by

$$\mathbf{M} = \mathbf{S}^T \mathbf{P}^{-1} \mathbf{S}, \quad (5.1)$$

where $\mathbf{S}$ is the sensitivity matrix and $\mathbf{P}$ is the variance-covariance matrix of the observation error [70].

In our case, the following system is considered:

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta}, \quad (5.2)$$

where, $\mathbf{y}$ is a vector of the $p\text{CO}_2$ values, $\mathbf{A}$ is the $m \times n$ matrix that contains m rows for the number of data points in the data set and n columns that each represent a term in the equation that is fitted to the data and $\boldsymbol{\beta}$ is an $n \times 1$ vector of the coefficients of these terms in the equation. The system is expanded as follows:

$$\begin{bmatrix} p\text{CO}_{2_1} \\ p\text{CO}_{2_2} \\ \vdots \\ p\text{CO}_{2_n} \end{bmatrix} = \begin{bmatrix} 1 & T_1 & \text{MLD}_1 & \text{Chl}_1 & \text{Lat}_1 & \dots \\ 1 & T_2 & \text{MLD}_2 & \text{Chl}_2 & \text{Lat}_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & T_n & \text{MLD}_n & \text{Chl}_n & \text{Lat}_n & \dots \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}. \quad (5.3)$$

In this case, there is a lack in information about the variance-covariance matrix of the observation error, $\mathbf{P}$. This is because the measurements are made at different points in the ocean and at different times with different conditions. In this dissertation it is assumed that the variance for all the data points are the same and that the covariance is zero everywhere, hence we can then write the matrix $\mathbf{P}$ as

$$\mathbf{P} = \sigma^2 \mathbf{I}, \quad (5.4)$$

where σ^2 is some constant variance on all the data points and $\mathbf{I}$ is the identity matrix.

The sensitivity matrix $\mathbf{S}$ can be written as

$$\mathbf{S} = \begin{bmatrix} \frac{\delta \mathbf{y}}{\delta \boldsymbol{\beta}} \end{bmatrix}, \quad (5.5)$$

where $\mathbf{y}$ is the model output and $\boldsymbol{\beta}$ is the vector of coefficients as in Equation (5.2).

Then the Fisher information matrix, $\mathbf{M}$, is given by:

$$\begin{aligned}\mathbf{M} &= \mathbf{S}^T \mathbf{P}^{-1} \mathbf{S} \\ &= \mathbf{A}^T \left[\frac{1}{\sigma} \mathbf{I} \right] \mathbf{A} \\ &= \frac{1}{\sigma} \mathbf{A}^T \mathbf{A}.\end{aligned}\tag{5.6}$$

In the D-optimal sampling in this application, the objective function for the optimization is given as follows:

$$\begin{aligned}\max(\det(\mathbf{M})) &= \max(\det(\frac{1}{\sigma} \mathbf{A}^T \mathbf{A})) \\ &= \max(\det(\mathbf{A}^T \mathbf{A})).\end{aligned}\tag{5.7}$$

By optimizing this objective function, the optimal distribution of points in the sample space will be selected, and the points obtained from this sampling will represent the variations in all the variables used in the particular equation. The volume of the covariance ellipsoid will be minimized with the optimization of the objective function. This optimization function is problem specific, and will be different for each different equation considered. This objective function ensures that the optimal distribution of all the variables will be selected in order to minimize the error in the pCO₂ estimations. The optimization method used to determine this optimal distribution of points is a genetic algorithm. This is explained in the next section.

5.2.3 Genetic algorithm application in D-optimal sampling

A genetic algorithm is used to select the optimal set of points to be used in the investigation of the relationship between pCO₂ and other ocean variables. The matrix $\mathbf{A}$ as described in Equations (5.2) and (5.3) is input to the genetic algorithm. $\mathbf{A}$ is an $m \times n$ matrix where m is the number of individual data points in the data set, in this case 6103 points, and n varies according to the equation that is considered. For example, for a linear curve fit with four variables, the equation will contain five terms and $n = 5$ in this case.

The genetic algorithm is set up to randomly select a number of lines (k) from the matrix $\mathbf{A}$, or similarly a number of points from the original ordered data set. This will be a corresponding set of points. The line i in matrix $\mathbf{A}$ will correspond to the data in line i in the original data set.

For the selected points, a matrix, $\mathbf{B}$, is created, similar to matrix $\mathbf{A}$ but with only the k selected lines in the matrix. For this matrix, $\mathbf{B}$, the objective function is calculated for each individual. An individual is defined as a set of points selected by the genetic algorithm. In one iteration of the genetic algorithm, the genetic algorithm creates a number of individuals, and then crossover takes place between these individuals in order to create individuals with (hopefully) better fitness and to obtain an optimal selection of points from the sample space, after several iterations.

The objective function that measures the fitness of an individual is defined by $\det(\mathbf{B}^T \mathbf{B})$, as specified by the D-optimal sampling method. For each individual the objective function is computed and the individuals are ranked according to their fitness values. A higher value for the objective function indicates a higher fitness value.

5.2.4 Selection

Selection is the step in the genetic algorithm that selects the individuals, from the current generation, that will be parents for reproduction.

At first roulette wheel selection is used to select the individuals (or parents) that should undergo reproduction to reproduce offspring. In the roulette wheel selection, the individuals are ranked according to their fitness, and the individual with a bigger fitness is given a greater probability to reproduce. This is done by dividing the fitness of an individual, f , by the total fitness of all the individuals $\sum f$ such that, $p = \frac{f}{\sum f}$. In this manner, the fitter individuals are more likely to be chosen as parents for reproduction, and the fitter genes are more likely to be carried over to the next generation of individuals. The problem encountered with roulette wheel selection, is that the matrices $\mathbf{A}$ and $\mathbf{B}$ that are considered yield determinants that are too big to ensure efficient roulette wheel selection. The value of the objective function of the individual with the highest fitness dominates all the other fitness values to such an extent that the fittest individual is given a 0.99 probability to be selected. This decreases the diversity of the genetic algorithm, and renders the genetic algorithm ineffective. Another selection technique has to be considered.

One way is to use the objective function $(\log(\det(\mathbf{A}^T \mathbf{A})))$. This way there can still be distinguished between the fitness values of the individuals, and all the individuals can be included in the selection process. This was however not successful. Another option is to use tournament selection.

Tournament selection is a selection process where a number of individuals “compete” against one another by means of their fitness values, and in every tournament, the fittest individual is selected for reproduction. This selection technique is known to be more effective than the roulette wheel selection. For this problem, the tournament selection is implemented as follows: Two parents are randomly selected from the population and the two parents’ fitness are then compared. The fitter individual is selected for reproduction. This is repeated until sufficient parents are selected for reproduction.

5.2.5 Crossover

Crossover enables two individuals to reproduce in order to create a new offspring. In this case, one point crossover is used. Two parents that are selected by the tournament selection undergoes crossover. A random point in the k number of rows (or data points) chosen is selected. The two parents crossover by exchanging the part of the individual behind this point. This creates new individuals or offspring that have the possibility of higher fitness than the fitness of the parents.

5.2.6 Mutation

Mutation adds to the diversity of the genetic algorithm. In this implementation, a mutation probability of 0.001 is used. If the mutation probability is too high, the method will resemble pure random sampling. If on the other hand the mutation probability is too low, then the diversity of the genetic algorithm is restricted. For mutation, a random number is generated. If this number is smaller than the mutation probability, then the child that is created undergoes mutation. If the number generated is bigger than the mutation probability, then no mutation takes place. For mutation to take place, a random point in the child or individual is selected, and this point or row is changed to a random other row in the full matrix. This ensures that the sample space is searched thoroughly.

5.2.7 Elitism

Elitism is used in the genetic algorithm to make sure that the fitness of the best individual in the next generation does not become less than the highest fitness in the current generation. It ensures that in each case the fittest parent of the previous generation is carried over, as it is, to the next generation. This way good genes are not lost in the process of crossover and mutation. The fittest individual in each case is selected, and a randomly chosen line of the newly created “offspring” matrix is replaced by this fittest parent.

5.2.8 Implementation

The steps in the genetic algorithm are iterated until some termination criterion is reached. In this implementation, the genetic algorithm is terminated after a maximum number of iterations, according to the requirements of the problem. The higher the number of allowed iterations of the algorithm, the more optimal the distribution of points will be.

An example of the implementation of the genetic algorithm to perform D-optimal sampling on the data set is the selection of the optimal 100 points from the total 6103 points, for the cubic equation in Equation (5.23) with the variables temperature (T), mixed layer depth (MLD), chlorophyll-a concentration (Chl) and latitude (Lat). The genetic algorithm implementation has the following specifications:

Number of generations:	100000	
Number of lines selected:	100	
Mutation probability:	0.001	(5.8)
Number of individuals selected per generation:	20	

The history of the fitness of the best individual is shown in Figure 5.1. The histogram in Figure 5.2 shows the distribution of the points selected by the genetic algorithm. As is expected for a cubic curve fit, the points are mainly selected from three areas. The points selected are grouped together as is shown by the histogram bars — there are two gaps in the bars separating the groups. Figures 5.3 to 5.8 give an indication of where the genetic algorithm selected the points in the search space. Note here that the points selected are equally distributed for all the possible pairs of variables, and that the extreme points are also selected, thus an accurate representation of the whole data set is given in this manner by the selected points.

The genetic algorithm as described in this section is used in this dissertation to select the optimal points for different equations and investigations. For each different equation that the genetic algorithm is used for, the genetic algorithm yields other results in terms of the best distribution of points in the sample space for the specific equation and number of points that are to be selected.

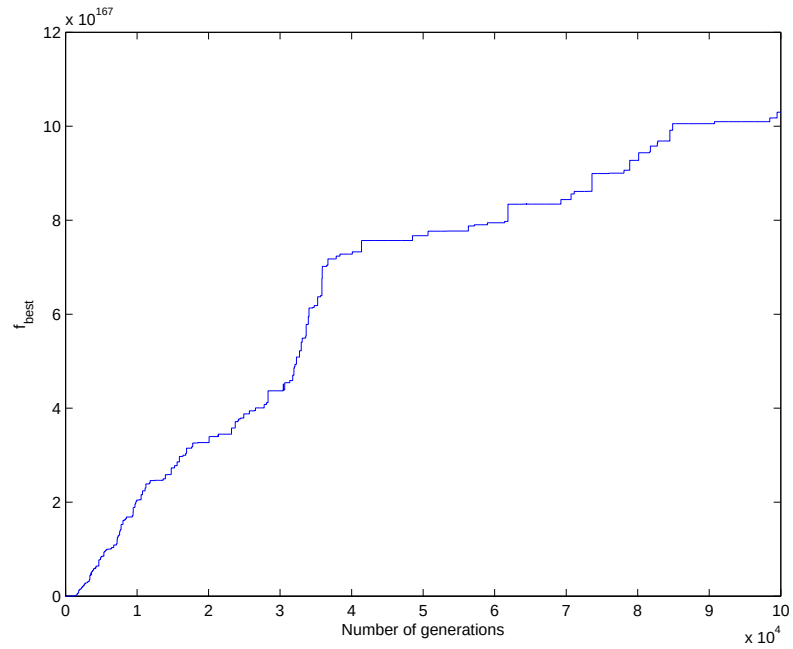


Figure 5.1: An illustration of how the best fitness of the individual increases with increasing iterations

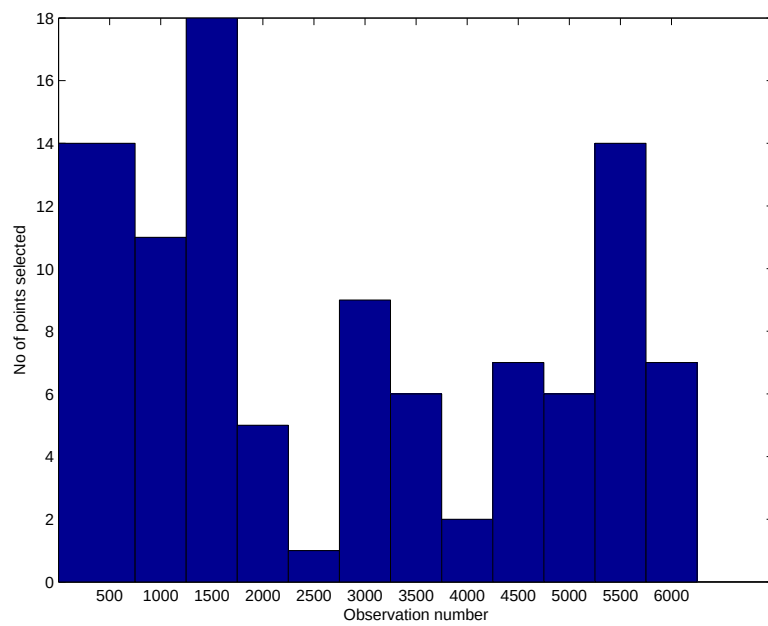


Figure 5.2: The histogram of the points selected by the individual with the best fitness in each iteration

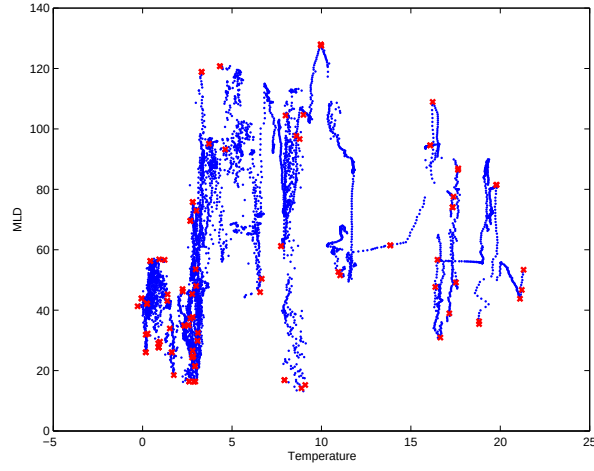


Figure 5.3: The distribution of the selected points in the temperature–MLD plane. The red crosses indicate the points selected by the genetic algorithm

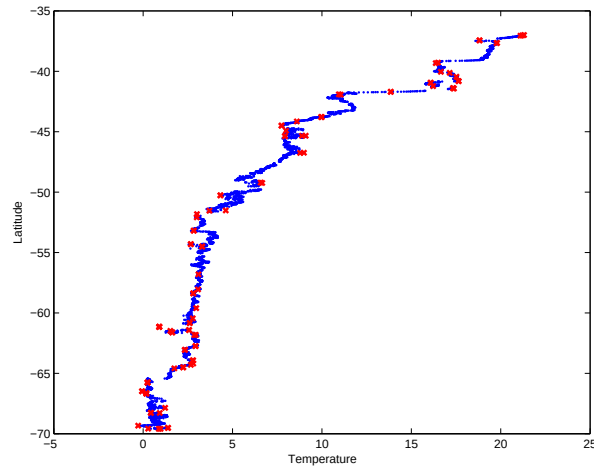


Figure 5.4: The distribution of the selected points in the temperature–latitude plane. The red crosses indicate the points selected by the genetic algorithm

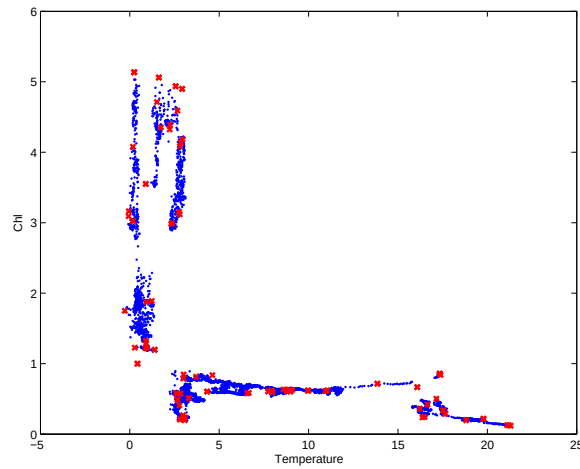


Figure 5.5: The distribution of the selected points in the temperature–chlorophyll plane. The red crosses indicate the points selected by the genetic algorithm

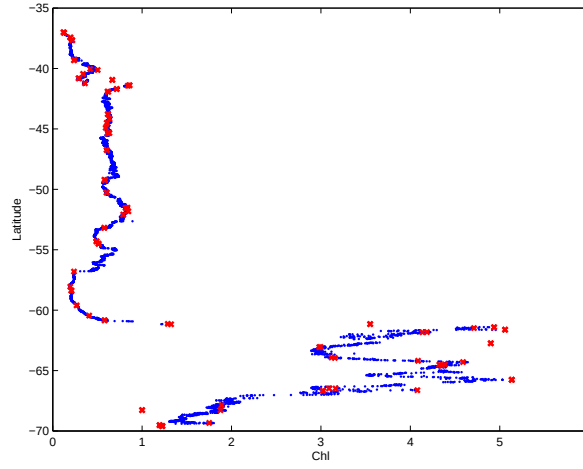


Figure 5.6: The distribution of the selected points in the latitude–chlorophyll plane. The red crosses indicate the points selected by the genetic algorithm

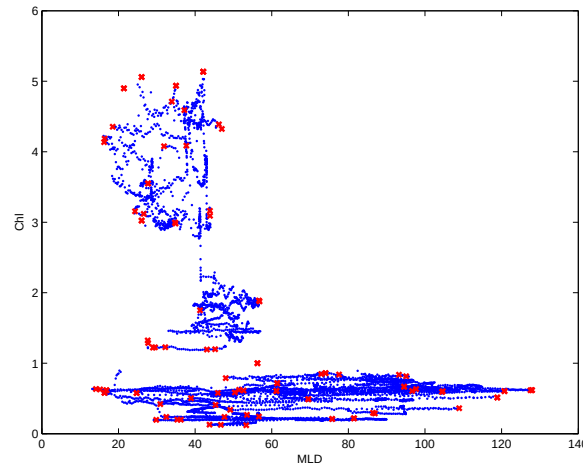


Figure 5.7: The distribution of the selected points in the MLD–chlorophyll plane. The red crosses indicate the points selected by the genetic algorithm

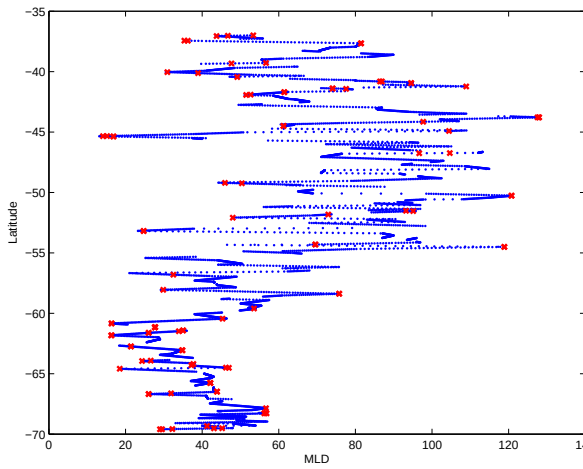


Figure 5.8: The distribution of the selected points in the MLD–latitude plane. The red crosses indicate the points selected by the genetic algorithm

5.3 Least Squares curve fitting

Least squares curve fitting was originally used to find the best possible solution for a system of overdetermined linear equations [64]. In this curve fitting, a number of data points and a class of function (i.e. exponential, polynomial etc.) is specified and the aim is to find the coefficients that fit the points such that the error between the actual data points and the values obtained by the equation is as small as possible [64].

The variables that can be measured by both satellites and *in situ* measurements include, amongst others, the temperature of the ocean at the sea surface, the mixed layer depth (MLD) and the chlorophyll-a concentration as well as the latitude and longitude values at which these measurements were made. Oceanographers in the field suggested that the temperature, mixed layer depth and chlorophyll-a concentration will be the variables that adhere to two important criteria: It can be measured by both *in situ* measurements and satellites; and these variables are expected to have an effect on the amount of pCO₂ in the ocean waters [36] [56]. In this implementation, the relationship between pCO₂ and the temperature (T), mixed layer depth (MLD) and chlorophyll-a concentration (Chl) is investigated using least squares optimization. The aim is to establish a relationship as follows:

$$\text{pCO}_2 = f(\text{T}, \text{MLD}, \text{Chl}). \quad (5.9)$$

There are various ways to determine such a relationship. This dissertation was started off by considering least squares optimization to find the equation, containing the mentioned variables, that fits the data with the smallest error. In every implementation of the least squares curve fit, the same steps were followed. Firstly, a number of points (k) is selected using both random and D-optimal sampling respectively. The least squares optimization is carried out on the k number of points selected, and the equation for pCO₂ is obtained from the least squares fit. Then this equation is applied to all 6103 points, and the error of this prediction of pCO₂ values on all the data points is the value that is to be minimized. Every time a different set of points is used to construct the equation, a different equation will be obtained for the estimation of the pCO₂. In this dissertation, however, the equation is obtained for $100 \times k$ different points, and the average coefficients are determined over the 100 different results, in order to obtain an improved result across the data points that is considered.

The least squares optimization is done as follows:

The system, $\mathbf{y} = \mathbf{B}\boldsymbol{\beta}$, is considered, where $\mathbf{y}$ is the $k \times 1$ vector of pCO₂ values, $\mathbf{B}$ is the $k \times n$ matrix

with k the number of points sampled from the data set and n the number of terms in the equation fitted to the data, and $\boldsymbol{\beta}$ is the $n \times 1$ vector of coefficients for the k equations in matrix $\mathbf{B}$. The aim is to minimize the sum of the square of the error. In this case the error is given by

$$\mathbf{E} = (\mathbf{y} - \tilde{\mathbf{y}}), \quad (5.10)$$

where $\mathbf{y}$ is a vector of the actual measurement of pCO_2 in the ocean, and $\tilde{\mathbf{y}}$ is a vector of modelled pCO_2 values. The objective function that is to be minimized for the least squares optimization then becomes:

$$f(\boldsymbol{\beta}) = (\mathbf{y} - \tilde{\mathbf{y}})^T (\mathbf{y} - \tilde{\mathbf{y}}) \quad (5.11)$$

$$= (\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{B}\boldsymbol{\beta}). \quad (5.12)$$

To minimize this function, the coefficients $\boldsymbol{\beta}$ should be determined such that $\frac{\delta f(\boldsymbol{\beta})}{\delta \boldsymbol{\beta}} = 0$. Taking $f(\boldsymbol{\beta})$ as in Equation (5.11), differentiating this and setting it equal to zero yields the solution:

$$\boldsymbol{\beta} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{y}. \quad (5.13)$$

The least squares optimization in this implementation is done by calculating $\boldsymbol{\beta}$ as described in Equation (5.13), where $\mathbf{B}$ in this case is the matrix containing only the k lines of data that represents the k data points selected from the complete data set.

5.3.1 Excluding latitude as a variable

The least squares optimization is applied to different equations and different numbers of k , in order to compare various equations of pCO_2 with one another. The variables used in this implementation include temperature (T), mixed layer depth (MLD) and chlorophyll-a concentration (Chl). For each curve fit, the curve fit is repeated 100 times, with a different selection of points each time, in order to accurately determine the variability of the error made in the prediction of pCO_2 . For every curve fit the root mean square (RMS) error is calculated, as follows:

$$RMS \text{ error} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \tilde{y}_i)^2}, \quad (5.14)$$

taking into account the error on all 6103 data points. The RMS error for a curve fit is calculated by taking the mean RMS error for the 100 different answers for the curve fit. Furthermore, the standard deviation, σ , of the error made in the prediction of the pCO_2 is also calculated for each curve fit by

$$\sigma = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - y_{avg})^2}. \quad (5.15)$$

The maximum under estimation and maximum over estimation in each case is also recorded, to give an indication of the variance of the data. The mean values for the coefficients for the curve fit for each implementation are also determined.

Linear curve fit

The least squares optimization for the linear equation optimally fits

$$\text{pCO}_2 = \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl}, \quad (5.16)$$

to the data selected by random and D-optimal sampling, respectively. The results of the different curve fits for the linear equations are shown in Figures A.1 to A.24 in Appendix A on page 181 to 188. The mean estimated pCO_2 value for all 100 curve fits in each case is plotted along with the 95% confidence interval in each case. The results for the linear least squares curve fit are summarized in Table 5.1. The mean values of coefficients β_1 to β_5 for the equation of the pCO_2 in each case are given in Table A.1. The standard deviations for coefficients β_1 to β_5 of all 100 curve fits are given in Table A.2. Table 5.2 shows the ratio of the standard deviation for the coefficients between the 200 randomly sampled points and the 200 D-optimally sampled points for the 100 curve fits. The ratio

$$\frac{\sigma_{\text{random}}}{\sigma_{\text{D-optimal}}}, \quad (5.17)$$

determines whether the standard deviation for the random sampling is bigger than the standard deviation of the D-optimal sampling.

It can be seen from the results that for a small number of points, the random sampling yields a RMS error of $27.92 \mu\text{atm}$ compared to a RMS error of $23.13 \mu\text{atm}$ with the D-optimal sampling. The D-optimal sampling performs better than the random sampling for a small number of points used and a smaller 95% confidence interval is obtained with the D-optimal sampling than with the random

Table 5.1: Results: Linear curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Es- timation	Max Unde Es- timation	σ	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ
10	27.92	390.7	-303.0	30.88	23.13	117.7	-146.5	23.23
50	19.90	97.98	-126.9	19.92	21.00	83.75	-136.1	21.04
100	19.37	93.17	-120.9	19.38	20.64	71.76	-117.7	20.67
200	19.14	69.89	-118.2	19.15	20.27	64.41	-113.7	20.27

Table 5.2: Ratio of standard deviation of coefficients of linear curve fitting for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	0.7780
β_2	0.7407
β_3	1.217
β_4	1.279

sampling (as can be seen in the figures). It can also be seen that the linear curve fit does not fit all the spikes in the data and could still be improved. In Table 5.1 it can be seen that for 200 sample points from the data set the random sampling yields a RMS error of $19.14 \mu\text{atm}$ compared to the RMS error of $20.27 \mu\text{atm}$ obtained with the D-optimal sampling. In Table 5.2 it is shown that although the RMS error is smaller for the 200 points sampled, the standard deviations of the coefficients determined from the curve fit is smaller with the random sampling for the first two coefficients and smaller in the D-optimal sampling for the last two coefficients. The linear curve fit yields results that are irregular for both random and D-optimal sampling and the inconsistent results makes it difficult to make conclusions from this curve fit.

Quadratic curve fit

The quadratic equation with the variables T, MLD and Chl is fitted to the data set in order to investigate the degree to which it improves on the linear fit. The quadratic equation that is fit to the data is given by:

$$\begin{aligned} \text{pCO}_2 = & \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 T^2 + \beta_6 \text{MLD}^2 + \beta_7 \text{Chl}^2 \\ & + \beta_8 T \cdot \text{MLD} + \beta_9 T \cdot \text{Chl} + \beta_{10} \text{MLD} \cdot \text{Chl} , \end{aligned} \quad (5.18)$$

and again the equation is fitted using points sampled both randomly and D-optimally.

The results of the curve fits obtained from the quadratic fit of the data are shown in Figures B.1 to B.24 in the Appendix B on pages 191 to 198. The mean estimated pCO₂ values for all 100 curve fits are plotted, along with the 95% confidence interval in each case. The results for the least squares optimization of the quadratic curve fits to the data set are summarized in Table 5.3. The mean values of the coefficients for all the terms in Equation (5.18) (β_1 to β_{15}) for all 100 curve fits in each case, for each k number of points chosen, are presented in Table B.1. The standard deviation for the coefficients on all 100 curve fits are given in Table B.2. The ratios of the standard deviation for the coefficients between the randomly sampled 200 points and the D-optimally sampled 200 points for the 100 curve fits are given in Table 5.4.

The RMS error on 25 random points sampled is 28.70 μatm in this case, compared to the RMS error of 19.59 μatm obtained with the 25 points sampled D-optimally. It can once again be seen that the D-optimal sampling yields better results than the random sampling for a small number of points sampled. Although the RMS error of 16.75 μatm for 200 random sampled points is smaller than the RMS error of 18.13 μatm for the 200 D-optimally sampled points, it can be seen in Table 5.4 that the standard deviations of the coefficients determined for the quadratic equation in 100 curve fits are larger with the random sampling than with the D-optimal sampling. This indicates that the D-optimal sampling is a more consistent and robust method to use than the random sampling. Comparing Table 5.1 and Table 5.3, it can be seen that the quadratic curve fit improves the error of the estimation from the linear curve fit. For 200 D-optimal sampled points a RMS error of 20.27 μatm is obtained with the linear curve fit compared to a RMS error of 18.13 μatm with the quadratic equation. The quadratic curve fit does not fit all the spikes in the data and the fit could still be improved to yield more accurate estimations of the pCO₂.

Table 5.3: Results: Quadratic curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Estimation	Max Under Estimation	σ	Mean RMSE	Max Over Estimation	Max Under Estimation	σ
25	28.70	271.3	-991.0	35.82	19.59	85.30	-136.7	19.58
50	19.02	162.8	-205.5	19.15	18.79	85.54	-113.4	18.81
100	17.38	126.6	-114.0	17.38	18.50	76.06	-101.0	18.50
200	16.75	89.13	-106.1	16.76	18.13	70.66	-95.51	18.11

Table 5.4: Ratio of standard deviation of coefficients of quadratic curve fitting for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	1.436
β_2	1.682
β_3	1.924
β_4	1.550
β_5	1.945
β_6	2.069
β_7	1.379
β_8	1.626
β_9	4.087
β_{10}	2.465

Cubic curve fit

A cubic equation with variables T, MLD and Chl is fitted to the data in order to find an equation that can estimate the pCO₂ values in the ocean with known values for T, MLD and Chl. The cubic equation that is fitted to the data is given by:

$$\begin{aligned}
 \text{pCO}_2 = & \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 T^2 + \beta_6 \text{MLD}^2 + \beta_7 \text{Chl}^2 + \beta_8 T \cdot \text{MLD} \\
 & + \beta_9 T \cdot \text{Chl} + \beta_{10} \text{MLD} \cdot \text{Chl} + \beta_{11} T^3 + \beta_{12} \text{MLD}^3 + \beta_{13} \text{Chl}^3 + \beta_{14} T^2 \cdot \text{MLD} \\
 & + \beta_{15} T^2 \cdot \text{Chl} + \beta_{16} T \cdot \text{MLD} \cdot \text{Chl} + \beta_{17} \text{MLD}^2 \cdot T + \beta_{18} \text{MLD}^2 \cdot \text{Chl} \\
 & + \beta_{19} \text{Chl}^2 \cdot T + \beta_{20} \text{Chl}^2 \cdot \text{MLD}.
 \end{aligned} \tag{5.19}$$

Equation (5.19) is fitted to the data by selecting a subset of the data both randomly or D-optimally and finding coefficients β_1 to β_{20} by means of least squares optimization. The results obtained are shown in Figures C.1 to C.24 in Appendix C on pages 202 to 209. The mean pCO₂ estimated values for all 100 curve fits in each case are plotted together with the 95% confidence interval in each case. The results for the cubic curve fits are summarized in Table 5.14. The mean values for coefficients β_1 to β_{35} obtained in each case are given in Table C.1. The standard deviations for the coefficients on the 100 curve fits are given in Table C.2. The ratios of the standard deviations of the 200 randomly sampled points compared to the standard deviations of the 200 D-optimally sampled points are given in Table 5.6.

Similar to the linear and quadratic curve fits, the RMS error of 33.23 μatm for 50 randomly sampled points is bigger than the RMS error of 18.40 μatm for the 50 D-optimally sampled points. It can be seen from the results that for a small number of points, the D-optimal sampling reduces the mean error on the pCO₂ estimation as well as the variance on the pCO₂ estimations. It can also be seen that the maximum over- and under estimation is smaller when using the D-optimal sampling even with 200 points. The RMS error of 15.90 μatm for 200 randomly sampled points is smaller than the RMS error of 17.94 μatm for the 200 D-optimal sampled points. However, considering Table 5.6 it can be seen that the standard deviations on the coefficients for the random sampling of 200 points is considerably larger than the standard deviations on the coefficients for the D-optimal sampling. This shows that the random sampling yields inconsistent results and that the D-optimal sampling is a more robust sampling method to use. The RMS error of 18.13 μatm for the 200 D-optimal sampled points for the quadratic curve fit is bigger than the RMS error of 17.94 μatm for the cubic curve fit with 200 D-optimal sampled points. This shows that the cubic curve fit further reduces the error

Table 5.5: Results: Cubic curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Estimation	Max Under Estimation	σ	Mean RMSE	Max Over Estimation	Max Under Estimation	σ
50	33.23	1543.9	-965.5	41.07	18.40	99.06	-70.22	18.2
55	26.90	782.2	-817.2	29.50	17.97	101.8	-69.50	17.80
100	18.40	198.3	-229.3	18.50	17.99	95.50	-62.50	17.70
200	15.90	115.2	-186.0	16.00	17.94	92.80	-60.70	17.70

of the estimation. The figures also show that the cubic equation fits more of the spikes in the data although some of the big spikes in the data are still not fit by the cubic equation.

Table 5.6: Ratio of standard deviation of coefficients of cubic curve fitting for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	6.313
β_2	4.334
β_3	5.639
β_4	4.769
β_5	3.463
β_6	5.040
β_7	3.200
β_8	5.784
β_9	6.230
β_{10}	7.126
β_{11}	3.219
β_{12}	4.372
β_{13}	2.461
β_{14}	5.585
β_{15}	4.359
β_{16}	7.822
β_{17}	5.376
β_{18}	7.953
β_{19}	5.685
β_{20}	6.320

Fourth order curve fit

The fourth order curve fit is performed in order to quantify the reduction in error by adding an additional 15 terms to the 20 terms in the cubic equation. The fourth order equation is given by:

$$\begin{aligned}
\text{pCO}_2 = & \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 T^2 + \beta_6 \text{MLD}^2 + \beta_7 \text{Chl}^2 + \beta_8 T \cdot \text{MLD} + \beta_9 T \cdot \text{Chl} \\
& + \beta_{10} \text{MLD} \cdot \text{Chl} + \beta_{11} T^3 + \beta_{12} \text{MLD}^3 + \beta_{13} \text{Chl}^3 + \beta_{14} T^2 \cdot \text{MLD} + \beta_{15} T^2 \cdot \text{Chl} \\
& + \beta_{16} T \cdot \text{MLD} \cdot \text{Chl} + \beta_{17} \text{MLD}^2 \cdot T + \beta_{18} \text{MLD}^2 \cdot \text{Chl} + \beta_{19} \text{Chl}^2 \cdot T + \beta_{20} \text{Chl}^2 \cdot \text{MLD} \\
& + \beta_{21} \text{Chl}^4 + \beta_{22} \text{Chl}^3 \cdot \text{MLD} + \beta_{23} \text{Chl}^3 \cdot T + \beta_{24} \text{Chl}^2 \cdot \text{MLD}^2 + \beta_{25} \text{Chl}^2 \cdot \text{MLD} \cdot T \\
& + \beta_{26} \text{Chl}^2 \cdot T^2 + \beta_{27} \text{Chl} \cdot \text{MLD}^3 + \beta_{28} \text{Chl} \cdot \text{MLD}^2 \cdot T + \beta_{29} \text{Chl} \cdot \text{MLD} \cdot T^2 + \beta_{30} \text{Chl} \cdot T^3 \\
& + \beta_{31} \text{MLD}^4 + \beta_{32} \text{MLD}^3 \cdot T + \beta_{33} \text{MLD}^2 \cdot T^2 + \beta_{34} \text{MLD} \cdot T^3 + \beta_{35} T^4.
\end{aligned}
\tag{5.20}$$

The results obtained for the fourth order equation that is fit to the data are shown in Figures D.1 to D.18 on pages 215 to 220 in Appendix D and the results for the 400 points sampled are shown in Figures 5.9 to 5.14. The first figure in each case shows the mean pCO₂ estimation values for all 100 curve fits which are plotted together with the 95% confidence interval. The second figure in each case shows the actual data vs. the modelled data with lines indicating the 10% error on the estimation, in order to see how many points fall outside this 10% error line. The last figure in each case shows the histogram of the prediction errors for the 100 curve fits in order to give an indication of the distribution of the errors. A summary of the results for the fourth order polynomial least squares optimization is given in Table 5.7. The mean for coefficients β_1 to β_{35} of all 100 fourth order curve fits are given in Table D.1. The standard deviations of the coefficients for the 100 curve fits of the fourth order equation is given in Table D.2. The ratio of the standard deviations on the coefficients of the 200 randomly sampled points compared to the standard deviations of the 200 D-optimal sampled points are given in Table 5.8.

The RMS error of 30.43 μatm for the 100 randomly sampled points is considerably larger than the RMS error of 15.41 μatm , confirming that for a small number of points the D-optimal sampling yields smaller RMS errors on the estimation than the random sampling. It can also be seen that the maximum over- and under estimation is also smaller with the D-optimal sampling. The standard deviations on the 100 different curves are much bigger with the randomly sampled points than with the D-optimal sampled points that can also be seen on the figures with the 95% confidence interval.

Table 5.7: Results: Fourth Order curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Estimation	Max Under Estimation	σ	Mean RMSE	Max Over Estimation	Max Under Estimation	σ
100	30.43	740.0	-2507	38.01	15.41	95.54	-69.79	11.11
120	24.55	903.6	-1161	27.90	15.58	99.11	-65.03	11.26
200	14.87	278.4	-268.2	15.04	15.47	93.72	-64.30	11.21
400	13.01	175.1	-126.8	13.02	15.36	92.67	-61.25	11.14

The RMS error of $13.01 \mu\text{atm}$ for the 200 randomly sampled points is smaller than the RMS error of $15.36 \mu\text{atm}$ for the 200 D-optimal sampled points for the fourth order fit. However, considering Table 5.8, it can be seen that the standard deviations of the coefficients for the random sampling is considerably larger than the standard deviations of the D-optimal sampling proving that the D-optimal sampling is more consistent and robust than the random sampling. The RMS error of $15.36 \mu\text{atm}$ of the 200 D-optimally sampled points for the fourth order curve fit is smaller than the RMS error of $17.94 \mu\text{atm}$ for the 200 D-optimally sampled points for the cubic curve fit. The fourth order curve fit thus yields an even bigger improvement on the error of the estimation. It can also be seen from the figures that the fourth order equation fits more spikes in the data than the cubic curve fit.

Table 5.8: Ratio of standard deviation of coefficients of fourth order curve fitting for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	4.500
β_2	3.528
β_3	2.989
β_4	4.697
β_5	2.737
β_6	2.547
β_7	4.380
β_8	5.860
β_9	4.158
β_{10}	3.974
β_{11}	2.700
β_{12}	2.540
β_{13}	4.107
β_{14}	4.546
β_{15}	4.190
β_{16}	5.045
β_{17}	4.261
β_{18}	3.820
β_{19}	3.978
β_{20}	4.312
β_{21}	3.843
β_{22}	4.915
β_{23}	4.287
β_{24}	4.299
β_{25}	4.338
β_{26}	4.646
β_{27}	3.820
β_{28}	5.912
β_{29}	4.418
Continued**	

Table 5.8 – **continued

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_{30}	4.721
β_{31}	2.477
β_{32}	2.780
β_{33}	6.708
β_{34}	3.322
β_{35}	2.862

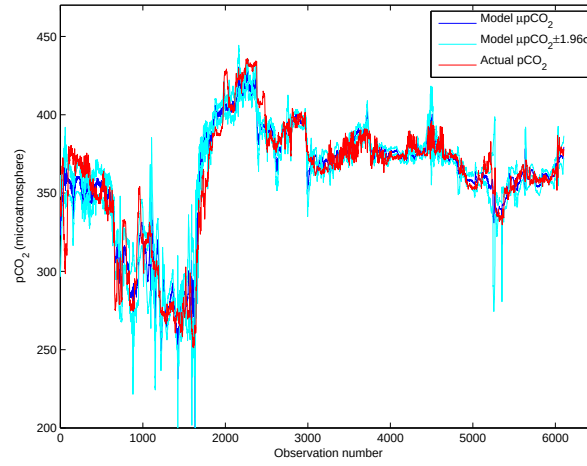


Figure 5.9: Fourth order curve fit results. 400 Random points sampled.

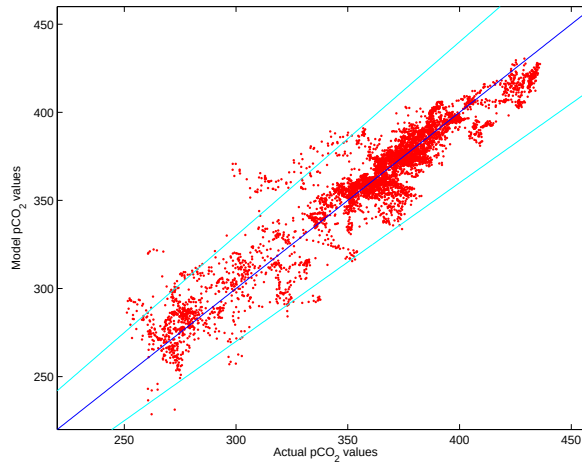


Figure 5.10: Fourth order curve fit results with 400 random points sampled. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

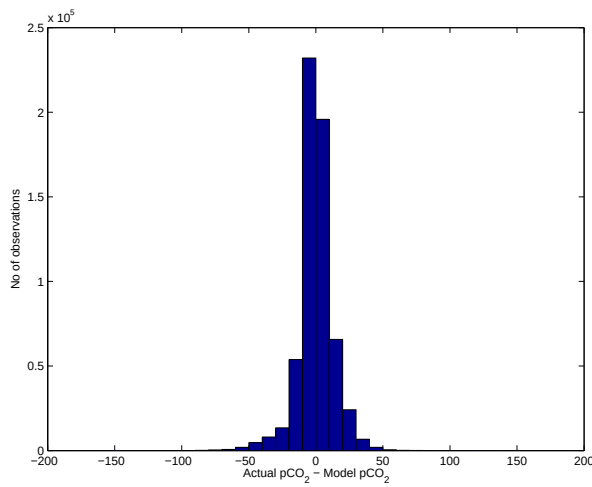


Figure 5.11: Fourth order curve fit results. Histogram of errors for all 100 runs and 400 random points sampled.

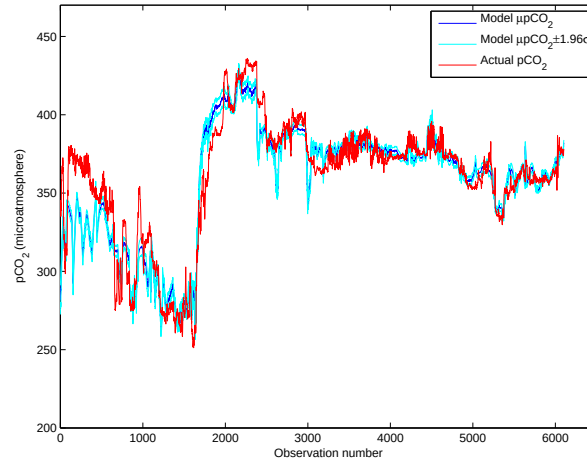


Figure 5.12: Fourth order curve fit results. 400 D-optimal sampled points.

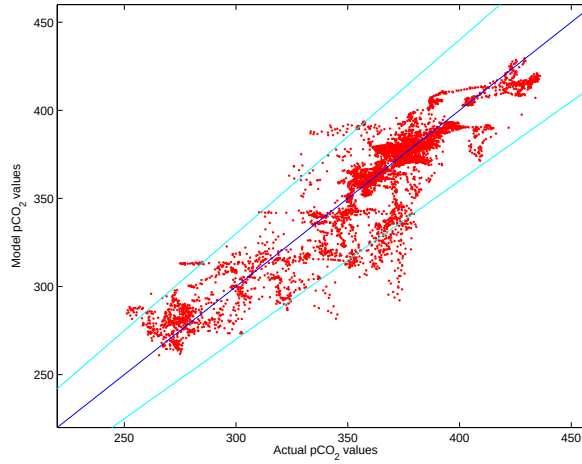


Figure 5.13: Fourth order curve fit results with 400 D-optimal sampled points. The degree to which the modelled points fall in the 10 % error margin is shown. The mean pCO_2 estimation for all 100 runs is shown here.

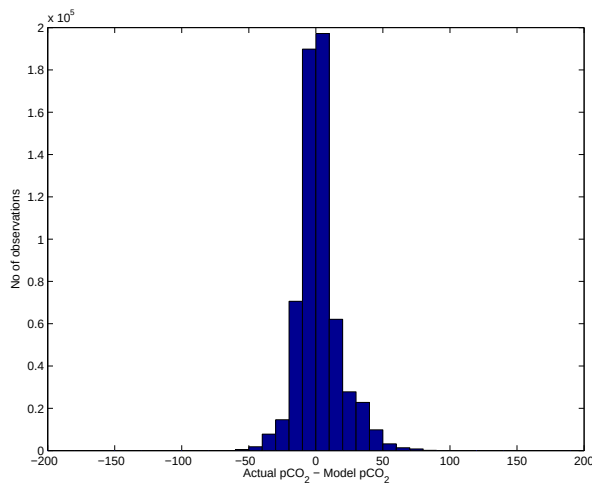


Figure 5.14: Fourth order curve fit results. Histogram of errors for all 100 runs and 400 D-optimal sampled points.

Removing terms

The fourth order equation yields a fit to the data with a reasonably small error. The terms in Equation (5.20) are systematically removed one by one in order to determine which of the terms in this equation plays the biggest role in the estimation of the $p\text{CO}_2$. The aim is to determine which of the terms can be removed without increasing the error on the estimation of $p\text{CO}_2$ to a great extent.

The error of the estimation of $p\text{CO}_2$ made by the fourth order equation is determined by fitting all 6103 data points with the fourth order equation, and calculating the RMSE of this fit. The terms are removed one-by-one in order to determine the error made in the $p\text{CO}_2$ prediction when each one of the terms are removed individually. The term that, when it is removed, makes the least difference on the error of the prediction, is removed from the equation. Once a term is removed, all the remaining terms are once again removed one-by one, in order to determine which term can be removed that will make the smallest difference on the error of the prediction. This is repeated until only one term is left. The results are given in Table 5.9. The error on the estimation of the $p\text{CO}_2$ as the terms are removed from the fourth order equation is shown in Figure 5.15.

From this table, the optimal n terms to estimate the $p\text{CO}_2$ can be determined. This can be used in applications where only a small number of terms can be used. Looking at this table, the optimal 20 terms yields an error of $13.65 \mu\text{atm}$ which is smaller than the error of $17.94 \mu\text{atm}$ obtained for the cubic equation in Table 5.5 on page 111. Thus with the same amount of terms, a better estimation of $p\text{CO}_2$ can be found in this manner.

Table 5.9: Table of RMSE when systematically removing terms for the fourth order equation.

No	Coefficient	Term removed	RMSE
1	β_{20}	$\text{Chl}^2 \cdot \text{MLD}$	11.967
2	β_{18}	$\text{MLD}^2 \cdot \text{Chl}$	11.973
3	β_{32}	$\text{MLD}^3 \cdot \text{T}$	11.997
4	β_{35}	T^4	12.031
5	β_{31}	MLD^4	12.081
6	β_{28}	$\text{Chl} \cdot \text{MLD} \cdot \text{T}^2$	12.154
7	β_1	1	12.253
8	β_{34}	$\text{MLD} \cdot \text{T}^3$	12.324
9	β_{10}	$\text{MLD} \cdot \text{Chl}$	12.440
Continued **			

Table 5.9 – ** continued

No	Coefficient	Term removed	RMSE
10	β_{22}	Chl ³ ·MLD	12.679
11	β_{24}	Chl ² ·MLD ²	12.773
12	β_{12}	MLD ³	12.875
13	β_{27}	Chl·MLD ³	13.057
14	β_{29}	Chl·MLD·T ²	13.224
15	β_{25}	Chl ² ·MLD·T	13.485
16	β_{16}	T·MLD·Chl	13.653
17	β_{33}	MLD ² ·T ²	14.450
18	β_{14}	T ² ·MLD	14.543
19	β_{17}	MLD ² ·T	15.266
20	β_8	T·MLD	15.511
21	β_3	MLD	15.993
22	β_6	MLD ²	16.023
23	β_{21}	Chl ⁴	17.725
24	β_{30}	Chl·T ³	18.9.5
25	β_{26}	Chl ² ·T ²	21.424
26	β_{23}	Chl ³ ·T	24.699
27	β_{13}	Chl ³	27.465
28	β_{19}	Chl ² ·T	48.485
29	β_{15}	T ² ·Chl	46.468
30	β_7	Chl ²	73.008
31	β_{11}	T ³	91.436
32	β_9	T·Chl	125.754
33	β_4	Chl	158.096
34	β_5	T ²	240.447
35	β_2	T	No model

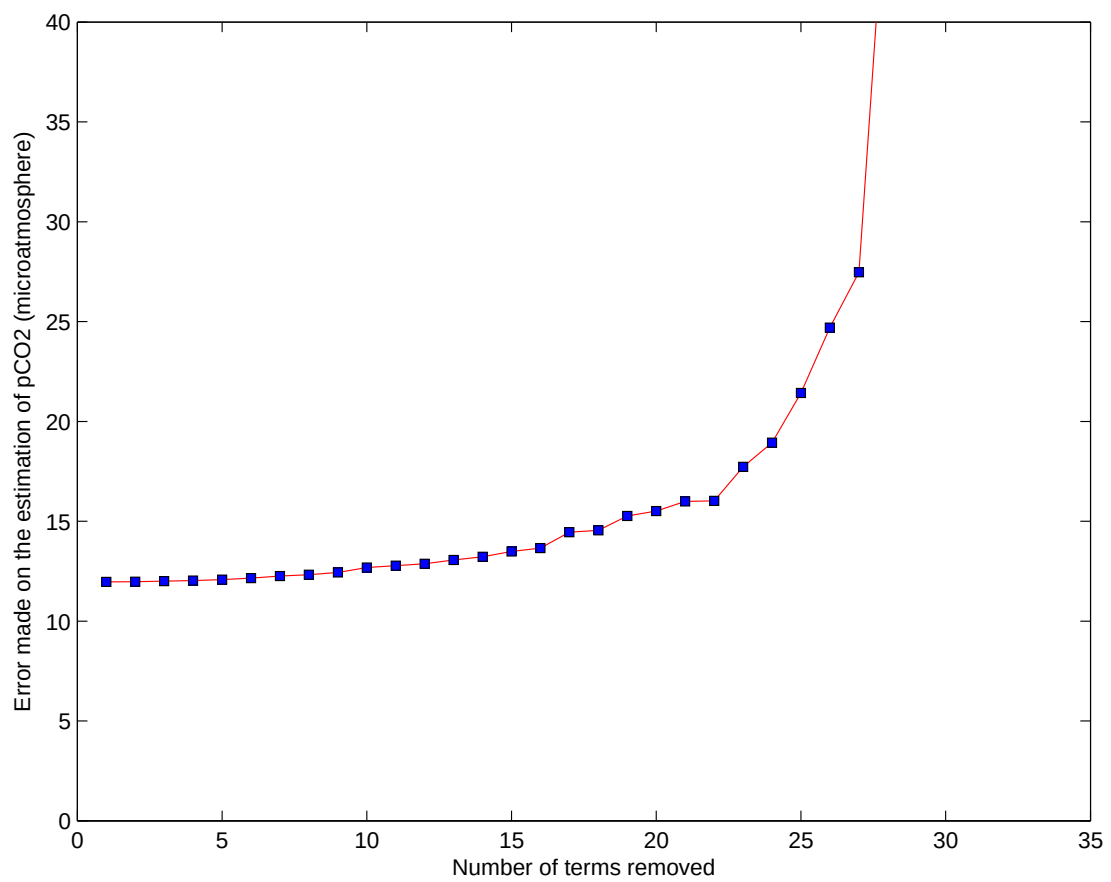


Figure 5.15: The increase of the error as the terms are removed from the fourth order polynomial.

5.3.2 Including latitude as variable

Different regions in the ocean are created by the flow of currents in different directions in the ocean. In the part of the ocean that is considered in this dissertation, some oceanic regions also exist. In the African sector of the Southern Ocean, essentially four regions exist [63]. The first region is the Sub-Antarctic zone that lies between the Subtropical front (40°S) and the Sub-Antarctic Front (45°S) [63]. The second region is the Polar Frontal zone, which is between the Sub-Antarctic Front (45°S) and the Antarctic Polar Front (50°S) [63]. The third region is the South ACC Frontal Zone, stretching from the Antarctic Polar Front (50°S) to the Southern ACC front (53°S) [63]. The fourth region in this stretch of the ocean is the Southern Boundary zone, which lies between the Southern ACC front (53°S) and the Southern boundary of the ACC (55°S) [63].

Oceanographers in the field suggest that the relationship between the oceanic variables differ in the different regions. One of the ways to approach this is to consider the regions separately, and create a unique equation for the estimation of pCO₂ for each region [36], [56]. Another approach is to include latitude as variable in the pCO₂ estimation equation. In this way the variability of the relationship between the variables is taken into account by the equation. The regions are then implicitly included in the equation that is created for the ocean. In the following sections, latitude (Lat) is included as a variable. This way, knowledge of where all the regions of the ocean are is not required, but the variabilities between the regions are implicitly taken into account.

Linear curve fit with latitude included as variable

The least squares optimization for the linear equation optimally fits

$$\text{pCO}_2 = \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 \text{Lat} \quad (5.21)$$

to the data points that are selected by random and D-optimal sampling respectively. The results of the different curve fits for the linear equations are shown in Figures E.1 to E.24 in Appendix E on pages 223 to 230. The mean estimated pCO₂ values for all 100 curve fits in each case are plotted together with the 95% confidence interval for the 100 runs. The results for the linear least squares curve fit are summarized in Table 5.10. The mean values of the coefficients β_1 to β_5 are given in Table E.1. The standard deviations for the coefficients on the 100 curve fits are given in Table E.2. The ratio of the standard deviations of the coefficients for the 200 randomly sampled points compared to the standard deviations of the coefficients for the 200 D-optimally sampled points are given in Table

Table 5.10: Results: Linear curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Estimation	Max Under Estimation	σ	Mean RMSE	Max Over Estimation	Max Under Estimation	σ
10	33.20	720.14	-201.7	43.26	20.67	72.05	-121.50	20.63
50	20.04	98.72	-129.0	20.07	20.33	56.31	-111.39	20.30
100	19.36	79.57	-118.2	19.37	20.33	54.23	-109.41	20.32
200	19.13	68.59	-116.4	19.13	20.30	53.33	-107.56	20.27

Table 5.11: The ratio of the standard deviation of coefficients of linear curve fitting with latitude included as variable for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	2.428
β_2	2.754
β_3	4.486
β_4	4.312
β_5	2.363

5.11.

The D-optimal sampling proves to be the sampling technique that yields smaller errors on the estimation for a small number of points and is also a more robust method when more points are sampled as shown in the previous section. The ratios of the standard deviations of the coefficients from the random sampling compared to the standard deviations of the coefficients from the D-optimal sampling shows that the D-optimal sampling has a smaller standard deviation on the coefficients. The RMS error of $19.13 \mu\text{atm}$ for the 200 randomly sampled points for the linear curve fit with latitude included as variable is almost the same as the RMS error of $19.14 \mu\text{atm}$ for the 200 randomly sampled points for the linear curve fit with latitude excluded as variable. The D-optimal results on the other hand show similar RMS error values when the latitude is included as variable for 200 D-optimally sampled points, but shows a smaller maximum over estimation which decreases from $64.41 \mu\text{atm}$ with latitude excluded to $53.33 \mu\text{atm}$ with latitude included as well as a smaller maximum under estimation that decreases from $-113.7 \mu\text{atm}$ with latitude excluded to $-107.56 \mu\text{atm}$ with latitude included as variable.

Quadratic curve fit with latitude included as variable

The quadratic equation with the variables T, MLD, Chl and Lat is fitted to the data set to investigate the degree to which it improves on the linear fit. The quadratic equation that is fit to the data is given by:

$$\begin{aligned} \text{pCO}_2 = & \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 \text{Lat} + \beta_6 T^2 + \beta_7 \text{MLD}^2 + \beta_8 \text{Chl}^2 + \beta_9 \text{Lat}^2 \\ & + \beta_{10} T \cdot \text{MLD} + \beta_{11} T \cdot \text{Chl} + \beta_{12} T \cdot \text{Lat} + \beta_{13} \text{MLD} \cdot \text{Chl} + \beta_{14} \text{MLD} \cdot \text{Lat} \\ & + \beta_{15} \text{Chl} \cdot \text{Lat}, \end{aligned} \quad (5.22)$$

and again the equation is fitted by points sampled both randomly and D-optimally.

Results of some of the curve fits obtained from the quadratic fit of the data are shown in Figures F.1 to F.24 in Appendix F on pages 234 to 241. The mean estimated pCO₂ value for all 100 curve fits are plotted in each case as well as the 95% confidence interval for the 100 runs. The results for the least squares optimization of the quadratic curve fits to the data set are summarized in Table 5.12. The mean values of the coefficients for all the terms in Equation (5.22) (β_1 to β_{15}) for all 100 curve fits in each case, for each k number of points chosen, are presented in Table F.1. The standard deviations of the coefficients are given in Table F.1. The ratio of the standard deviations of the coefficients for the 200 randomly sampled points compared to the standard deviations of coefficients for the 200 D-optimally sampled points are given in Table F.2.

The RMS error for the 200 D-optimal sampled points improved from 20.30 μatm for the linear curve fit to 18.19 μatm for the quadratic curve fit. The RMS error for the quadratic curve fit is approximately 18.2 μatm for the case where the latitude is excluded and included as variable. The inclusion of latitude thus does not make a big difference for the quadratic curve fit.

Table 5.12: Results: Quadratic curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Estimation	Max Under Estimation	σ	Mean RMSE	Max Over Estimation	Max Under Estimation	σ
25	35.36	358.5	-3483	78.93	19.72	93.07	-122.2	19.48
50	18.80	152.8	-164.2	18.89	18.18	82.29	-95.16	17.71
100	17.07	113.5	-127.8	17.09	18.21	80.83	-79.09	17.63
200	16.40	77.51	-106.6	16.40	18.19	81.83	-76.68	17.52

Table 5.13: The ratio of the standard deviation of coefficients of quadratic curve fitting with latitude included as variable for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	1.705
β_2	1.556
β_3	1.872
β_4	4.462
β_5	1.672
β_6	1.680
β_7	2.342
β_8	1.893
β_9	1.687
β_{10}	2.172
β_{11}	3.426
β_{12}	1.480
β_{13}	2.557
β_{14}	2.107
β_{15}	3.963

Cubic curve fit with latitude included as variable

The cubic equation with variables T, MLD, Chl and Lat is fitted to the data, in order to find an equation that can estimate the pCO₂ values in the ocean with known values for T, MLD, Chl and Lat. The cubic equation that is fitted to the data is given by:

$$\begin{aligned}
 \text{pCO}_2 = & \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 \text{Lat} + \beta_6 T^2 + \beta_7 \text{MLD}^2 + \beta_8 \text{Chl}^2 + \beta_9 \text{Lat}^2 + \beta_{10} T \cdot \text{MLD} \\
 & + \beta_{11} T \cdot \text{Chl} + \beta_{12} T \cdot \text{Lat} + \beta_{13} \text{MLD} \cdot \text{Chl} + \beta_{14} \text{MLD} \cdot \text{Lat} + \beta_{15} \text{Chl} \cdot \text{Lat} + \beta_{16} T^3 + \beta_{17} \text{MLD}^3 \\
 & + \beta_{18} \text{Chl}^3 + \beta_{19} \text{Lat}^3 + \beta_{20} T^2 \cdot \text{MLD} + \beta_{21} T^2 \cdot \text{Chl} + \beta_{22} T^2 \cdot \text{Lat} + \beta_{23} T \cdot \text{MLD} \cdot \text{Chl} \\
 & + \beta_{24} T \cdot \text{MLD} \cdot \text{Lat} + \beta_{25} T \cdot \text{Chl} \cdot \text{Lat} + \beta_{26} \text{MLD}^2 \cdot T + \beta_{27} \text{MLD}^2 \cdot \text{Chl} + \beta_{28} \text{MLD}^2 \cdot \text{Lat} \\
 & + \beta_{29} \text{MLD} \cdot \text{Chl} \cdot \text{Lat} + \beta_{30} \text{Chl}^2 \cdot T + \beta_{31} \text{Chl}^2 \cdot \text{MLD} + \beta_{32} \text{Chl}^2 \cdot \text{Lat} + \beta_{33} \text{Lat}^2 \cdot T \\
 & + \beta_{34} \text{Lat}^2 \cdot \text{MLD} + \beta_{35} \text{Lat}^2 \cdot \text{Chl}.
 \end{aligned}
 \tag{5.23}$$

Equation (5.23) is fitted to the data by selecting a subset of the data both randomly and D-optimally and finding the coefficients β_1 to β_{35} by means of least squares optimization. The results obtained are shown in Figures G.1 to G.24 in Appendix G on pages 247 to 254. The mean pCO₂ estimated values for all 100 curve fits in each case are plotted here together with the 95% confidence interval for the 100 runs. The results for the cubic curve fits are summarized in Table 5.14. The mean values for coefficients β_1 to β_{35} obtained in each case of k points selected are given in Table G.1. The standard deviations of the coefficients for the 100 curve fits are given in Table G.2. The ratio of the standard deviations for 200 randomly sampled points versus the standard deviations of the coefficients for the 200 D-optimally sampled points are given in Table 5.15.

The RMS error for 200 D-optimally sampled points for the cubic equation with latitude excluded as variable is 17.94 μatm , compared to the RMS error of 13.52 μatm for the 200 D-optimally sampled points for the cubic equation with latitude included as variable. The error is thus improved by 4.42 μatm , i.e. an improvement of 25%, when the latitude is included as variable. The maximum over estimation is also improved from 92.80 μatm with latitude excluded as variable to 56.71 μatm with latitude included as variable. Similarly the maximum under estimation is improved from -60.70 μatm with latitude excluded to -57.78 μatm with latitude included as variable. The inclusion of latitude as variable also improves the random sampling, where the RMS error for 200 randomly sampled points and latitude excluded as variable is 15.90 μatm and is improved to 13.51 μatm when latitude

Table 5.14: Results: Cubic curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Estimation	Max Under Estimation	σ	Mean RMSE	Max Over Estimation	Max Under Estimation	σ
50	71.68	1080	-3886	95.71	23.13	250.8	-486.2	25.28
55	53.38	1250	-3079	63.82	13.97	62.15	-76.04	13.99
100	20.86	570.9	-492.4	22.54	13.63	61.81	-62.25	13.63
200	13.51	193.5	-375.1	13.72	13.52	56.71	-57.78	13.53

is included as variable with the cubic curve fit.

Table 5.15: The ratio of the standard deviation of coefficients of cubic curve fitting with latitude included as variable for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	3.389
β_2	3.791
β_3	4.515
β_4	6.845
β_5	3.152
β_6	4.187
β_7	5.493
β_8	5.283
β_9	3.011
β_{10}	4.821
β_{11}	8.540
β_{12}	3.646
β_{13}	8.447
β_{14}	4.098
β_{15}	6.107
β_{16}	4.610
β_{17}	3.349
β_{18}	4.938
β_{19}	2.961
Continued **	

Table 5.15 – ** continued

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_{20}	5.287
β_{21}	9.301
β_{22}	3.985
β_{23}	7.371
β_{24}	4.268
β_{25}	7.410
β_{26}	4.992
β_{27}	5.479
β_{28}	4.665
β_{29}	8.217
β_{30}	6.016
β_{31}	4.921
β_{32}	5.027
β_{33}	3.674
β_{34}	3.890
β_{35}	5.412

Fourth order curve fit with latitude included as variable

The fourth order curve fit is done on the data set in order to compare the improvement of the error by adding an additional 35 terms to the 35 terms in the cubic equation. The fourth order equation is given by:

$$\begin{aligned}
 \text{pCO}_2 = & \beta_1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 \text{Lat} + \beta_6 T^2 + \beta_7 \text{MLD}^2 + \beta_8 \text{Chl}^2 + \beta_9 \text{Lat}^2 + \beta_{10} T \cdot \text{MLD} \\
 & + \beta_{11} T \cdot \text{Chl} + \beta_{12} T \cdot \text{Lat} + \beta_{13} \text{MLD} \cdot \text{Chl} + \beta_{14} \text{MLD} \cdot \text{Lat} + \beta_{15} \text{Chl} \cdot \text{Lat} + \beta_{16} T^3 + \beta_{17} \text{MLD}^3 \\
 & + \beta_{18} \text{Chl}^3 + \beta_{19} \text{Lat}^3 + \beta_{20} T^2 \cdot \text{MLD} + \beta_{21} T^2 \cdot \text{Chl} + \beta_{22} T^2 \cdot \text{Lat} + \beta_{23} T \cdot \text{MLD} \cdot \text{Chl} \\
 & + \beta_{24} T \cdot \text{MLD} \cdot \text{Lat} + \beta_{25} T \cdot \text{Chl} \cdot \text{Lat} + \beta_{26} \text{MLD}^2 \cdot T + \beta_{27} \text{MLD}^2 \cdot \text{Chl} + \beta_{28} \text{MLD}^2 \cdot \text{Lat} \\
 & + \beta_{29} \text{MLD} \cdot \text{Chl} \cdot \text{Lat} + \beta_{30} \text{Chl}^2 \cdot T + \beta_{31} \text{Chl}^2 \cdot \text{MLD} + \beta_{32} \text{Chl}^2 \cdot \text{Lat} + \beta_{33} \text{Lat}^2 \cdot T \\
 & + \beta_{34} \text{Lat}^2 \cdot \text{MLD} + \beta_{35} \text{Lat}^2 \cdot \text{Chl} + \beta_{36} \text{Chl}^4 + \beta_{37} \text{Chl}^3 \cdot \text{Lat} + \beta_{38} \text{Chl}^3 \cdot \text{MLD} + \beta_{39} \text{Chl}^3 \cdot T \\
 & + \beta_{40} \text{Chl}^2 \cdot \text{Lat}^2 + \beta_{41} \text{Chl}^2 \cdot \text{Lat} \cdot \text{MLD} + \beta_{42} \text{Chl}^2 \cdot \text{Lat} \cdot T + \beta_{43} \text{Chl}^2 \cdot \text{MLD}^2 \\
 & + \beta_{44} \text{Chl}^2 \cdot \text{MLD} \cdot T + \beta_{45} \text{Chl}^2 \cdot T^2 + \beta_{46} \text{Chl} \cdot \text{Lat}^3 + \beta_{47} \text{Chl} \cdot \text{Lat}^2 \cdot \text{MLD} \\
 & + \beta_{48} \text{Chl} \cdot \text{Lat}^2 \cdot T + \beta_{49} \text{Chl} \cdot \text{Lat} \cdot \text{MLD}^2 + \beta_{50} \text{Chl} \cdot \text{Lat} \cdot \text{MLD} \cdot T + \beta_{51} \text{Chl} \cdot \text{Lat} \cdot T^2 \\
 & + \beta_{52} \text{Chl} \cdot \text{MLD}^3 + \beta_{53} \text{Chl} \cdot \text{MLD}^2 \cdot T + \beta_{54} \text{Chl} \cdot \text{MLD} \cdot T^2 + \beta_{55} \text{Chl} \cdot T^3 + \beta_{56} \text{Lat}^4 \\
 & + \beta_{57} \text{Lat}^3 \cdot \text{MLD} + \beta_{58} \text{Lat}^3 \cdot T + \beta_{59} \text{Lat}^2 \cdot \text{MLD}^2 + \beta_{60} \text{Lat}^2 \cdot \text{MLD} \cdot T \\
 & + \beta_{61} \text{Lat}^2 \cdot T^2 + \beta_{62} \text{Lat} \cdot \text{MLD}^3 + \beta_{63} \text{Lat} \cdot \text{MLD}^2 \cdot T + \beta_{64} \text{Lat} \cdot \text{MLD} \cdot T^2 \\
 & + \beta_{65} \text{Lat} \cdot T^3 + \beta_{66} \text{MLD}^4 + \beta_{67} \text{MLD}^3 \cdot T + \beta_{68} \text{MLD}^2 \cdot T^2 + \beta_{69} \text{MLD} \cdot T^3 + \beta_{70} T^4 .
 \end{aligned} \tag{5.24}$$

The results obtained for the fourth order equation that is fit to the data are shown in Figures H.1 to H.18 on pages 262 to 267 in Appendix H, and the results for the 400 sampled points are shown in Figures 5.16 to 5.21. The pCO₂ estimation values for all 100 curve fits in each case are plotted. A summary of the results for the fourth order polynomial least squares optimization is given in Table 5.16. The mean values of coefficients β_1 to β_{70} of the fourth order fit for all 100 curve fits are given in Table H.1. The standard deviations of the coefficients are given in Table H.2. The ratio of the standard deviations of 200 randomly sampled points versus the standard deviations of the 200 D-optimally sampled points are given in Table 5.17.

Table 5.16: Results: Fourth Order curve fitting.

No of points k	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Estimation	Max Under Estimation	σ	Mean RMSE	Max Over Estimation	Max Under Estimation	σ
100	89.25	9986	-5587	124.1	9.680	74.05	-98.2	9.669
120	52.01	2219	-1791	61.24	9.418	75.82	-73.3	6.530
200	17.77	3664	-831.1	19.35	8.969	68.38	-48.76	6.182
400	10.76	302.9	-458.0	11.02	8.797	58.92	-45.06	6.018

The RMS error obtained from the 200 D-optimal sampled points for the fourth order curve fit with latitude excluded as variable is $15.36 \mu\text{atm}$ compared to the RMS error of $8.797 \mu\text{atm}$ that is obtained for the 200 D-optimal sampled points for the fourth order curve fit with latitude included as variable. The mean RMS error is thus reduced by $6.56 \mu\text{atm}$, which is an improvement of about 40% on the error when latitude is included as variable. Even for the 200 randomly sampled points, the error is improved from $13.01 \mu\text{atm}$ when latitude is excluded as variable to $10.76 \mu\text{atm}$ when latitude is included as variable. The maximum over estimation for 200 D-optimal sampled points, when latitude is excluded as variable is $92.07 \mu\text{atm}$ and is improved to $58.92 \mu\text{atm}$ when latitude is included as variable. Similarly the maximum under estimation for 200 D-optimal sampled points when latitude is excluded as variable is improved from $-61.25 \mu\text{atm}$ to $-45.06 \mu\text{atm}$ when latitude is included as variable.

Table 5.17: The ratio of the standard deviation of coefficients of fourth order curve fitting with latitude included as variable for 200 sampled points.

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_1	3.943
β_2	3.951
β_3	4.807
β_4	3.952
β_5	3.881
β_6	4.007
β_7	3.793
β_8	6.885
β_9	3.811
β_{10}	4.647
β_{11}	4.633
β_{12}	3.830
β_{13}	2.989
β_{14}	4.935
β_{15}	4.005
β_{16}	4.213
β_{17}	4.029
β_{18}	5.359
β_{19}	3.728
β_{20}	4.647
β_{21}	5.150
β_{22}	3.769
β_{23}	3.620
β_{24}	4.665
β_{25}	4.655
β_{26}	3.711
β_{27}	2.945
β_{28}	3.783
β_{29}	2.956
Continued **	

Table 5.17 – ** continued

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_{30}	8.038
β_{31}	5.652
β_{32}	6.817
β_{33}	3.726
β_{34}	4.972
β_{35}	4.036
β_{36}	3.171
β_{37}	4.677
β_{38}	4.755
β_{39}	6.774
β_{40}	6.511
β_{41}	5.771
β_{42}	7.792
β_{43}	3.483
β_{44}	6.173
β_{45}	6.072
β_{46}	4.039
β_{47}	2.912
β_{48}	4.652
β_{49}	3.001
β_{50}	3.584
β_{51}	5.041
β_{52}	3.124
β_{53}	3.571
β_{54}	4.107
β_{55}	5.769
β_{56}	3.629
β_{57}	4.908
β_{58}	3.641
β_{59}	3.653
β_{60}	4.642
Continued **	

Table 5.17 – ** continued

Coefficient	$\frac{\sigma_{random}}{\sigma_{D-optimal}}$
β_{61}	3.568
β_{62}	4.383
β_{63}	3.514
β_{64}	4.617
β_{65}	3.851
β_{66}	2.583
β_{67}	4.380
β_{68}	3.939
β_{69}	4.677
β_{70}	4.606

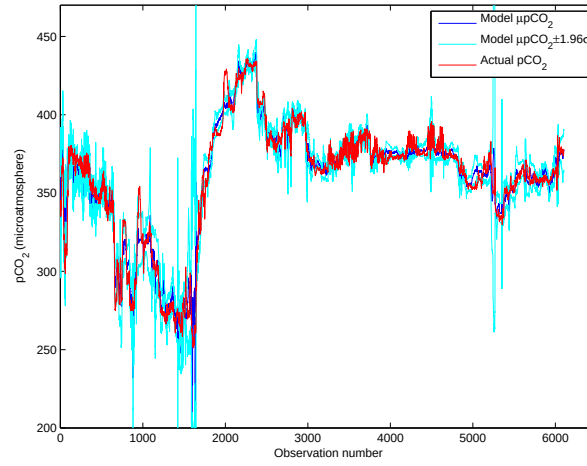


Figure 5.16: Fourth order curve fit results. 400 Random points sampled.

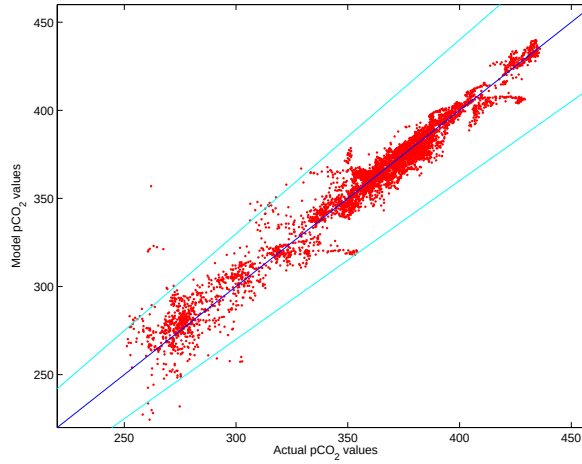


Figure 5.17: Fourth order: curve fit results with 400 random points sampled. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

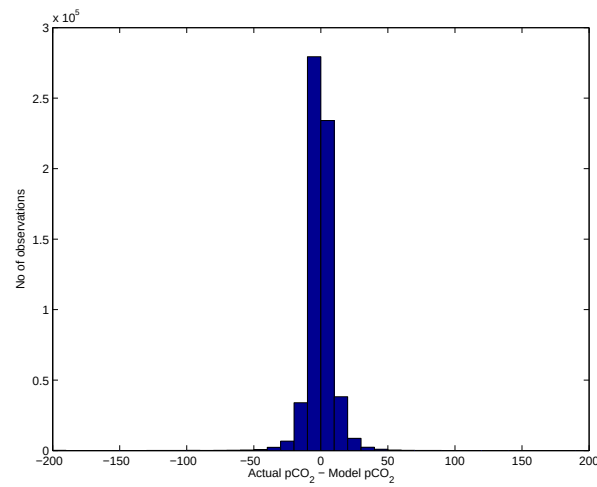


Figure 5.18: Fourth order curve fit results. Histogram of errors for all 100 runs and 400 random points sampled.

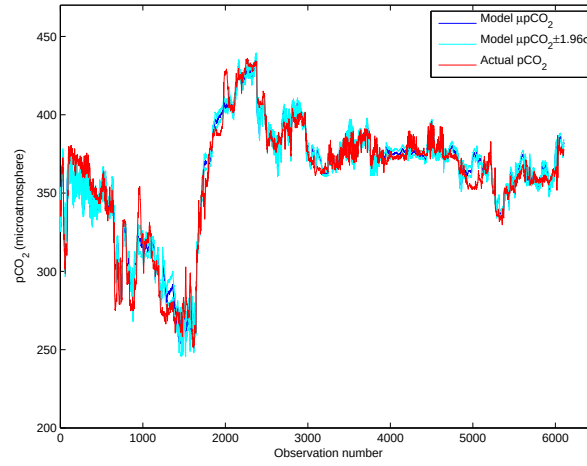


Figure 5.19: Fourth order curve fit results. 400 D-optimal sampled points.

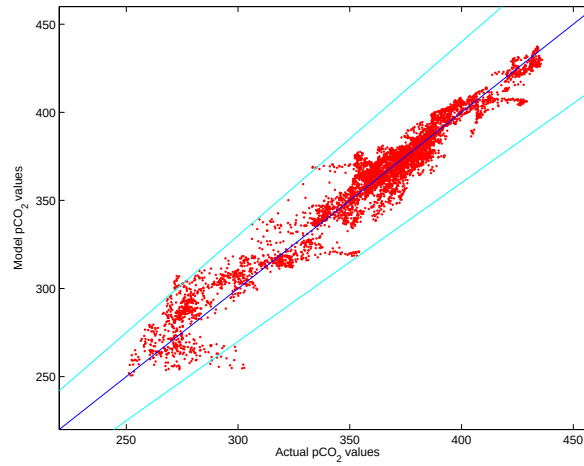


Figure 5.20: Fourth order: curve fit results with 400 D-optimal sampled points. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

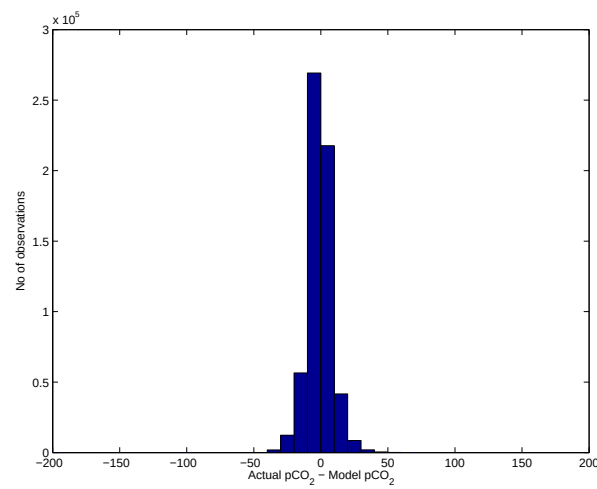


Figure 5.21: Fourth order curve fit results. Histogram of errors for all 100 runs and 400 D-optimal sampled points.

Removing terms

The fourth order equation fits the data with a reasonably small error. The terms in the fourth order equation are removed one by one in order to determine which of the terms has the biggest influence on the estimation of the $p\text{CO}_2$.

The fourth order equation is considered. The error of the estimation of $p\text{CO}_2$ made by the fourth order equation is determined by fitting all 6103 data points with the fourth order equation, and calculating the RMSE of this fit. The terms are removed one-by-one in order to determine the error made in the $p\text{CO}_2$ prediction when each one of the terms are removed individually. The term that, when it is removed, makes the least difference on the error of the prediction, is removed from the equation. Once the term is removed, all the remaining terms are once again removed one-by one, in order to determine which term can be removed that will make the smallest difference on the error of the prediction. This is repeated until only one term is left. The results are given in Table 5.18. The error on the estimation of the $p\text{CO}_2$ as the terms are removed from the fourth order equation is shown in Figure 5.22.

In Table 5.18 it can be seen that the terms that are removed first do not play such a big role in the estimation of the $p\text{CO}_2$, whereas the terms that are removed last plays the biggest role in the estimation of the $p\text{CO}_2$. It can also be seen that the last 35 terms that are removed, already yields a RMS error of $8.602 \mu\text{atm}$ which is a smaller error on the $p\text{CO}_2$ estimation than the RMS error of $13.52 \mu\text{atm}$ that is obtained from the 200 D-optimally sampled points with the 35 terms of the cubic equation. It is particularly noteworthy how important latitude is as a variable. Note that 14 of the 20 most important terms contain latitude, as well as 23 of the most important 35 terms. The six most important terms all contain latitude. Note that the 18 most important terms do not contain MLD, but 12 of these terms contain latitude.

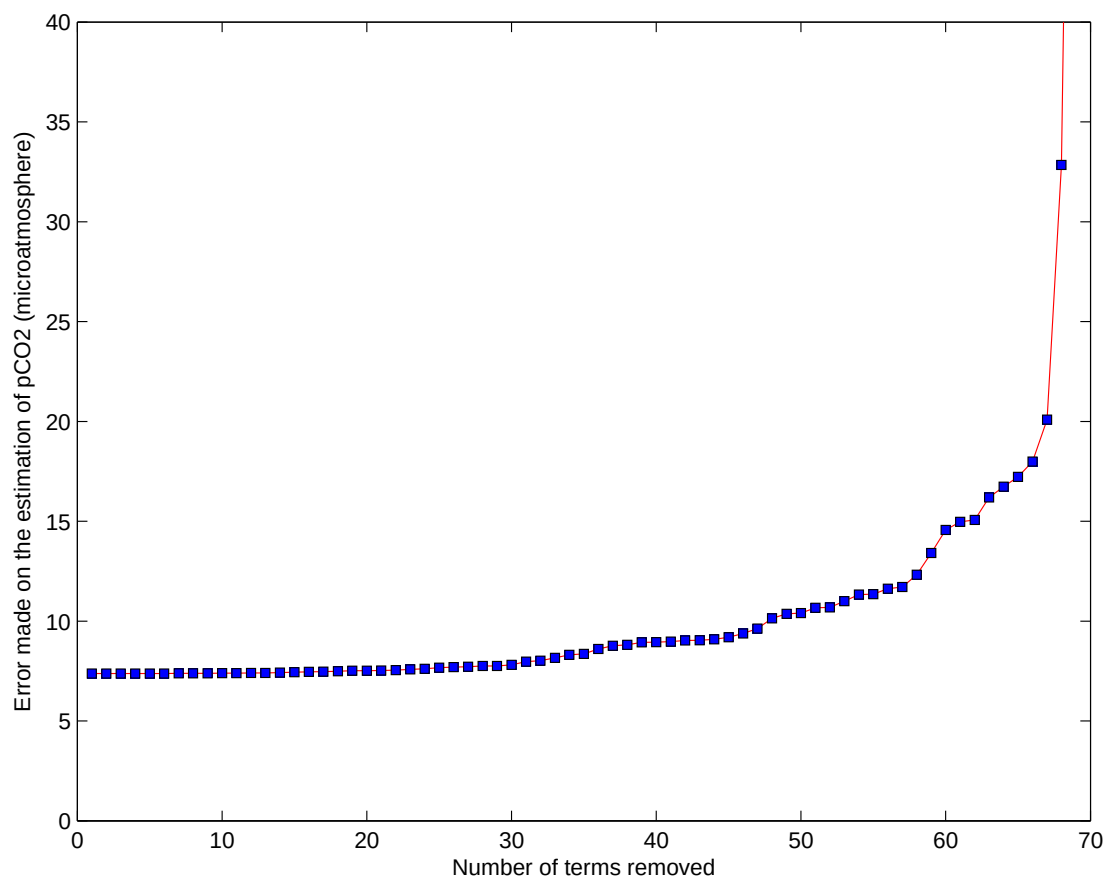


Figure 5.22: The increase of the error as the terms are removed from the fourth order polynomial.

Table 5.18: Table of RMSE when systematically removing terms from the fourth order equation with latitude included.

No	Coefficient	Term	RMSerror
1	β_{19}	Lat^3	7.375
2	β_{21}	$\text{T}^2 \cdot \text{Chl}$	7.375
3	β_6	T^2	7.376
4	β_{44}	$\text{Chl}^2 \cdot \text{MLD} \cdot \text{T}$	7.376
5	β_{66}	MLD^4	7.376
6	β_{57}	$\text{Lat}^3 \cdot \text{MLD}$	7.382
7	β_{39}	$\text{Chl}^3 \cdot \text{T}$	7.385
8	β_{58}	$\text{Lat}^3 \cdot \text{T}$	7.389
9	β_{55}	$\text{Chl} \cdot \text{T}^3$	7.391
10	β_{48}	$\text{Chl} \cdot \text{Lat}^2 \cdot \text{T}$	7.394
11	β_{49}	$\text{Chl} \cdot \text{Lat} \cdot \text{MLD}^2$	7.399
12	β_{63}	$\text{Lat} \cdot \text{MLD}^2 \cdot \text{T}$	7.408
13	β_{68}	$\text{MLD}^2 \cdot \text{T}^2$	7.426
14	β_{53}	$\text{Chl} \cdot \text{MLD}^2 \cdot \text{T}$	7.446
15	β_{54}	$\text{Chl} \cdot \text{MLD} \cdot \text{T}^2$	7.458
16	β_{38}	$\text{Chl}^3 \cdot \text{MLD}$	7.467
17	β_3	MLD	7.467
18	β_{43}	$\text{Chl}^2 \cdot \text{MLD}^2$	7.487
19	β_{26}	$\text{MLD}^2 \cdot \text{T}$	7.512
20	β_{67}	$\text{MLD}^3 \cdot \text{T}$	7.513
21	β_{52}	$\text{Chl} \cdot \text{MLD}^3$	7.529
22	β_{27}	$\text{MLD}^2 \cdot \text{Chl}$	7.543
23	β_{17}	MLD^3	7.584
24	β_{62}	$\text{Lat} \cdot \text{MLD}^3$	7.609
25	β_2	T	7.665
26	β_{70}	T^4	7.7001
27	β_{16}	T^3	7.712
28	β_{59}	$\text{Lat}^3 \cdot \text{MLD}^2$	7.754
29	β_7	MLD^2	7.757
Continued **			

Table 5.18 – ** continued

No	Coefficient	Term	RMSerror
30	β_{28}	MLD ² ·Lat	7.813
31	β_{33}	Lat ² ·T	7.968
32	β_{60}	Lat ² ·MLD·T	8.015
33	β_8	Chl ²	8.164
34	β_{50}	Chl·Lat·MLD·T	8.316
35	β_{23}	T·MLD·Chl	8.354
36	β_{65}	Lat·T ³	8.602
37	β_{64}	Lat·MLD·T ²	8.765
38	β_{69}	MLD·T ³	8.817
39	β_{13}	MLD·Chl	8.939
40	β_{20}	T ² ·MLD	8.948
41	β_{24}	T·MLD·Lat	8.977
42	β_{42}	Chl ² ·Lat·T	9.032
43	β_{45}	Chl ² ·T ²	9.035
44	β_{25}	T·Chl·Lat	9.085
45	β_{51}	Chl·Lat·T ²	9.191
46	β_{41}	Chl ² ·Lat·MLD	9.382
47	β_{31}	Chl ² ·MLD	9.624
48	β_{10}	T·MLD	10.13
49	β_{47}	Chl·Lat ² ·MLD	10.36
50	β_{29}	MLD·Chl·Lat	10.40
51	β_{34}	Lat ² ·MLD	10.67
52	β_{14}	MLD·Lat	10.69
53	β_{36}	Chl ⁴	11.00
54	β_{11}	T·Chl	11.33
55	β_{30}	Chl ² ·T	11.35
56	β_{37}	Chl ³ ·Lat	11.62
57	β_{18}	Chl ³	11.71
58	β_{46}	Chl·Lat ³	12.33
59	β_{56}	Lat ⁴	13.41
60	β_{12}	T·Lat	14.47
Continued **			

Table 5.18 – ** continued

No	Coefficient	Term	RMSerror
61	β_{22}	$T^2 \cdot \text{Lat}$	14.97
62	β_1	1	15.07
63	β_{40}	$\text{Chl}^2 \cdot \text{Lat}^2$	16.19
64	β_4	Chl	16.72
65	β_{32}	$\text{Chl}^2 \cdot \text{Lat}$	17.23
66	β_{61}	$\text{Lat}^2 \cdot T^2$	17.99
67	β_{35}	$\text{Lat}^2 \cdot \text{Chl}$	20.09
68	β_{15}	$\text{Chl} \cdot \text{Lat}$	32.84
69	β_9	Lat^2	84.85
70	β_5	Lat	No Model

5.4 Radial basis function interpolation

5.4.1 Introduction

Radial basis function interpolation is a method that can be used to create a smooth interpolation function in a domain. Global radial basis functions consider all the points in the domain when interpolating. Local radial basis functions only consider points within a certain radius from a point when interpolating. Global radial basis functions yield large dense matrices which result in time consuming and costly computations, especially when large data sets are considered. Local radial basis functions have less dense matrices, because the interpolation function becomes zero at a certain radius from a point. When considering large data sets, the computational costs will be less if local radial basis functions are used to construct the interpolation function. Local RBFs may also yield more accurate interpolation than global RBFs for a large data set with variability of the relationship between the variables across the domain. In this case the local RBFs only consider points in the region of a point, and can thus interpolate the behaviour in the region around the point. In this research, the main aim is to use the data from the entire ocean to construct an equation that will estimate the pCO₂ in the ocean. The data set that is used in this case will thus be a large data set and global support functions may be inadequate for this purpose. Local radial basis functions will allow for an interpolation function to be set up for large data sets with reasonable computational costs and increased accuracy. Local radial basis functions will also allow the interpolation to take regional variations in the data into account.

There are various local radial basis functions that can be used. In this research, the aim is to construct an equation with variables T, MLD, Chl and possibly Lat that will be able to predict the pCO₂ in the ocean. The variables are thus in $\mathbb{R}^5$. Thus, a local support function that is radial on $\mathbb{R}^5$ is needed. Wu's compactly supported functions have polynomial support and are strictly positive definite functions on their support [35]. For $\mathbb{R}^5$, Wu's local radial basis function is given by:

$$\phi = (1 - r)^5(8 + 40r + 48r^2 + 25r^3 + 5r^4) , \quad (5.25)$$

where $r_j = \| \mathbf{x} - \mathbf{x}_{b_j} \|$ and $\mathbf{x}_{b_j} = [T_{b_j}, \text{MLD}_{b_j}, \text{Chl}_{b_j}, \text{Lat}_{b_j}]$. This is the local support function that is used in this dissertation. The interpolation equation is given by

$$s(\mathbf{x}) = \sum_{j=1}^{n_b} \alpha_j \phi(\| \mathbf{x} - \mathbf{x}_{b_j} \|) + p(\mathbf{x}) , \quad (5.26)$$

where $\mathbf{x} = (\text{T}, \text{MLD}, \text{Chl}, \text{Lat})$ in this case and n_b is the number of points that the RBF is constructed with.

In this part of the research the polynomial p is considered to be a linear polynomial with the variables T, MLD, Chl and Lat, as follows

$$p = \beta_1 1 + \beta_2 \text{T} + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 \text{Lat} = \begin{bmatrix} 1 & \text{T} & \text{MLD} & \text{Chl} & \text{Lat} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}. \quad (5.27)$$

The system considered for the radial basis function interpolation can be written as

$$\begin{bmatrix} \mathbf{d}_b \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{b,b} & \mathbf{P}_b \\ \mathbf{P}_b^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}. \quad (5.28)$$

In this system, $\boldsymbol{\alpha}$ contains the coefficients α_j and $\boldsymbol{\beta}$ contains the coefficients for the polynomial p . $\mathbf{M}_{b,b}$ is an $n_b \times n_b$ matrix that contains the calculated radial basis functions $\phi_{b,b_j} = \phi(\|\mathbf{x}_b - \mathbf{x}_{b_j}\|)$ and $\mathbf{P}_b$ is an $n_b \times 5$ matrix with row j represented by $\mathbf{x}_{b_j} = [x_{b_j}^1, x_{b_j}^2, \dots, x_{b_j}^d]$. In this dissertation, the system in Equation (5.28) is as follows:

$$\begin{bmatrix} \text{pCO}_{2_1} \\ \text{pCO}_{2_2} \\ \vdots \\ \text{pCO}_{2_m} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \phi(\|\mathbf{x}_1 - \mathbf{x}_1\|) & \phi(\|\mathbf{x}_1 - \mathbf{x}_2\|) & \dots & \phi(\|\mathbf{x}_1 - \mathbf{x}_m\|) & 1 & T_1 & M_1 & C_1 & L_1 \\ \phi(\|\mathbf{x}_2 - \mathbf{x}_1\|) & \phi(\|\mathbf{x}_2 - \mathbf{x}_2\|) & \dots & \phi(\|\mathbf{x}_2 - \mathbf{x}_m\|) & 1 & T_2 & M_2 & C_2 & L_2 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi(\|\mathbf{x}_m - \mathbf{x}_1\|) & \phi(\|\mathbf{x}_m - \mathbf{x}_2\|) & \dots & \phi(\|\mathbf{x}_m - \mathbf{x}_m\|) & 1 & T_m & M_m & C_m & L_m \\ 1 & 1 & \dots & 1 & 0 & 0 & \cdot & \cdot & 0 \\ T_1 & T_2 & \dots & T_m & 0 & 0 & \cdot & \cdot & 0 \\ M_1 & M_2 & \dots & M_m & 0 & 0 & \cdot & \cdot & 0 \\ C_1 & C_2 & \dots & C_m & 0 & 0 & \cdot & \cdot & 0 \\ L_1 & L_2 & \dots & L_m & 0 & 0 & \cdot & \cdot & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}. \quad (5.29)$$

In Equation (5.29), ϕ is the radial basis function, in this case the local radial basis function of Wu given in Equation (5.25). Furthermore, T refers to the sea surface temperature, M refers to the mixed layer depth, C refers to the chlorophyll-a concentration and L refers to the latitude. In the vector on

the left, pCO_2_j refers to the j^{th} pCO_2 value that is known, with the corresponding (T_j, M_j, C_j, L_j) values from the SANAE49L6 data set.

In this case, the interpolation function will be

$$\text{pCO}_2(\mathbf{x}) = \sum_{j=1}^n \alpha_j ((1-r_j)^5 (8+40r_j+48r_j^2+25r_j^3+5r_j^4) + \beta_1 1 + \beta_2 T + \beta_3 \text{MLD} + \beta_4 \text{Chl} + \beta_5 \text{Lat}), \quad (5.30)$$

where $r_j = \|\mathbf{x} - \mathbf{x}_j\|$ and $\mathbf{x} = (T, \text{MLD}, \text{Chl}, \text{Lat})$.

The system in Equation (5.29) is solved using matrix manipulations in order to find the coefficients α_j and β_i for $j = 1, 2, \dots, m$ and $i = 1, 2, \dots, 5$. This system can be solved directly since the size of the matrices are moderate.

5.4.2 D-optimal sampled points

One way to sample the n_b points from the complete data set, that are used as interpolation points in the RBF, is to use D-optimal sampling (as previously described).

The following system is considered:

$$\mathbf{y} = \mathbf{R}\boldsymbol{\gamma}, \quad (5.31)$$

which can be written as:

$$\begin{bmatrix} \mathbf{d}_b \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{b,b} & \mathbf{P}_b \\ \mathbf{P}_b^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}. \quad (5.32)$$

The Fisher information matrix, $\mathbf{M}$, is then given by:

$$\begin{aligned} \mathbf{M} &= \mathbf{S}^T \mathbf{P}^{-1} \mathbf{S} \\ &= \mathbf{R}^T \left[\frac{1}{\sigma} \mathbf{I} \right] \mathbf{R} \\ &= \frac{1}{\sigma} \mathbf{R}^T \mathbf{R}, \end{aligned} \quad (5.33)$$

where $\mathbf{S}$ is the sensitivity matrix and $\mathbf{P}$ is the variance-covariance matrix of the observation error.

It is assumed that the observations all have a constant variance, such that the variance-covariance matrix of the observations becomes $\mathbf{P} = \sigma \mathbf{I}$. The sensitivity matrix is defined as $\mathbf{S} = \frac{\delta \mathbf{y}}{\delta \gamma}$.

The objective function for the D-optimal sampling in this case then becomes

$$\begin{aligned} \max(\det(\mathbf{M})) &= \max\left(\det\left(\frac{1}{\sigma} \mathbf{R}^T \mathbf{R}\right)\right) \\ &= \max(\det(\mathbf{R}^T \mathbf{R})) , \end{aligned} \quad (5.34)$$

or expanded,

$$\max\left(\det\left(\begin{bmatrix} \mathbf{M}_{b,b} & \mathbf{P}_b \\ \mathbf{P}_b^T & \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{M}_{b,b} & \mathbf{P}_b \\ \mathbf{P}_b^T & \mathbf{0} \end{bmatrix}\right)\right) . \quad (5.35)$$

In the D-optimal sampling with the original data set, the distances between the points become too large, and because of computational restrictions, the determinant of the Fisher information matrix has the value of infinity. This causes the D-optimal sampling to be inefficient in sampling the optimal points in the data set. There thus exists a need to scale the data set in order to perform efficient sampling. The data set is scaled, using the following scaling:

$$x_{\text{scaled}} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \cdot 2 , \quad (5.36)$$

in order to scale all the values for the variables between zero and two, where x can be any one of the variables (T, MLD, Chl, Lat). This scaled data set is then used to do the D-optimal sampling, as uniform scaling on all variables will not influence the points chosen in the sample space. No matter what the uniform scaling is, the optimal distribution of points that will reduce the effect of noise on the data will still be selected using D-optimal sampling. The RBF interpolation for both the random and the D-optimal sampled points are then done on the scaled data set in order to make an equal comparison of the performance of each.

5.4.3 RBF interpolation with variables T, MLD and Chl

RBF interpolation is done on the data set using the variables T, MLD, and Chl. The interpolation function is as in Equation (5.26) with $\mathbf{x} = [\text{T}, \text{MLD}, \text{Chl}]$, $\mathbf{x}_{b_j} = [\text{T}_{b_j}, \text{MLD}_{b_j}, \text{Chl}_{b_j}]$ and n_b is selected as 50, 100 and 200.

Table 5.19: Results of RBF interpolation with variables T,MLD and Chl with the 0-2 scaled data set.

No of points n_b	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ
50	22.63	230.87	-277.5	23.38	15.95	112.8	-103.4	15.98
100	18.48	254.9	-353.4	19.27	12.54	88.97	-87.28	12.58
200	14.73	424.9	-406.4	14.93	9.617	92.27	-81.26	9.63

The results are obtained by selecting n_b points from the data set by means of random or D-optimal sampling, and creating the interpolation function from these points. This is repeated 100 times and the mean RMS error is recorded, as well as the maximum over estimation and the maximum under estimation. The standard deviation, σ , of the error on the estimation is also computed. The mean pCO₂ estimation value is obtained and plotted in the figures along with the 95% confidence interval on the 100 runs.

The results of the random and D-optimal sampling are summarized in Table 5.19. Examples of the results are shown in Figures I.1 to I.12 in Appendix I and the results for 200 sampled points are shown in Figures 5.23 to 5.28. The first figure in each case is the mean pCO₂ value for the 100 curve fits plotted as well as the 95% confidence interval in each case. The second figure is the plot of the actual data versus the model output, with the 10% error lines that indicate the amount of mean pCO₂ values that fall inside the 10% error margin. The third figure in each case is the histogram of the prediction errors in the pCO₂ estimation to indicate the distribution of the prediction errors of the model.

From the results it can be seen that as the number of points sampled from the data set increases, the error on the estimation of the pCO₂ from the interpolation function created from these points decreases. The RMS error of 14.73 μatm obtained with 200 randomly sampled points is larger than the error of 9.617 μatm that is obtained with 200 D-optimal sampled points. This once again proves that the D-optimal sampling performs better than the random sampling in this application. Also, the D-optimal sampling yields results with less variance and from the figures it can be seen that the D-optimal sampled points yield interpolation functions that estimate the pCO₂ with a smaller error. The RMS error of 9.617 μatm obtained with 200 D-optimally sampled points with the RBF interpolation function is smaller than the RMS error of 15.36 μatm obtained with 200 D-optimally sampled points with the fourth order curve fit with latitude excluded as variable.

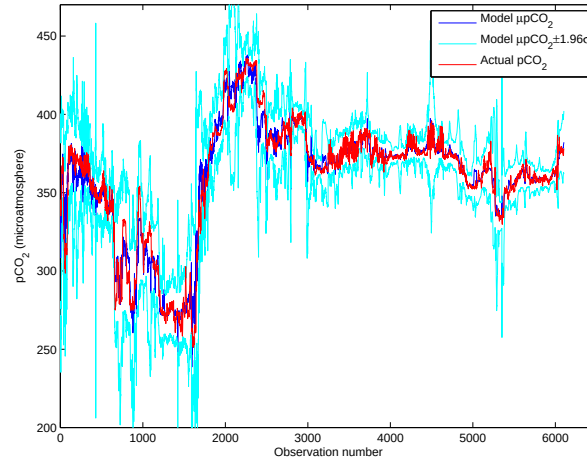


Figure 5.23: RBF interpolation results. 200 Random points sampled with the 0-2 scaled data set.

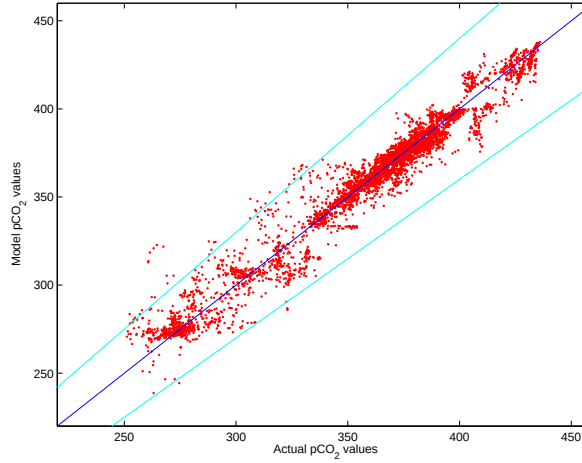


Figure 5.24: RBF interpolation results with the 0-2 scaled data set and 200 randomly sampled points. The degree to which the modelled points fall in the 10 % error margin is shown. The mean pCO_2 estimation for all 100 runs is shown here.

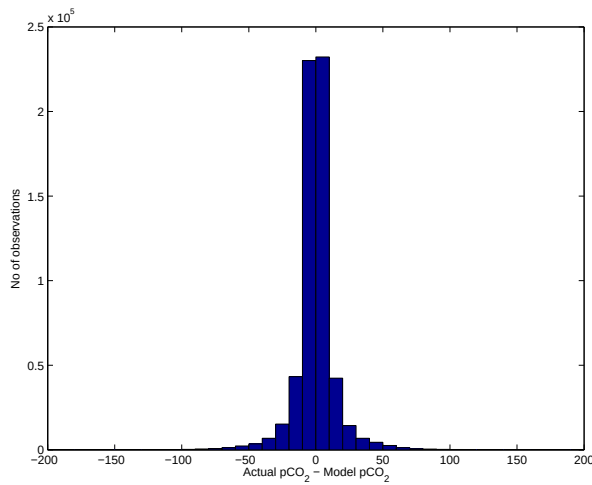


Figure 5.25: RBF interpolation results. Histogram of errors for all 100 runs and 200 Random points sampled with the 0-2 scaled data set.

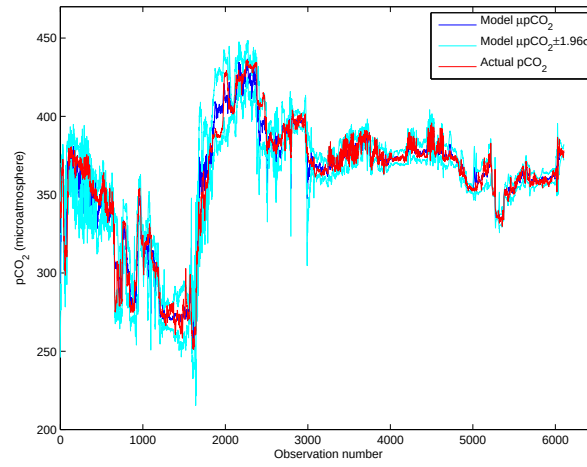


Figure 5.26: RBF interpolation results. 200 D-optimal sampled points with the 0-2 scaled data set.

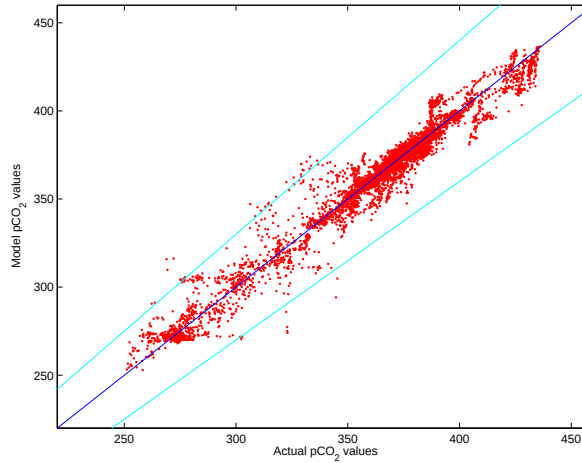


Figure 5.27: RBF interpolation results with D-optimal sampled points with the 0-2 scaled data set. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

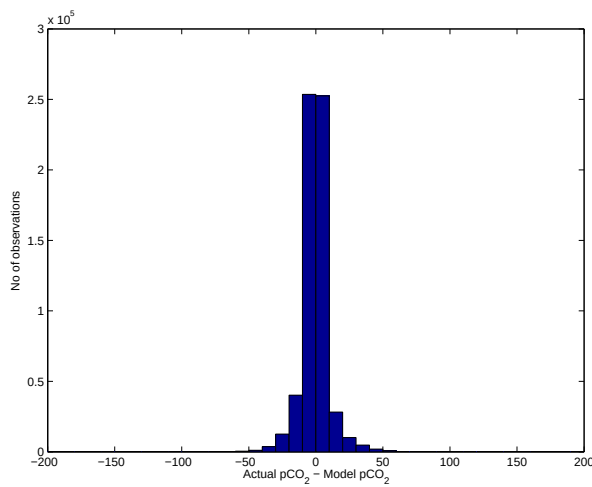


Figure 5.28: RBF interpolation results. Histogram of errors for all 100 runs and 200 D-optimal sampled points with the 0-2 scaled data set.

Optimal scaling of variables

The variables in the original data set all have a different range of values, as shown in Table 5.21. When using the locally supported radial basis function it has a radial influence of one. This means that a point is influenced by all points that is within a radius of one from this point. This definition of a distance of one may differ between all of the variables. Since for example, the chlorophyll-a concentration values has a range of five, and the MLD values on the other hand has a range of more than 100, the radius of one will not have the same meaning for all of the variables. It is suggested that by scaling the variables optimally, the effect of each variable on the pCO₂ as well as the variability in the values of each variable will be taken into account in order to reduce the error made in estimating pCO₂ values from these variables.

To find this optimal scaling an optimization algorithm is used in order to find the scaling that minimizes the error made on the pCO₂ estimation. As initial guess for the optimization algorithm, the scaling is such that all the variables are scaled between zero and two. The scaled variables (*) are then specified in the RBF interpolation as

$$\begin{aligned} T_j^* &= a_1 \left[\frac{T_j - T_{min}}{T_{max} - T_{min}} \right] \\ MLD_j^* &= a_2 \left[\frac{MLD_j - MLD_{min}}{MLD_{max} - MLD_{min}} \right] \\ Chl_j^* &= a_3 \left[\frac{Chl_j - Chl_{min}}{Chl_{max} - Chl_{min}} \right], \end{aligned} \tag{5.37}$$

where the values a_1 , a_2 and a_3 are determined by minimizing the error of the estimation of pCO₂. Once an optimal scaling is found, a new subset containing k points are determined with D-optimal sampling. The optimization algorithm to find a_1 , a_2 and a_3 is repeated for the new k points using the previously determined scaling as initial guess. This is repeated until the scaling indicates convergence, as seen in Table 5.20. The optimal scaling shown in Table 5.20 is expected to vary since for each iteration a new set of points are selected to determine the optimal scaling. To avoid this variation would require a much more complex optimization problem formulation, but it falls outside the scope of this dissertation. Nevertheless, representative values for the optimal scaling are given in Table 5.21. When this scaling is applied to the variables, the error made by the estimation of pCO₂ is shown in Figures J.1 to J.12 in Appendix J and the results for the 200 sampled points are shown in Figures 5.29 to 5.34. These results are summarized in Table 5.22.

Table 5.20: Optimal scaling found for different points selected by D-optimal sampling.

Iteration	a_1	a_2	a_3
1	2	2	2
2	2.899	0.977	3.511
3	2.737	2.552	1.830
4	3.270	-0.931	3.126
5	4.366	-1.594	1.262
6	4.4909	1.162	4.250
7	3.423	1.178	5.406
8	3.399	0.770	5.406
9	2.089	1.193	4.410

Table 5.21: Range of values for the variables.

Scaling	T_{min}	T_{max}	MLD_{min}	MLD_{max}	Chl_{min}	Chl_{max}
Original Data	0.2791	21.3027	13.1542	127.9258	0.1200	5.1350
Scaled Data	0	3.5	0	1	0	4.5

The results show that the D-optimal sampled points reduces the variance on the results and also reduces the error on the estimation of the pCO_2 . As the number of points sampled increases, the error on the estimation decreases. The RMS error of $13.66 \mu atm$ obtained with 200 randomly sampled points with the optimal scaling is smaller than the RMS error of $14.73 \mu atm$ obtained with 200 randomly sampled points without the optimal scaling. However, the RMS error of $9.012 \mu atm$ for the 200 D-optimal sampled points with the optimal scaling is only slightly smaller than the RMS error of $9.617 \mu atm$ for 200 D-optimal sampled points without the scaling. The optimal scaling on the three variables, T, MLD and Chl does not make a significant difference in the error on the pCO_2 estimation.

Table 5.22: Results of RBF interpolation for the variables T,MLD and Chl with optimally scaled variables.

No of points n_b	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ
50	17.40	164.5	-235.4	17.58	14.98	122.07	-108.6	15.02
100	16.12	201.0	-409.0	16.98	11.20	124.0	-109.6	11.23
200	13.66	389.7	-309.6	13.97	9.012	101.88	-86.88	9.026

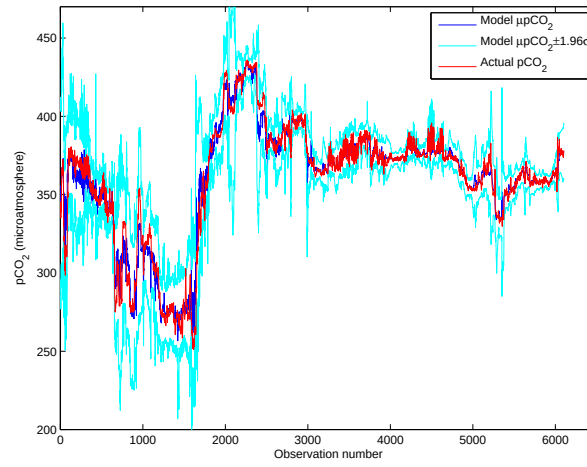


Figure 5.29: RBF interpolation results. 200 Random points sampled with the optimally scaled variables.

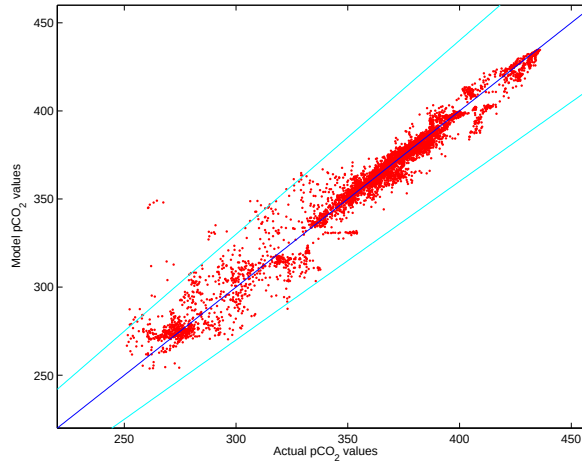


Figure 5.30: RBF interpolation results with 200 random points sampled and optimally scaled variables. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

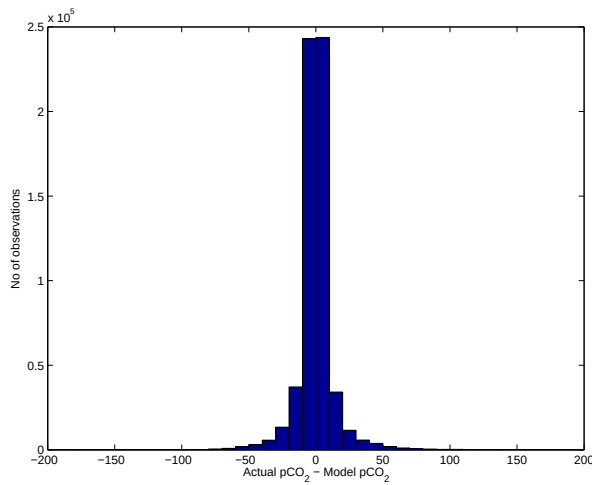


Figure 5.31: RBF interpolation results. Histogram of errors for all 100 runs and 200 Randomly sampled points with optimally scaled variables

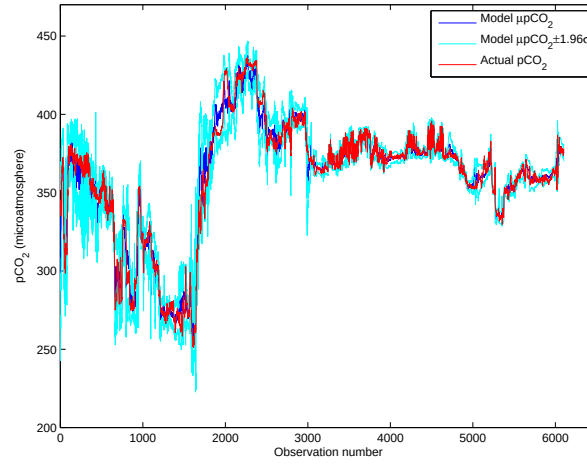


Figure 5.32: RBF interpolation results. 200 D-optimal sampled points with the optimally scaled variables.

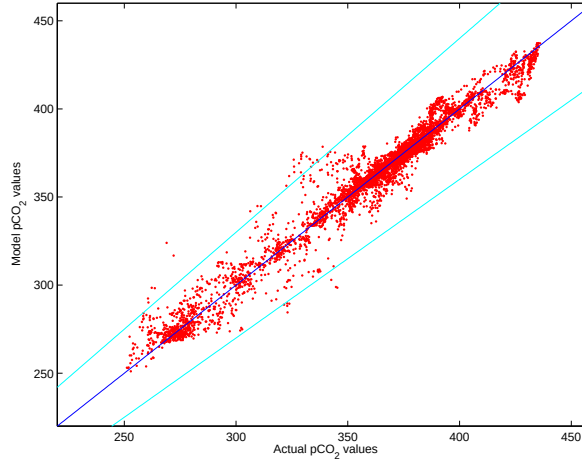


Figure 5.33: RBF interpolation results with 200 D-optimal sampled points and optimally scaled variables. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

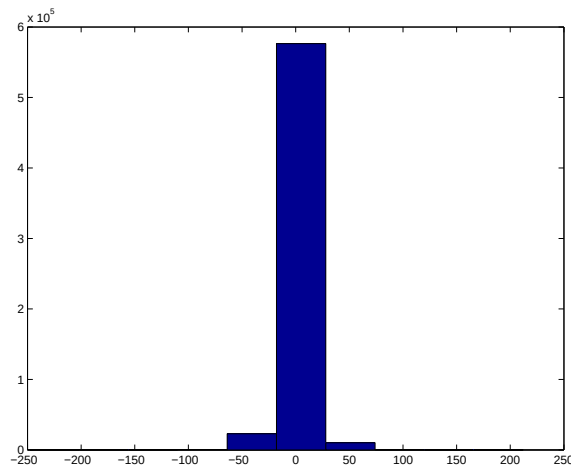


Figure 5.34: RBF interpolation results. Histogram of errors for all 100 runs and 200 Randomly sampled points with optimally scaled variables

5.4.4 RBF interpolation with variables T, MLD, Chl and Lat

The ocean is divided into regions in which the relationship between the variables differ. There are regions that are separated by the ocean currents and the oceanic properties that differ in these regions. Oceanographers suggest that the variability between the regions should be taken into account in the investigation of the relationship between the oceanic properties. The local support radial basis functions have the ability to take into account only the points at a specified radius from a point when interpolating. By including latitude as variable in the RBF interpolation, the regions can implicitly be accounted for by taking into account only the points within a specified radius in latitude as well, and therefore automatically grouping the areas in the ocean into regions where the relationship between the variables differ. The properties of the local support RBFs thus allows for different regions in the ocean to be taken account for. Thus, latitude is included as variable in the RBF interpolation.

RBF interpolation is done on the data set using the variables T, MLD, Chl and Lat. The interpolation function is as in Equation (5.26) with $\mathbf{x} = [T, \text{MLD}, \text{Chl}, \text{Lat}]$, $\mathbf{x}_{b_j} = [T_{b_j}, \text{MLD}_{b_j}, \text{Chl}_{b_j}, \text{Lat}_{b_j}]$ and n_b in this case is the k points sampled.

The results are obtained by selecting k points from the data set by means of random or D-optimal sampling, and creating the interpolation function from these points. This is repeated 100 times and the mean RMS error is recorded, as well as the maximum over estimation and the maximum under estimation by the interpolation function and the standard deviation σ of the error on the estimation. The mean pCO_2 estimation value is obtained and plotted in the figures along with the 95% confidence interval on the 100 runs. The results for the random and D-optimal sampling are summarized in Table 5.23.

Examples of the results are shown in Figures K.1 to K.12 in Appendix K and the results for 200 sampled points are shown in Figures 5.35 to 5.40. The first figure in each case is the mean pCO_2 values for the 100 curve fits plotted as well as the 95% confidence interval in each case. The second figure is the plot of the actual data versus the model output, with the 10% error lines that indicate the amount of mean pCO_2 values that fall inside the 10% error margin. The third figure in each case is the histogram of the prediction errors in the pCO_2 estimation to indicate the distribution of the prediction errors of the model.

The RMS error of $15.61 \mu\text{atm}$ obtained with 200 randomly sampled points with latitude included as variable in the RBF interpolation is larger than the RMS error of $14.73 \mu\text{atm}$ obtained with 200 randomly sampled points in Table 5.19 for the RBF interpolation with latitude excluded as variable. The D-optimal sampling, however, shows improved results when the latitude is included as variable.

Table 5.23: Results of RBF interpolation with variables T,MLD, Chl and Lat with the 0-2 scaled data set.

No of points n_b	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ
50	21.39	324.0	-549.9	23.54	12.48	81.22	-86.26	12.52
100	18.54	496.3	-326.2	19.47	9.784	72.45	-63.38	9.766
200	15.61	705.4	-319.2	16.11	6.761	121.7	-63.13	6.776

The RMS error is reduced from $9.617 \mu\text{atm}$ when latitude is excluded to $6.761 \mu\text{atm}$ when latitude is included with 200 D-optimal sampled points with the RBF interpolation.

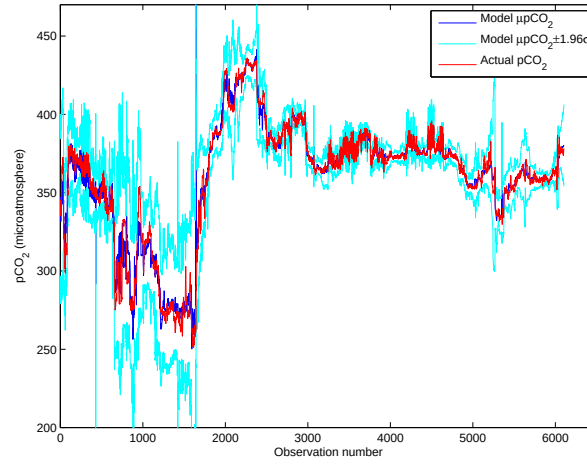


Figure 5.35: RBF interpolation results. 200 Random points sampled with the 0-2 scaled data set.

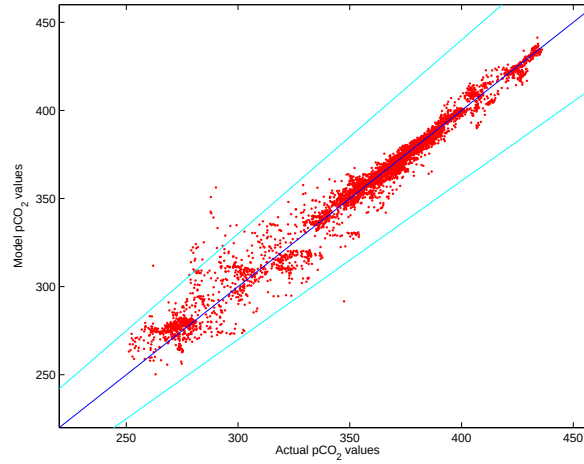


Figure 5.36: RBF interpolation results with the 0-2 scaled data set and 200 randomly sampled points. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

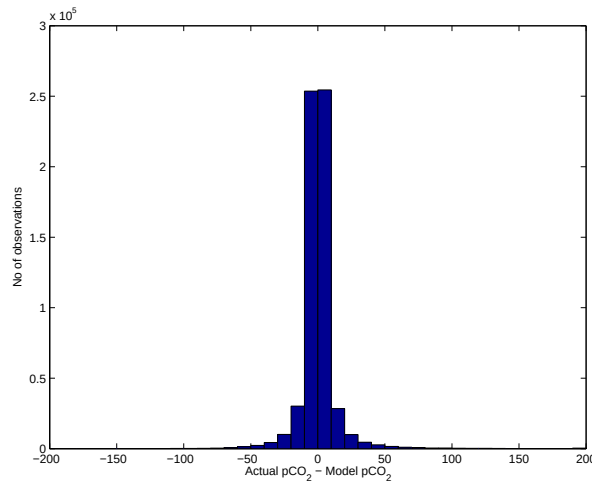


Figure 5.37: RBF interpolation results. Histogram of errors for all 100 runs and 200 Random points sampled with the 0-2 scaled data set.

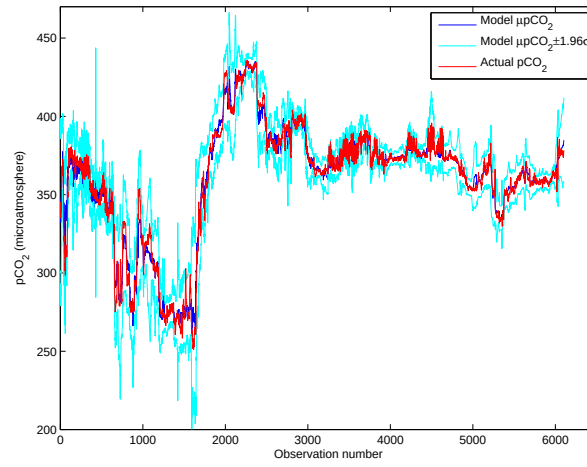


Figure 5.38: RBF interpolation results. 200 D-optimal sampled points with the 0-2 scaled data set.

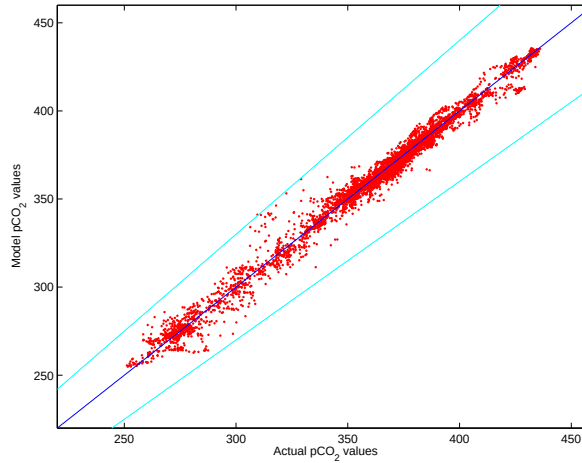


Figure 5.39: RBF interpolation results with D-optimal sampled points with the 0-2 scaled data set. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

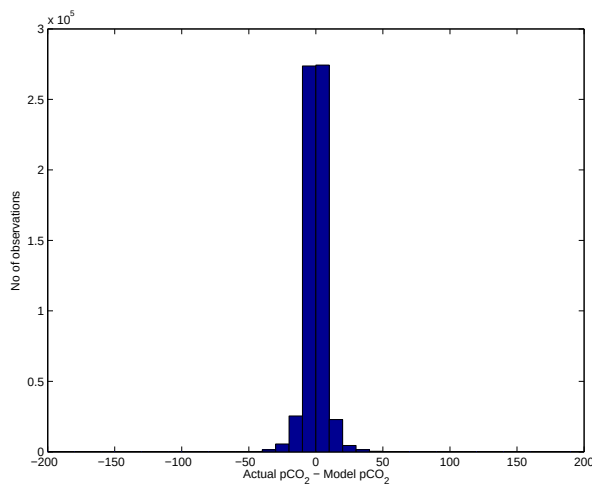


Figure 5.40: RBF interpolation results. Histogram of errors for all 100 runs and 200 D-optimal sampled points with the 0-2 scaled data set.

Optimal scaling of variables

The variables in the original data set all have a different range of values, as shown in Table 5.25. When using the locally supported radial basis function it has a radial influence of one. This means that a point is influenced by the points that is at a radius of one from this point. As before, it is suggested that by scaling the variables optimally, the effect of each variable on the pCO₂ as well as the variability in the values of each variable will be taken into account in order to reduce the error made in estimating the pCO₂ values from these variables.

To find this optimal scaling an optimization algorithm is used in order to find the scaling that minimizes the error made on the pCO₂ estimation. As initial guess for the optimization algorithm, the scaling is such that all the variables are scaled between zero and two. The variables are then specified in Equation (5.37) with the addition of the latitude:

$$\text{Lat}_j^* = a_4 * \left[\frac{\text{Lat}_j - \text{Lat}_{\min}}{\text{Lat}_{\max} - \text{Lat}_{\min}} \right]. \quad (5.38)$$

The values a_1 to a_4 are then determined as described in the previous section. Representative values for the optimal scaling are given in Table 5.25. It is clear from this scaling in Table 5.24 and Table 5.25 that the latitude requires local support. Temperature also requires local support, but considerably less than latitude. Mixed layer depth and chlorophyll-a concentration indicate global support on the domain. This shows that for especially the latitude, the regional differences in the relationship between the variables are taken into account with the local support radial basis functions. In this way, with the optimal scaling, the regional variations in the relationship between the various oceanic regions are taken into account in the pCO₂ estimations with the RBF interpolation function.

The optimal scaling on the variables is tested by sampling k points randomly as well as with D-optimal sampling. When this optimal scaling is used for the variables, the error made by the estimation of pCO₂ is shown in Figures L.1 to L.12 in Appendix L and the results for 200 sampled points are shown in Figures 5.41 to 5.46. These results are summarized in Table 5.26.

From the results it can be seen that the D-optimal sampling produces interpolations functions that estimate the pCO₂ with a smaller error than the interpolation function produced by the random sampling. As the number of points sampled are increased, the error on the estimation is decreased. The RMS error is reduced from 6.761 μatm without the optimal scaling to 4.065 μatm when the optimal scaling is applied to the variables for 200 D-optimally sampled points for the RBF interpolation with latitude included as variable. The optimal scaling takes into account the variability of the variables

Table 5.24: Optimal scaling for consecutive iterations of the optimization.

Iteration	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$	$\mathbf{a_4}$
1	2	2	2	2
2	2.128	1.933	1.811	7.931
3	6.235	-0.485	1.943	16.048
4	4.861	-0.705	1.313	19.294
5	6.612	-0.745	1.544	23.69
6	0.733	-0.496	1.617	24.357
7	2.735	0.135	1.912	24.39

Table 5.25: Range of values for the variables.

Scaling	$\mathbf{T_{min}}$	$\mathbf{T_{max}}$	$\mathbf{MLD_{min}}$	$\mathbf{MLD_{max}}$	$\mathbf{Chl_{min}}$	$\mathbf{Chl_{max}}$	$\mathbf{Lat_{min}}$	$\mathbf{Lat_{max}}$
Original Data	0.2791	21.3027	13.1542	127.9258	0.1200	5.1350	37.0004	69.5998
Scaled Data	0	3	0	0.5	0	1.8	0	24

across the domain as well as the importance of each variable in the estimation of the pCO_2 . By applying this optimal scaling to the variables, the results for the RBF interpolation is improved. The RMS error of the 200 D-optimal sampled points with the optimal scaling on the variables with the RBF interpolation is improved from $9.012 \mu\text{atm}$ with latitude excluded as variable to $4.065 \mu\text{atm}$ with latitude included as variable i.e. an improvement of 45% when latitude is included as variable. From Figure 5.44 it can be seen that the 200 D-optimal sampled points with the latitude included as variable and the optimal scaling fits practically all the spikes in the data with a small 95% confidence interval. Looking at the results for the 200 D-optimally sampled points with the optimal scaling, it can be seen that the RMS error of $4.065 \mu\text{atm}$ as well as the maximum over estimation of $39.32 \mu\text{atm}$ and the maximum under estimation of $-36.34 \mu\text{atm}$ is smaller than the required error of $60 \mu\text{atm}$.

Table 5.26: Results of RBF interpolation for variables T, MLD, Chl and Lat with optimally scaled variables.

No of points n_b	Random Sampling				D-optimal Sampling			
	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ	Mean RMSE	Max Over Es- timation	Max Un- der Esti- mation	σ
50	17.33	322.1	-255.1	18.09	9.560	103.0	-56.68	9.585
100	13.12	1043.4	-267.76	14.23	6.394	86.70	-53.57	6.423
200	9.849	493.8	-477.5	10.56	4.065	39.32	-36.34	4.070

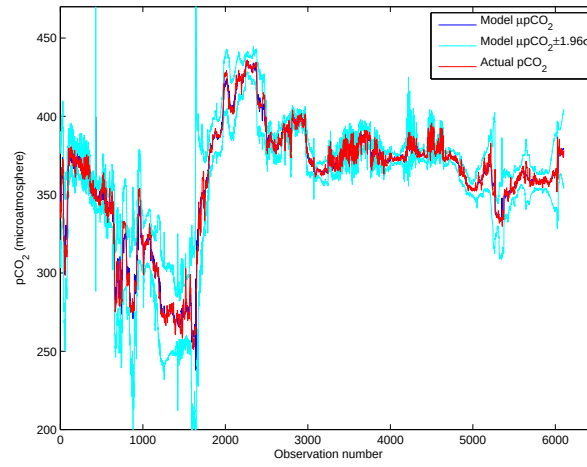


Figure 5.41: RBF interpolation results. 200 Random points sampled with the optimally scaled variables.

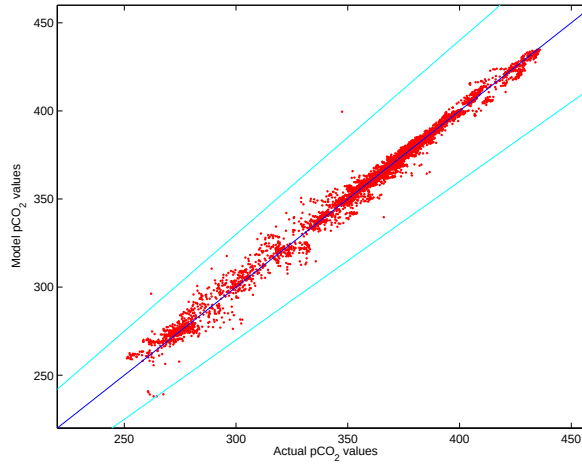


Figure 5.42: RBF interpolation results with 200 random points sampled and optimally scaled variables. The degree to which the modelled points fall in the 10 % error margin is shown. The mean $p\text{CO}_2$ estimation for all 100 runs is shown here.

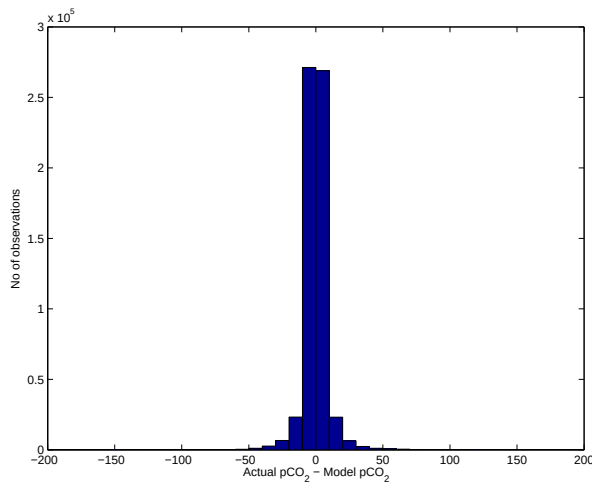


Figure 5.43: RBF interpolation results. Histogram of errors for all 100 runs and 200 Randomly sampled points with optimally scaled variables

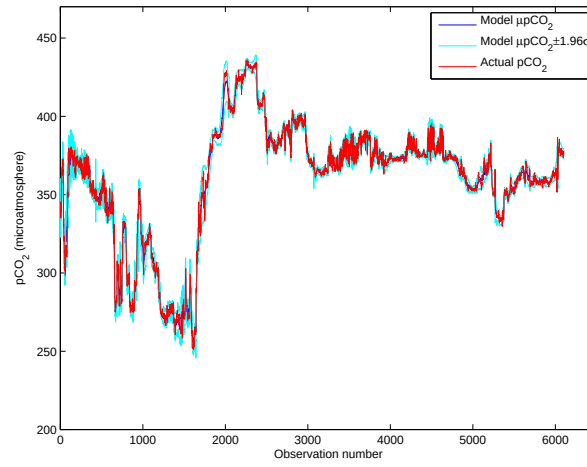


Figure 5.44: RBF interpolation results. 200 D-optimal sampled points with the optimally scaled variables.

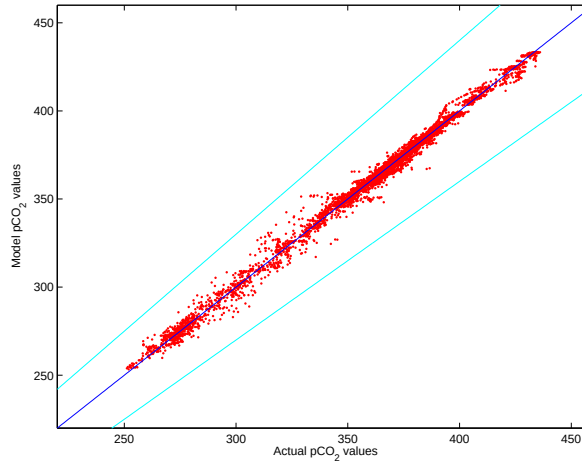


Figure 5.45: RBF interpolation results with 200 D-optimal sampled points and optimally scaled variables. The degree to which the modelled points fall in the 10 % error margin is shown. The mean pCO_2 estimation for all 100 runs is shown here.

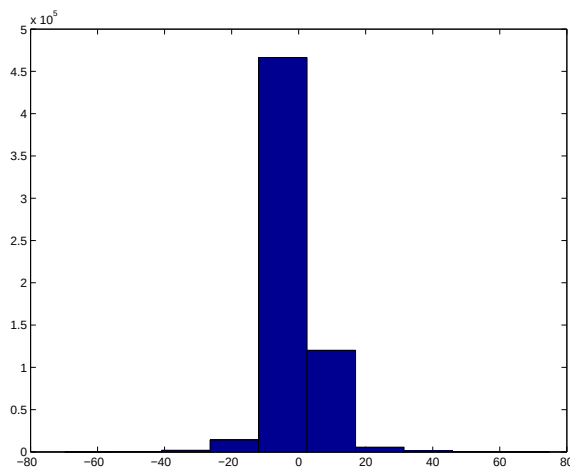


Figure 5.46: RBF interpolation results. Histogram of errors for all 100 runs and 200 Randomly sampled points with optimally scaled variables

5.5 Conclusion

In this chapter various methods were used in order to find the relationship between the ocean variables that yield the smallest RMS error on the estimation of the $p\text{CO}_2$ in the ocean. The least squares curve fitting with T, MLD and Chl as variables is one method that is used to establish the empirical relationship between the mentioned variables and the $p\text{CO}_2$. In each case a number of the points in the data set is used to create the equation, and the equation is tested on the remainder of the data points. The sampling of these points is done by random sampling and D-optimal sampling respectively and the error of the estimation in each case is compared. Latitude is then also added as a variable such that the variables T, MLD, Chl and Lat are fitted to the data in order to determine the empirical relationship between the $p\text{CO}_2$ and these variables. The same methods are followed to determine linear, quadratic, cubic and fourth order curve fits, and the results are presented in this chapter.

The fourth order equation in each case is used and the terms are removed one-by-one and the error of the estimation of the $p\text{CO}_2$ is calculated in each case. This is done in order to determine which terms in the fourth order equation makes the biggest contribution to the accurate estimation of the $p\text{CO}_2$.

Another method of investigating the relationship between the ocean variables to estimate the $p\text{CO}_2$ is to use radial basis function interpolation. Once again the sampling of the points that are used to create the equation is done by random sampling and D-optimal sampling respectively. Compact support radial basis functions are used in this case, since a large data set is used and several variables are considered. The variables are optimally scaled so that the radial basis functions take into account the variability of each variable separately. The RBF interpolation is then applied to these scaled variables T, MLD and Chl. The latitude is again included as a variable in order to include the effect of the regions in the ocean in the equation and the same methods are repeated with the variables T, MLD, Chl and Lat. The results of the RBF interpolation are also presented in this chapter.

These methods and procedures were used in order to find the optimal estimation equation for the CO_2 in the ocean. The aim is to find an equation that is simple, generalizable and yields a small RMS error on the estimation. From the above investigations several different equations for the estimation of the $p\text{CO}_2$ in the ocean is found. Furthermore, D-optimal sampling yields smaller RMS errors than random sampling. A discussion and overview of the results are given in the following chapter.