

Non-parametric regression modelling of *in situ* fCO₂ in the Southern Ocean

by

Wesley Byron Pretorius

*Thesis presented in partial fulfilment of the requirements for
the degree of Master of Commerce in the Faculty of
Economics and Business Management at Stellenbosch*



Supervisors:

Prof. Paul J. Mostert
Statistics and Actuarial Sciences
University of Stellenbosch

Dr. Sonali Das
Built Environment
CSIR

December 2012

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date:

Copyright © 2012 Stellenbosch University
All rights reserved.

Abstract

The Southern Ocean is a complex system, where the relationship between CO₂ concentrations and its drivers varies intra- and inter-annually. Due to the lack of readily available *in situ* data in the Southern Ocean, a model approach was required which could predict the CO₂ concentration proxy variable, fCO₂. This must be done using predictor variables available via remote measurements to ensure the usefulness of the model in the future. These predictor variables were sea surface temperature, log transformed chlorophyll-a concentration, mixed layer depth and at a later stage altimetry. Initial exploratory analysis indicated that a non-parametric approach to the model should be taken. A parametric multiple linear regression model was developed to use as a comparison to previous studies in the North Atlantic Ocean as well as to compare with the results of the non-parametric approach. A non-parametric kernel regression model was then used to predict fCO₂ and finally a combination of the parametric and non-parametric regression models was developed, referred to as the mixed regression model. The results indicated, as expected from exploratory analyses, that the non-parametric approach produced more accurate estimates based on an independent test data set. These more accurate estimates, however, were coupled with “zero” estimates, caused by the curse of dimensionality. It was also found that the inclusion of salinity (not available remotely) improved the model and therefore altimetry was chosen to attempt to capture this effect in the model. The mixed model displayed reduced errors as well as removing the “zero” estimates and hence reducing the variance of the error rates. The results indicated that the mixed model is the best approach to use to predict fCO₂ in the Southern Ocean and that altimetry’s inclusion did improve the prediction accuracy.

Opsomming

Die Suidelike Oseaan is 'n komplekse sisteem waar die verhouding tussen CO₂ konsentrasies en die drywers daarvoor intra- en interjaarlik varieer. 'n Tekort aan maklik verkrygbare in situ data van die Suidelike Oseaan het daartoe gelei dat 'n model benadering nodig was wat die CO₂ konsentrasie plaasvervanger-veranderlike, fCO₂, kon voorspel. Dié moet gedoen word deur om gebruik te maak van voorspellende veranderlikes, beskikbaar deur middel van afgeleë metings, om die bruikbaarheid van die model in die toekoms te verseker. Hierdie voorspellende veranderlikes het ingesluit see-oppervlaktetemperatuur, log getransformeerde chlorofil-a konsentrasie, gemengde laag diepte en op 'n latere stadium, hoogtemeting. 'n Aanvanklike, ondersoekende analise het aangedui dat 'n nie-parametriese benadering tot die data geneem moet word. 'n Parametriese meerfoudige lineêre regressie model is ontwikkel om met die vorige studies in die Noord-Atlantiese Oseaan asook met die resultate van die nie-parametriese benadering te vergelyk. 'n Nie-parametriese kern regressie model is toe ingespan om die fCO₂ te voorspel en uiteindelik is 'n kombinasie van die parametriese en nie-parametriese regressie modelle ontwikkel vir dieselfde doel, wat na verwys word as die gemengde regressie model. Die resultate het aangetoon, soos verwag uit die ondersoekende analise, dat die nie-parametriese benadering meer akkurate beramings lewer, gebaseer op 'n onafhanklike toets datastel. Dié meer akkurate beramings het egter met "nul"beramings gepaartgegaan wat veroorsaak word deur die vloek van dimensionaliteit. Daar is ook gevind dat die insluiting van soutgehalte (nie beskikbaar oor via sateliet nie) die model verbeter en juis daarom is hoogtemeting gekies om te poog om hierdie effek in die model vas te vang. Die gemengde model het kleiner fout getoon asook die "nul"beramings verwyder en sodoende die variasie van die foutkoerse verminder. Die resultate het dus aangetoon dat die gemengde model die beste benadering is om te gebruik om die fCO₂ in die Suidelike Oseaan te beraam en dat die insluiting van altimetry die akkuraatheid van hierdie beraming verbeter.

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations ...

To the almighty God for all His guidance and grace He has towards me in my imperfection.

To Prof. Paul Mostert and Dr. Sonali Das, my supervisors, I give all my thanks for your guidance and instruction when I had no clue what to do next.

To Stellenbosch University and the department of Statistics and Actuarial Sciences for allowing me the opportunity to complete my masters here.

To the CSIR for financial support.

To the SOCCO group for all their input and support in my thesis research.

To Dr. Pedro Monteiro, Dr. Nicolas Faucherau, Sebastian Swart and Sandy Thomalla for all their inputs and data collection.

To Marizelle van der Walt for all our discussions and brainstorming regarding this study.

To my loving wife Chantelle for all her patience and love and for believing in me even when I didn't.

To my mother and father for all their support, financially, emotionally and spiritually over the last 24 years and for the many years still to come.

To my brother, Warren, and sister, Suria, as well as their families for all their advice and support in my work.

To the le Roux and Carstens families, skoonma, skoonpa, Adelma en Thinus, for all your love and support in all areas of my life

And finally to all my friends and family, my sincerest gratitude for all your kind words, helping gestures and motivational moments. They will never be forgotten.

Dedications

*This thesis is dedicated to:
Chantelle Pretorius (My loving wife)
Leo (Luigi) O'Connor (10/03/1977 – 17/07/2012) (“At the going down of
the sun and in the morning, we will remember them”)*

Contents

Declaration	i
Abstract	ii
Opsomming	iii
Acknowledgements	iv
Dedications	v
Contents	vi
List of Figures	xi
List of Tables	xiv
List of Abbreviations and Symbols	xvi
1 Introduction	1
1.1 Background	1
1.1.1 The Global Carbon Cycle	1
1.1.2 Carbon Sinks and the Southern Ocean	2
1.2 Focus of the Study	3
1.2.1 Research Objectives	4
1.2.2 Potential Obstacles of the Study	5
1.2.3 Contribution of the Study	5

<i>CONTENTS</i>	vii
1.3 Outline of Thesis	6
2 Overview of Anthropogenic CO₂	7
2.1 Introduction	7
2.2 Concentration of CO ₂ and its distribution	8
2.3 Main Factors Influencing CO ₂ Solubility in the SO	14
2.4 SANAE49 data set	15
2.4.1 Introduction	15
2.4.2 Description of the data set	15
2.4.3 Data cleaning	17
2.4.3.1 Locating spikes in the data	17
2.5 SANAE49L6-MLD data set	19
2.5.1 Introduction	19
2.5.2 Reducing latitude values	20
2.5.3 Description and exploratory analysis	22
2.5.4 Graphical approach to exploratory analysis	26
2.6 Summary	29
3 Parametric regression model for CO₂ concentrations	31
3.1 Introduction	31
3.2 Modelling ocean CO ₂ with MLR	32
3.3 Theoretical background of linear regression	34
3.3.1 Simple linear regression	34
3.3.2 Multiple linear regression	35
3.3.3 Least Squares Minimisation	36
3.3.4 Matrix notation of the least squares minimisation	37
3.4 Linear models to predict fCO ₂	38
3.5 Multiple linear regression results to predict fCO ₂	41
3.5.1 Optimising the regression model	41
3.5.2 Assessing the regression model	42

3.5.2.1	Training-Test Data Splits	47
3.5.2.2	Standardised regression model to predict fCO_2 .	47
3.5.2.3	Simulating the regression error	48
3.6	Discussion of the linear regression results	50
3.6.1	Model parameter interpretation	51
3.6.2	Discussion of error statistics	52
3.6.2.1	Training-Test Data splits	53
3.6.2.2	Standardised regression models	54
3.6.2.3	Simulating the regression error	55
3.7	Summary	55
4	Non-parametric Kernel Regression	57
4.1	Introduction	57
4.2	Review on non-parametric research of CO_2 data	58
4.3	Using non-parametric kernel methods for predicting fCO_2	59
4.3.1	Theoretical overview of non-parametric kernel regression	61
4.3.2	Specifying the kernel and bandwidth optimisation	64
4.4	Non-parametric results to predict fCO_2	68
4.4.1	Optimising the non-parametric regression model	69
4.4.2	Assessing the non-parametric regression model	70
4.4.2.1	Training-Test Data Splits	71
4.4.2.2	Standardised non-parametric regression models	75
4.4.2.3	Simulating the non-parametric regression error	75
4.5	Discussion of the non-parametric regression results	80
4.5.1	Model bandwidth interpretation	80
4.5.2	Model error rates	81
4.5.2.1	Training-Test Data splits	83
4.5.2.2	Standardising non-parametric regression models	84
4.5.2.3	Simulating non-parametric regression error	85
4.6	Summary	86

<i>CONTENTS</i>	ix
5 Sea surface topography and the mixed regression model	88
5.1 Introduction	88
5.2 Sea surface topography	89
5.2.1 Background on sea surface topography	89
5.2.2 Altimetry data collection	92
5.3 Regression models to include altimetry	93
5.3.1 Developing the regression model	94
5.3.2 Mixture of parametric and non-parametric regression models	95
5.4 Regression results to predict $f\text{CO}_2$	97
5.4.1 Estimating the pure parametric regression model and non-parametric regression model including altimetry	98
5.4.2 Assessing the parametric and non-parametric regression models with altimetry	98
5.4.3 Estimating the mixed regression model	98
5.4.4 Assessing the mixed regression model	100
5.4.4.1 Training-test data splits	100
5.4.4.2 Simulating the mixed regression model error	104
5.5 Discussion	106
5.5.1 NPKR and MLR models including altimetry	106
5.5.1.1 Estimating the models	106
5.5.1.2 Assessing the models	107
5.5.2 Mixed Models	108
5.5.2.1 Subset Division	109
5.5.2.2 Model Simulation	112
5.6 Conclusion	114
6 Summary, conclusions and future research	115
6.1 Summary	115
6.1.1 Exploratory Analysis	116

*CONTENTS***x**

6.1.2	Multiple Linear Regression	116
6.1.3	Non-parametric kernel regression	117
6.1.4	Including altimetry into the regression model	118
6.1.5	Mixed regression model	119
6.2	Conclusion	120
6.3	Future research	121
6.3.1	Removal of spatial dependency	121
6.3.2	Small area modelling	121
6.3.3	Expanding the model to remote sensing data	122

Appendix

R Code		123
Data cleaning		123
Exploratory Analysis		124
Multiple linear regression		127
Models M1 to M10		127
MLR error simulation		132
Non-parametric kernel regression		134
Models M1 - M10		134
NPKR error simulation		140
Mixed regression model		141
MLR and NPKR model M11 and Mixed models M1, M3 and M11		141
Mixed regression model subset division		144
Mixed regression model error simulation		152

List of References**154**

List of Figures

2.1	Mean annual net air-sea flux for CO ₂ for 1995.	11
2.2	Location of LDEO V2009 master database of sea surface pCO ₂ observations (Takahashi <i>et al.</i> , 2009a)	13
2.3	Traveling path of the SANAE49 ship	16
2.4	Plots of variables from SANAE49L6-EQU	19
2.5	Euclidean distance weighted interpolation of MLD values	21
2.6	Despiked variable plots from SANAE49L6-final	23
2.7	Histogram of fCO ₂ , MLD, pH and salinity	27
2.8	Histogram of Intake Temperature, Chlorophyll-a Concentration, Oxygen (ppm) and Oxygen (Saturation)	28
3.1	Multiple linear regression observed and predicted fCO ₂ for model M1 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	44
3.2	Multiple linear regression observed and predicted fCO ₂ for model M2 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	45
3.3	Multiple linear regression observed and predicted fCO ₂ for model M3 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	46
3.4	Histogram of 100 MLR model MSEs	49
3.5	Histogram of 100 MLR model MAEs	49

3.6	Histogram of 100 MLR model RMSEs	50
4.1	Non-parametric kernel regression observed and predicted fCO ₂ for model M1 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	72
4.2	Non-parametric kernel regression observed and predicted fCO ₂ for model M2 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	73
4.3	Non-parametric kernel regression observed and predicted fCO ₂ for model M3 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	74
4.4	Non-parametric kernel regression observed and predicted fCO ₂ for model M8 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	76
4.5	Non-parametric kernel regression observed and predicted fCO ₂ for model M9 (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	77
4.6	Histogram of 100 non-parametric kernel regression model MSEs	78
4.7	Histogram of 100 non-parametric kernel regression model MAEs	79
4.8	Histogram of 100 non-parametric kernel regression model RMSEs	79
5.1	1992-2002 Mean dynamic ocean topography on a 0.5° grid Maximenko and Niiler (2011)	91
5.2	Line plot of altimetry versus latitude	94
5.3	Prediction plots of fCO ₂ versus latitude for the mixed model (left) and NPKR model (right) for the 30% - 70% subset division (blue dots represent observed fCO ₂ while the red line represents predicted fCO ₂)	102

5.4	Prediction plots of $f\text{CO}_2$ versus latitude for the mixed model (left) and NPKR model (right) for the 20% - 80% subset division (blue dots represent observed $f\text{CO}_2$ while the red line represents predicted $f\text{CO}_2$)	103
5.5	Histogram of mean square errors for 100 different subset divisions using the mixed model M11	104
5.6	Histogram of mean absolute errors for 100 different subset divisions using the mixed model M11	105
5.7	Histogram of root mean square errors for 100 different subset divisions using the mixed model M11	105

List of Tables

2.1	Variables of SANAE49L6-EQU	17
2.2	Explanation of new variables in SANAE49L6-final	22
2.3	Descriptive statistics of SANAE49L6-final	24
2.4	Shape and range descriptive statistics	25
3.1	MLR Models Investigated	39
3.2	MLR model parameter estimates	42
3.3	Multiple linear regression model error rates	43
3.4	Multiple Linear Regression Subset Division Error Rates	47
3.5	Standardised model error rates for Multiple Linear Regression Models	48
3.6	MLR error rate statistics for 100 subset divisions	50
4.1	Non-parametric kernel regression bandwidth estimates	69
4.2	Non-parametric kernel regression model error rates	70
4.3	Non-parametric kernel regression subset division error rates	71
4.4	Standardised model error rates for non-parametric kernel regression models	75
4.5	Non-parametric kernel regression error rate statistics for 100 subset divisions	80
5.1	Descriptive statistics of altimetry data	92
5.2	Shape and range statistics of altimetry data	93
5.3	Optimised MLR parameter estimates	98
5.4	Optimised NPKR optimal bandwidth estimates	98

5.5	Error rates for model M11 including altimetry using MLR and NPKR approaches	99
5.6	Multiple linear regression parameter estimates for differing subset divisions	99
5.7	Non-parametric kernel regression bandwidth estimates for differing subset divisions	99
5.8	Error rates for mixed models M1, M3 and M11	100
5.9	Error rates for mixed models developed and assessed on varying subset sizes	101
5.10	Descriptive statistics of the error rates of 100 repetitions of the mixed model M11	106

List of Abbreviations and Symbols

Constants

$\pi =$ 3.141 592 654

$e =$ 2.718 281 828

Abbreviations

CO₂ Carbon dioxide

SO Southern ocean

VOS Volunteer observing ships

MLR Multiple linear regression

NPKR Nonparametric kernel regression

SOCAT Southern Ocean CO₂ atlas

pCO₂ Partial pressure of carbon dioxide

fCO₂ Fugacity of carbon dioxide

xCO₂ Mole fraction of carbon dioxide in the space of air above the sea water

SANAE49L6 South African national Antarctic Expedition 49 leg 6

μatm Micro atmospheres

ppm Parts per million

SST Sea-surface temperature

MLD Mixed layer depth

$\mu\text{g/l}$	Micrograms per litre
<i>SOCCO</i>	Southern ocean carbon and climate observatory
<i>COV</i>	Coefficient of variation
<i>DIC</i>	Dissolved inorganic carbon
<i>SSS</i>	Sea-surface salinity
<i>SLR</i>	Simple linear regression
<i>RSS</i>	Residual sum of squares
<i>MSE</i>	Mean square error
<i>MAE</i>	Mean absolute error
<i>RMSE</i>	Root mean square error
<i>NN</i>	Neural network
<i>SOM</i>	Self-organising map
<i>CV</i>	Cross-validation
<i>SSH</i>	Sea-surface height
<i>SSA</i>	Sea-surface altimetry
<i>NCEP</i>	National centers for environmental prediction
<i>GRACE</i>	Gravity recovery and climate experiment
<i>CSIR</i>	Council for industrial and scientific research

Chapter 1

Introduction

1.1 Background

This project focuses on applying statistical techniques, in particular non-parametric kernel regression modelling, in order to provide an understanding of the relationships between the physical and bio-geochemical properties and the concentration of carbon dioxide (CO_2) (described by the fugacity of CO_2) in the Southern Ocean (SO). These relationships are used to form an understanding of the distribution of oceanic sinks and sources of CO_2 in the SO in order to predict carbon concentrations in areas of the ocean which have not yet been observed *in situ*.

1.1.1 The Global Carbon Cycle

CO_2 is widely attributed as being the leading factor in the increasing, negative effects of the global climate change phenomenon affecting all parts of the world. Focus has, therefore, increasingly been placed on reaching an agreement to not only stabilise, but actively reduce CO_2 emissions in order to curb the impact on the climate. These agreements and strategies make the assumption that the natural global carbon cycle's fluxes (which makes a much larger contribution to the global cycle) will remain in balance (Monteiro, 2010). This is, however,

not a certainty and the assumption is by no means accurate. Naturally CO₂ rich areas (such as the CO₂ sinks in the SO) are very complex systems relying on a number of climatic conditions and are therefore very sensitive to changes in any of these factors. These systems are not well understood in the SO due to very little complete data being available and therefore an effort is made in this study to shed some light on the problem of understanding the system as well as providing a model which can predict CO₂ concentrations in areas of the ocean which are not yet available for sampling.

1.1.2 Carbon Sinks and the Southern Ocean

Humankind has been responsible for an increase of more than 30% in the non-natural CO₂ emissions (known as anthropogenic CO₂) since the Industrial Revolution. This has caused emissions of CO₂ to reach higher levels than ever before in recorded history and has been attributed to humankind's role in the burning of carbon rich fossil fuels such as coal, natural gas and oil (Sarmiento and Gruber, 2002). These sources provide us with energy to produce electricity, heat and also to power forms of transportation and industrial production. The removal of forests and harvesting of wood by human beings have added to the already increasing CO₂ levels in the atmosphere, however, the rate at which atmospheric CO₂ has been increasing is less than 50% of the rate expected if all anthropogenic CO₂ produced remained in the atmosphere. This reduced rate of the retention of CO₂ in the atmosphere is due to a significant uptake of CO₂ by plants, soils and water sources such as the ocean. In essence, these natural elements act as terrestrial and oceanic "sinks", absorbing CO₂ and storing it for many years. The threat of global climate change may, in fact, worsen from initial ideas if climatic conditions caused by humans reduce the absorption of CO₂ by the terrestrial biosphere and the ocean (Sarmiento and Gruber, 2002; Monterey Bay Aquarium Research Institute, 2005). The CO₂ behaviour in the SO is of particular interest to this study. The size and strength of this

specific oceanic sink has been heavily disputed since it is highly variable and is influenced, to a large extent, by the climate. This results in changes being observed with regards to CO₂ absorption at different times of the year. A major reason for the debate surrounding the size and strength of the SO carbon sink is due to the fact that data regarding the SO is especially sparse in comparison to the total surface area being discussed (Le Quéré *et al.*, 2007). A study of air-sea CO₂ fluxes by Takahashi *et al.* (2002) suggested a significantly large CO₂ sink exists in the SO, contributing up to 20% of the total annual oceanic CO₂ uptake flux, while representing an area of ocean covering only approximately 10% of the total area of the global ocean. The SO is also of direct importance since it is the only place where a direct exchange of CO₂ between CO₂ rich deep waters and the atmosphere takes place (Monteiro, 2010). Carbon fluxes can be described as a process taking place between two carbon reservoirs, in this case the ocean and the atmosphere, where a transfer occurs between the systems on a connecting surface i.e. the surface of the water (Bye, 1996).

1.2 Focus of the Study

The main objective of this study is two-pronged. The first focus area addresses the topic of the seasonal cycle of the oceanic CO₂ fluxes and how accurate current knowledge of this cycle is as well as testing how sensitive the SO carbon-climate system is to changes in the annular wind and fresh water fluxes. However this is not the main focus of this thesis. The second focus area is the main objective of this particular thesis, which is to develop a model that can be used to predict CO₂ concentrations in the SO based on *in situ* observations. This thesis focuses specifically on identifying an accurate and reliable method for predicting fCO₂ which can then, in future studies, be expanded to areas of the SO where *in situ* measurements are unavailable.

1.2.1 Research Objectives

To determine an accurate and reliable, statistical approach to estimate the concentrations of CO_2 in the SO in a way that is understandable and explainable to persons not involved in the model building process.

The objectives of this thesis are divided into short, medium and long term goals. The short term goals include the initial analyzing of the data in an attempt to understand the distribution and profile of the variables and the response which, in this study, is the concentration of CO_2 in the ocean. This concentration can be represented by the partial pressure of CO_2 ($p\text{CO}_2$), fugacity of CO_2 ($f\text{CO}_2$) or the mole fraction of CO_2 in the space of air above the sea water ($x\text{CO}_2$). This process also serves as a method of determining any irregularities in the data and thereby cleaning the data set. In the medium term, the goal is to determine any interesting and notable relationships between the variables in the data set, in order to briefly describe how they relate to one another and the response in order to develop a model which can replicate these relationships. The long term objective is to reduce the current uncertainty in the predicted CO_2 concentrations from 50% to around 10% of the average CO_2 concentration using numerical methods to produce a model which is then used to predict the CO_2 ocean-atmosphere fluxes in the SO (Monteiro, 2010). In this thesis, the objective is approached by applying non-parametric kernel regression (NPKR) models to the data obtained from the SO. These models could then be used to predict the CO_2 concentration values for measurements of the (possible) predictors obtained through satellites in areas where *in situ* measurements were not available, since these are obtained only along lines run by the voluntary observing ships (VOS) used in the measurement process. The generalisation of these models to remote sensing data obtained from satellites, however, is not part of the objectives for this thesis.

1.2.2 Potential Obstacles of the Study

Measurements with regards to CO₂ concentrations are sparse in terms of the total surface area of the SO, since data are collected only along lines traveled by the ships. This, therefore, provides little (if not no) information regarding the measurements in the rest of the SO region. Measurements of CO₂ concentrations in the SO are also very difficult to obtain in the winter months due to oceanic shipping paths to Antarctica being frozen and treacherous climatic conditions which make traveling by ship in the area practically impossible (Böning *et al.*, 2008; Le Quéré *et al.*, 2009). This results in the measurements also being seasonally biased (as well as being sparse) and therefore empirical relationships between CO₂ concentrations and other measurable oceanic variables, for which less sparse and more seasonally regular measurements are available (especially due to remote sensing), need to be investigated.

1.2.3 Contribution of the Study

This analysis will develop a model on *in situ* data and assess the optimised model on an independent test data set not used in the model development. Once the model has been identified to predict unseen, *in situ* data, the entire data set can then be used to develop an optimised model which can be used to produce a fCO₂ flux map for the SO. This will contribute towards an understanding of the distribution of CO₂ sinks and sources in the SO as well as the magnitude of the overall sink/source present. The R code used for this analysis was also written to be able to adapt to any similar, future data set. Although the model estimation programs were obtained from existing code, the data cleaning, model development and model assessment functions, as well as the mixed model programs were all written and made available for future research in this area.

1.3 Outline of Thesis

Chapter 2 provides the background to the problem of CO₂ concentrations in ocean waters as well as particularly in the SO. The data set South African national Antarctic Expedition 49 leg 6 (SANAE49L6) and method of data cleaning used in this thesis is also described in this chapter along with an exploratory analysis of the data. Chapter 3 introduces the first (parametric) approach to develop a model which can predict fCO₂. Multiple linear regression (MLR) is first discussed in detail (along with a similar previous study where MLR was used) and then applied to the SANAE49L6 data set. The results are discussed along with shortfalls in the approach. Chapter 4 presents the non-parametric kernel regression (NPKR) method as an alternative to the MLR approach. The chapter begins by introducing other non-parametric techniques used in estimating CO₂ concentrations and discussing why an alternative is needed. The NPKR method is then discussed in detail and applied to the SANAE49L6 data. The results obtained are then discussed and compared to those from the MLR approach. Chapter 5 introduces the sea surface topography (altimetry) as an independent variable for both the MLR and NPKR approaches. A combination of these two regression models is then investigated while including altimetry in the regression functions. The results of these mixed models are compared to the individual MLR and NPKR approaches and final conclusions as well as future research opportunities are discussed in Chapter 6.

Chapter 2

Overview of Anthropogenic CO₂

2.1 Introduction

The interaction, with regards to anthropogenic carbon dioxide, between the ocean and the atmosphere (known as the carbon flux) has a large impact on the amount of CO₂ measured in the atmosphere. *In situ* measurements made from ships traveling the SO suggest a large sink for atmospheric CO₂ exists (Rangama, 2005). The following section explains this in more detail.

The increase in CO₂ levels observed in the atmosphere has caught the attention of the research world due to its role in trapping radiation emitted from the surface of the earth. More than half of this increased trapping of radiation, since the beginning of the industrial age, by the earths atmosphere can be attributed to CO₂. The implications of this depends on many other factors, but general consensus is that it will lead to global warming. This implies a warming in overall temperature readings combined with the associated environmental changes such as an increased sea-level. These factors will not only have a negative impact on global terrestrial and marine ecosystems, but will also impact on the global socio-economic condition of human beings (Sarmiento and Gruber, 2002; Takahashi *et al.*, 2009b).

CO₂ is, however, nonreactive in the earths atmosphere and for this rea-

son it remains (resides) there for a long period of time. Initial impressions of the CO₂ levels in the atmosphere, based on anthropogenic emissions, were therefore very worrying, but were fortunately found to be unsubstantiated. The reason for this is due to the terrestrial and oceanic carbon sinks. Researchers have suggested that the carbon not measured in the atmosphere is approximately equally divided between these two natural sinks (Sarmiento and Gruber, 2002). It is roughly suggested that of the approximate 7 billion tons of anthropogenic carbon produced by humans every year, only half remains in the atmosphere to act as a reflector for radiation waves. ± 1.5 billion tons of this human produced carbon is absorbed by the terrestrial biosphere, while a further ± 2 billion tons are dissolved into the ocean. The carbon which is dissolved into the ocean, although not directly contributing to global warming anymore, disrupts the ecological system in the ocean by creating a more acidic environment. This disruption may indirectly influence climate change if it affects the oceanic carbon cycle, specifically by reducing the absorption of CO₂ by the ocean (Monterey Bay Aquarium Research Institute, 2005). The debates continue with regards to the spacial location, distributions and mechanisms of these sinks. New research performed by Deng and Chen (2011) suggested that the CO₂ retention of the atmosphere is even less than previously suspected, indicating that only about 40% of the anthropogenic CO₂ produced is retained. It is imperative that the behaviours of these sinks be understood in order to control the impact of future anthropogenic emissions of CO₂.

2.2 Concentration of CO₂ and its distribution

The concentration of CO₂ present in the ocean cannot be measured directly from the oceanic waters. For this reason a proxy must be determined in order to obtain a quantitative measure of this concentration of greenhouse gas in the ocean waters. The proxy measurement used is fCO₂ and can be defined as the

concentration of dissolved CO₂ gas in the ocean measured directly from the ship. These values are then used to derive the pCO₂ which takes into account that CO₂ does not act as an ideal gas in the ocean system (Dickson and Goyet, 1994; Weiss, 1974). Due to the consistent and reliable use of fCO₂ as a measurement of CO₂ in the ocean by the Southern Ocean CO₂ atlas (SOCAT) database which collects all *in situ* measurements of fCO₂, pCO₂ and xCO₂ into a single, common format, this analysis uses fCO₂ as a response variable. This allows for future studies to expand to using the models developed to predict fCO₂ using remote sensing of the independent variables in order to compare to the *in situ* measurements in the SOCAT database. Lueker *et al.* (2000) indicated that the net air-sea flux of pCO₂ must be determined to indicate the net uptake of CO₂ by the ocean. This is done by determining the $\Delta p\text{CO}_2$ or in equivalent terms, the difference between the pCO₂ levels in the atmosphere, to that on the ocean surface. For clarity purposes, a definition of $\Delta p\text{CO}_2$ is given as $(p\text{CO}_2)_{\text{Water}} - (p\text{CO}_2)_{\text{Atmosphere}}$. Determining this value is done using dissolved inorganic carbon (DIC) and Total Alkalinity (TA), as well as using the first and second dissociation constants of carbonic acid (K1 and K2). The details of this relationship can be found in Lueker *et al.* (2000). The need to understand this imbalance between the pCO₂ levels in the atmosphere and the pCO₂ levels in the ocean is described by Takahashi *et al.* (2009b) in which the potential existing in the ocean surface for the transfer of CO₂ is described. The potential for a carbon sink exists when the $(p\text{CO}_2)_{\text{Atmosphere}}$ is larger than the $(p\text{CO}_2)_{\text{Water}}$ resulting in a negative $\Delta p\text{CO}_2$ value. In this case the excess atmospheric CO₂ is absorbed by the ocean thereby creating a carbon sink. The opposite, however, can also occur. When the $\Delta p\text{CO}_2$ is positive in which case the excess CO₂ in the ocean is released into the atmosphere resulting in a carbon source. The seasonal variation in the levels of pCO₂ (and fCO₂) measured in the ocean is generally much higher than the seasonal variation of the pCO₂ (and fCO₂) measured in the atmosphere. For this reason the

magnitude and the direction of the interaction and transfer of CO₂ between the ocean and the atmosphere depends mainly on the oceanic pCO₂ (and therefore also fCO₂) measurements (Takahashi *et al.*, 2009b, 2002). A definite need, therefore, exists to determine the distribution and strength of the sinks and sources in the SO (Bakker *et al.*, 1997).

An analysis performed by Takahashi *et al.* (2002) indicated that the area of the SO situated between approximately 40°S and 60°S of the equator seems to contain large anthropogenic CO₂ sinks. They suggest a mixing effect of the warm south bound waters and the nutrient rich sub-polar waters as a possible reason for the observed sink. An increase in atmospheric CO₂ over the past few centuries has been attributed to an increase in industries and technology producing large amounts of anthropogenic CO₂, since CO₂ is produced in the use of fossil fuels such as coal, oils and other naturally rich carbon sources. The ocean represents a large CO₂ sink, absorbing an estimated 33% of the anthropogenic CO₂ per annum (this figure has been debated among researchers of the oceanic carbon flux cycle). Understanding the carbon flux in the ocean (along with the carbon flux in the terrestrial biospheres) can allow for the prevention of dangerous climate changes as well as the prediction of expected climate changes based on historical data (Deng and Chen, 2011). Since oceans cover over two thirds of the planet and have a significant effect on the absorption of anthropogenic CO₂, the ocean can be perceived as playing an integral role in controlling our climate. However, measuring and detecting small changes in fCO₂ levels in the ocean represents a formidable challenge and, along with the large seasonal variations in fCO₂ in the ocean and its vast size, complicates the task of directly measuring oceanic CO₂ concentrations. A further complicating factor in the measuring of changes in the oceanic fCO₂ regards the large spatial variations of carbon dioxide in the ocean (Goyet, 1998). This is especially true in the SO, where spatial variability and uncertainty is present which will be seen later. Uncertainties in the measurements of pCO₂ and fCO₂ in the ocean

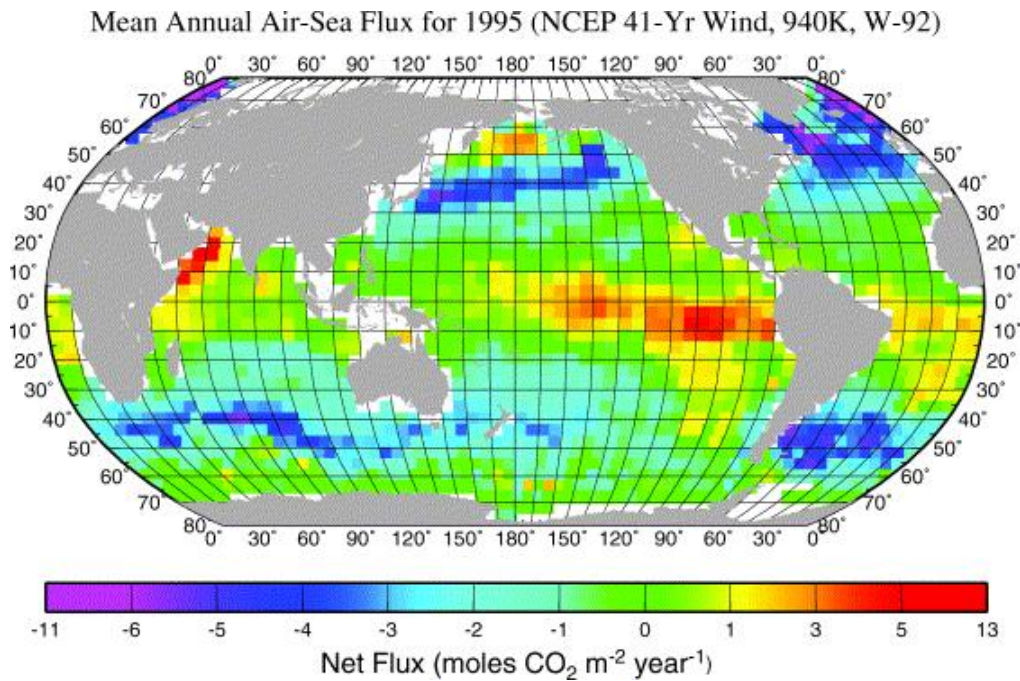


Figure 2.1: Mean annual net air-sea flux for CO₂ for 1995.

Mean annual net air-sea flux for CO₂ (mole CO₂ m⁻² year⁻¹) for 1995. The following information has been used; (a) climatological distribution of surface-water pCO₂ for the reference year 1995, (b) the NCEP/NCAR 41-year mean wind speeds, (c) the long-term wind-speed dependence of the sea-air CO₂ transfer velocity by Wanninkhof (1992) (d) the concentration of atmospheric CO₂ in dry air in 1995 (GLOBALVIEW-CO₂, 2000), and (e) the climatological barometric pressure and sea surface temperature (Atlas of Surface Marine Data, 1994; Takahashi *et al.*, 2002)

have been estimated, but can be reduced by introducing new measurements as they become available. This is particularly evident in Takahashi *et al.* (1997), (2002) and (2009b), in which adding to the data available helped not only obtain new estimates of the oceanic uptake of pCO₂ in different areas of the ocean, but also helped reduce uncertainties in these measurements .

Figure 2.1 is obtained from Takahashi *et al.* (2002). It depicts the estimated average annual air-sea flux of CO₂ (measured in moles of CO₂ per square meter per year) for the year 1995. Very low values, which are depicted by the blue or purple pixels, indicate the ocean acting as a sink for atmospheric CO₂. This was described earlier as areas of the ocean where atmospheric pCO₂ is higher than oceanic pCO₂ (or equivalently for fCO₂ values) resulting in a

dissolving of the CO₂ into the ocean waters. Areas coloured yellow or red indicate the ocean areas acting as a source of CO₂. What is clearly visible in the figure is the carbon sink evident south of South Africa, between 40°S and 60°S of the equator. After this, a neutral (neither a sink nor source) area is indicated further south towards Antarctica. An important observation is that even though the SO (defined by them as all areas of ocean below 50°S of the equator) only takes up about 10% of the earth's ocean area, it is responsible for approximately 20% of the earth's annual total oceanic CO₂ uptake. This places high importance on attempting to understand and control the fluxes of CO₂ in the SO (Takahashi *et al.*, 2002).

Takahashi *et al.* (2002) identify the importance of developing a model to understand and predict the CO₂ concentrations in the SO. Due to the observed increase in anthropogenic CO₂ emissions, it is imperative that such a model is developed and that the understanding of the relationship between the oceanic and atmospheric systems is improved. A large stumbling block in the SO is the lack of data available in comparison to other areas of the world, especially the northern hemisphere oceans. Ships taking measurements in the northern hemisphere cover almost the entire oceanic region, whilst in the southern hemisphere observations are restricted not only to certain areas, but to certain times of the year as well. This is due, not solely, to extreme maritime conditions in the SO especially in the winter months. Another factor that plays a large role in the under sampling of the SO are the very cold climates, which drops well below freezing, causing enormous areas of frozen water which prohibits the sailing of ships in certain areas of the ocean and creates treacherous conditions in other areas (Böning *et al.*, 2008; Le Quéré *et al.*, 2009). This also inhibits *in situ* measurements being made during the winter months and therefore most (if not all) data available from the SO is seasonally biased. The SO is also limited in terms of commercial ships which travel from which *in situ* observations could be made. This is in comparison

to the busy, maritime trade routes in the North Atlantic, which has therefore been extensively sampled. Figure 2.2 is taken from a data base published by Takahashi and Sutherland (2007) and later updated again in Takahashi *et al.* (2009a). The figure shows the traveling paths of ships in the global ocean from which data was collected regarding the sea surface fCO₂ levels (Takahashi *et al.*, 2009a).

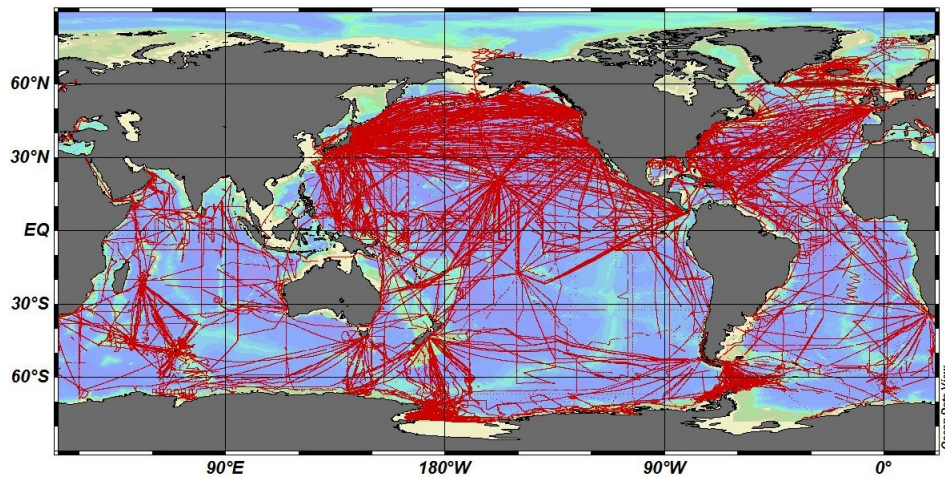


Figure 2.2: Location of LDEO V2009 master database of sea surface pCO₂ observations (Takahashi *et al.*, 2009a)

From Figure 2.2 it is evident that there is a lack of measurements of sea surface fCO₂ in the southern hemisphere. The northern hemisphere, in the figure, seems to be widely sampled with large areas covered in red indicating ship activity and fCO₂ measurements in these areas are high and frequent. The majority of the southern hemisphere oceans, however, are not sampled, leaving large areas where currently no *in situ* data is available. This creates large problem areas for modelling since being able to reliably predict the sea surface fCO₂ levels in those parts of the ocean which were not sampled becomes very difficult and results in large uncertainties in the predictions as well as having little data with which to assess the predictive ability of the models.

2.3 Main Factors Influencing CO₂ Solubility in the SO

The concentration of CO₂ absorbed (or released) in the exchange between the SO and the atmosphere is determined by many factors. These factors may vary not only spatially, but also on a time scale, such as intra-seasonally. This section discusses some of these factors.

Takahashi *et al.* (2002), introduced a broad spectrum of factors affecting the pCO₂ (and fCO₂) levels in what is referred to as the mixed-layer. This is the layer of water exchanging CO₂ directly with the atmosphere. They suggest that the pCO₂ levels in this layer are directly influenced by changes to the temperature, total concentration of CO₂ in the mixed-layer and the alkalinity of the ocean. These three variables are, in turn, influenced by other, more readily measurable variables. Water temperature is mainly affected by physical factors (such as solar-energy input and mixed-layer thickness) while the total CO₂ concentration and alkalinity of the ocean are determined mainly by biological processes (such as photosynthesis, respiration and calcification) and also by the upwelling of CO₂ that brings nutrient rich deep-waters to the surface which then directly exchanges CO₂ with the atmosphere (Takahashi *et al.*, 2002). These deep-waters can contain CO₂ absorbed by the ocean from many years before, and stored deep within the depths of the ocean (Sigman *et al.*, 2010; Takahashi and Chipman, 1982). A potential problem that faces humans with regards to this form of storage of carbon is that if the amount of CO₂ stored in deep waters by the ocean diminishes, or stops altogether, it could result in a much higher percentage of the anthropogenic CO₂ being released into the atmosphere.

The system for the solubility of CO₂ is a complex process due to a number of factors as discussed above, as well as their interactions with one another and their joint effect on CO₂ solubility. It is well documented that as the

temperature of the solution, in this case the sea surface temperature (SST), increases, it forces the CO₂ gas to become more soluble, while the “salting-out effect” refers to a reduction in the solubility of CO₂ gas in solutions due to the presence of salts (in this case the salt refers to the salinity of the SO) (Al-anezi *et al.*, 2008; Markham and Kobe, 1941; Yasunishi and Yoshida, 1979). This complex system, therefore, poses a problem in modelling the effect of the physical and bio-geochemical factors on the levels of fCO₂ in the ocean and the models proposed later in this thesis will be shown to closely capture this complex relationship in the *in situ* data used.

2.4 SANAE49 data set

2.4.1 Introduction

This section gives a detailed description of the data used in the subsequent analyses and also explains the methods involved in cleaning the data. In the latter part of the chapter, a preliminary descriptive data analysis, as well as exploratory plots of the variables, are provided.

2.4.2 Description of the data set

The data is obtained from the 2009-2010 data collection trip in the ocean south of South Africa, where measurements are taken from the South African National Antarctic Expedition (SANAE) 49 ship. The SANAE ship’s complete course begins in Cape Town where it travels south to Antarctica, avoiding large patches of ice which it cannot move through (Leg 1). A team of experts in areas such as Oceanography onboard continually take measurements of certain conditions and aspects of the ocean such as temperature and salinity among others. Upon reaching Antarctica, the cargo of the ship is unloaded to take to those people living and working in the Antarctic base (Leg 2). The ship then performs a round trip North-West to the island of South Georgia (Leg

3) and back to Antarctica (Leg 4) where the ship once again is docked (Leg 5). This specific data was obtained from Leg 6 of this journey which is the return leg to South Africa and runs from Antarctica to Cape Town. This data set will henceforth be referred to as SANAE49L6. The data set consists of 9215 observations, each of which contains 27 variables which are measured *in situ* from the ship. This return leg started on 12 February 2010 at GPS time 00:04:48 and ended on 22 February 2010 at GPS time 23:55:54 traveling north from (70.6245°S, 0.0001°W) to (34.073°S, 17.4585°E). Figure 2.3 indicates the traveling path of the SANAE49 ship. The plot on the left indicates the path traveled in Leg 1 from Cape Town to Antarctica, while the right hand side plot indicates the Leg 6 traveling path.

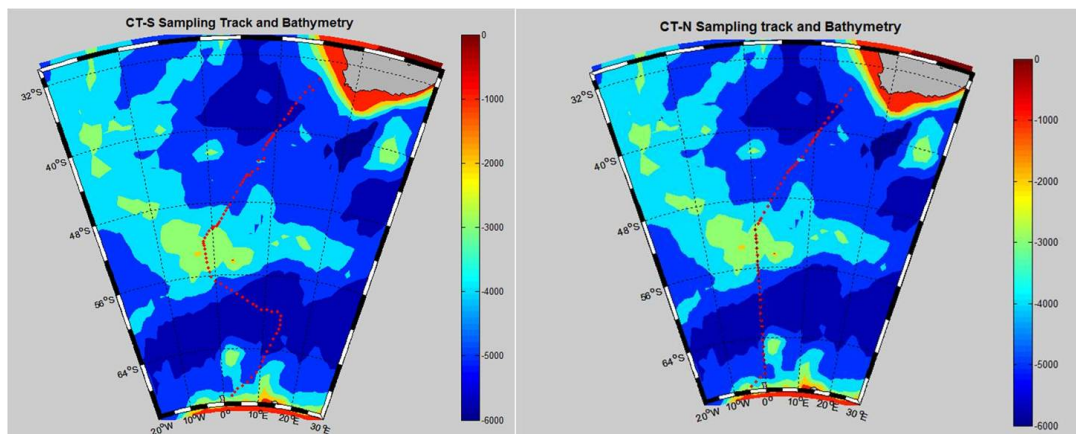


Figure 2.3: Traveling path of the SANAE49 ship

The data set is reduced to include only the variables of interest which in this case are given by: Date, GPS Time, Latitude, Longitude, fCO₂ water, Intake Temperature, Salinity, O₂ sat, O₂(ppm), pH and chlorophyll-a Concentration. This creates the reduced data set, henceforth referred to as SANAE49L6-EQU, comprising of 8424 observations of 11 variables each. The variables used are described in Table 2.1

Table 2.1: Variables of SANAE49L6-EQU

Variable	Explanation
Date	Date of Measurement (mm/dd/yyyy)
gps time	Time of Measurement (hh:mm:ss)
latitude	Latitude Measurement (Negative = South) where observation was taken in degrees
longitude	Longitude Measurement (Negative = West) where observation was taken in degrees
fCO ₂ water	Sea surface Water fugacity of CO ₂ used to calculate pCO ₂ in micro atmospheres (μatm)
Intake.Temperature	Outside SST in degrees centigrade ($^{\circ}\text{C}$)
Salinity	Salt content of the Water in parts per million (ppm)
O ₂ (%sat)	Oxygen Concentration in % Saturation (about right but not calibrated)
O ₂ (ppm)	Oxygen Concentration in micrograms per litre ($\mu\text{g}/\text{l}$) (about right but not calibrated)
pH	Water pH on a scale from 0 to 14 (Not accurate but diagnostically useful in relative units)
Ch.conc	Chlorophyll-a Concentration: Fluorescence Units in $\mu\text{g}/\text{l}$ (not calibrated)

2.4.3 Data cleaning

Although in the previous section, the full data set was reduced to a smaller, more relevant data set, it was still necessary to check that the data was “clean”. Data cleaning was performed to improve the quality of the data, involving correction or removal of large, obvious errors in the data due to machine or human error.

2.4.3.1 Locating spikes in the data

The primary variable of interest, i.e. the response, is fCO₂ water (henceforth referred to as fCO₂) or the measured level of fCO₂ in the sea surface water, measured in μatm as described in Table 2.1. Recall that the “f” in fCO₂ refers

to the fugacity, which relates to the concentration of CO₂ in the ocean. The rest of the variables are taken to be explanatory variables (independent variables), where our interest lies in modelling the behaviour of the *in situ* fCO₂ in terms of *in situ* explanatory variables. Figure 2.4 depicts the exploratory plots of the data in the SANAE48L6-EQU data set where each of the observed values of fCO₂, chlorophyll-a concentration, pH, oxygen saturation, oxygen parts per million, salinity and intake temperature are plotted against the latitude at which the measurement is made. Each plot has its own scale and set of axes, but may be plotted on the same graph. It is clear that significant “spikes” (indicated within the red ovals) in the data occur around 60°S, 50°S and 40°S in one or more of the plots of the variables. These “spikes” do not follow the pattern of the rest of the measurements for the respective variables and therefore these observations are identified as being potentially erroneous measurements. Although the plots for the Oxygen saturation and parts per million seem to have many “spikes”, it has been advised that these variables do tend to vary much more than the others as well as the fact that, as was seen in the variable description in Table 2.1, the Oxygen measurements were not calibrated.

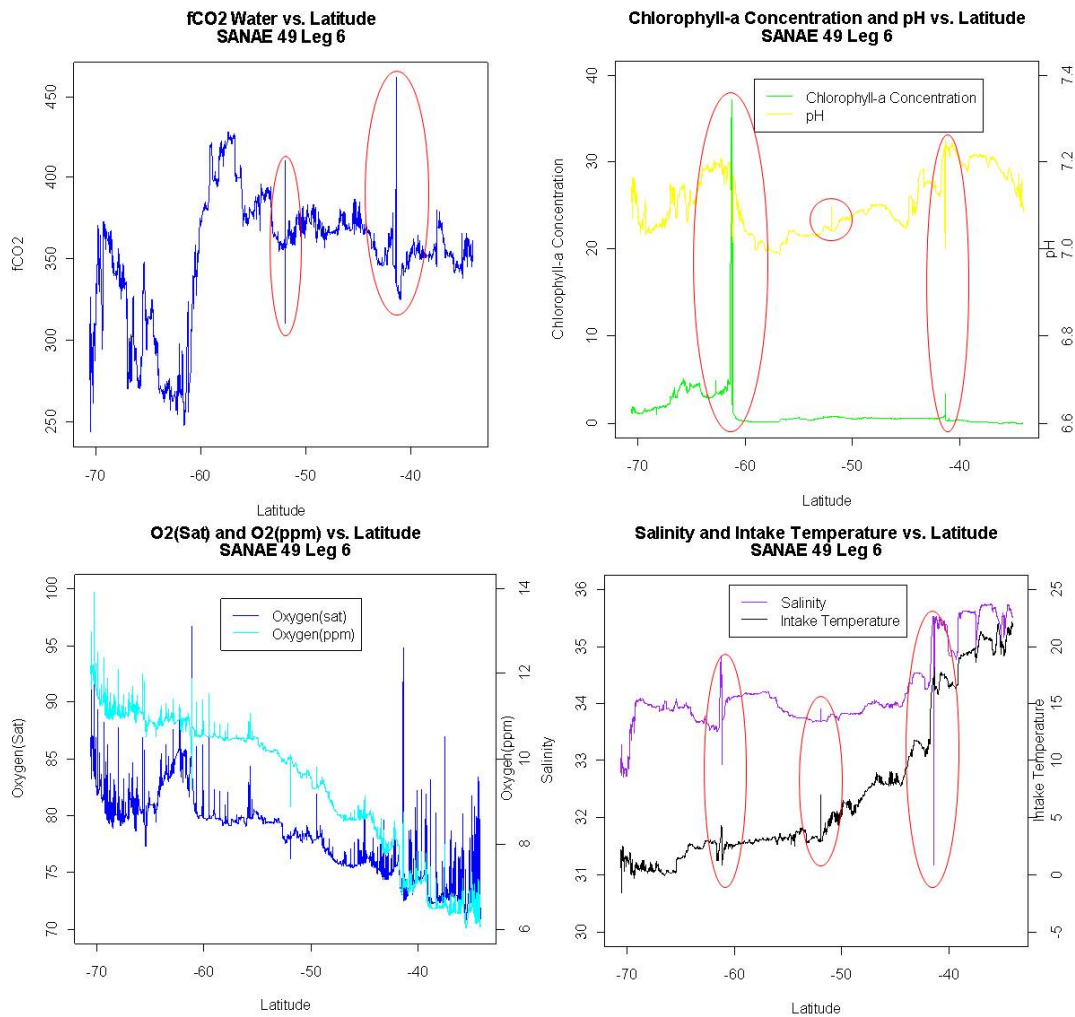


Figure 2.4: Plots of variables from SANA E49L6-EQU

The identified observations in the data were identified from the data set and queried with the domain experts from the SOCCO group. The data points, when confirmed to be faulty, were eliminated completely from the data set.

2.5 SANA E49L6-MLD data set

2.5.1 Introduction

A second data set, containing the mixed layer depth (MLD), described as the depth at which a change in ocean temperature of 0.5°C was obtained from the SOCCO group. Many other methods of measuring MLD exist, however

all methods tend to produce similar measurements of MLD. This data set is henceforth referred to as SANAE49L6-MLD and contains 3 columns and 244 rows. Measurements in this data set begin at (69.5998°S, 5.9036°W) and end at (34.073°S, 17.4585°E). These measurements were not taken at the same spatial locations (with regards to latitude and longitude) or intervals as the fCO₂ data. It is therefore required to interpolate the MLD measurements in order to obtain MLD values on the same scale as the fCO₂ observations so as to include them in a data set using the spatial scale of the *in situ* fCO₂ measurements. This section describes the methods used in reducing and combining the SANAE49L6-EQU2 and SANAE49L6-MLD data sets into a single data set, which will henceforth be referred to as SANAE49L6-EQU3.

2.5.2 Reducing latitude values

An initial reduction of the fCO₂ data set is required in order to interpolate the MLD measurements and also to comply with information received from experts in the field from the SOCCO group. On the Antarctica side, all observations south of the first MLD latitude value are deleted in order to eliminate problems with interpolation beyond boundaries. This implies that all data observations south of 69.5998°S were ignored. Secondly, on the Cape Town side of the trip, all observations north of 37°S were ignored because they may be affected by the continental shelf. This is done in compliance with expert guidance and according to research indicating an effect of the continental shelf on measurements of fCO₂ (Tsunogai *et al.*, 1999).

The next step in the process of obtaining a single data set is to interpolate the MLD measurements in the SANAE49L6-MLD to correspond to the latitude measurements in the SANAE49L6-EQU3 data set. This is done using an Euclidian distance weighted averaging method, where the Euclidean distance from each GPS co-ordinate in the SANAE49L6-EQU3 is calculated to its nearest GPS co-ordinates on either side (i.e. North and South of the tar-

get GPS co-ordinate) in the SANAE49L6-MLD data set. The total Euclidean distance between the 2 co-ordinates in the SANAE49L6-MLD data set via the co-ordinate in the SANAE49L6-EQU3 data set is calculated and then the ratio (with respect to the total distance) of distances from each GPS co-ordinate in the SANAE49L6-MLD data set to the GPS co-ordinate in the SANAE49L6-EQU3 data set are calculated and used as the weights for the 2 MLD measurements corresponding to the GPS co-ordinates in the SANAE49L6-MLD data set. The MLD value closer in Euclidean distance to the target GPS co-ordinate in the SANAE49L6-EQU3 data set is assigned the heavier weight. Finally this weighted average of the 2 MLD values using the corresponding weights assigned is calculated and assigned to the MLD measurement at the target point in space. This Euclidean distance weighting method is graphically represented in Figure 2.5.

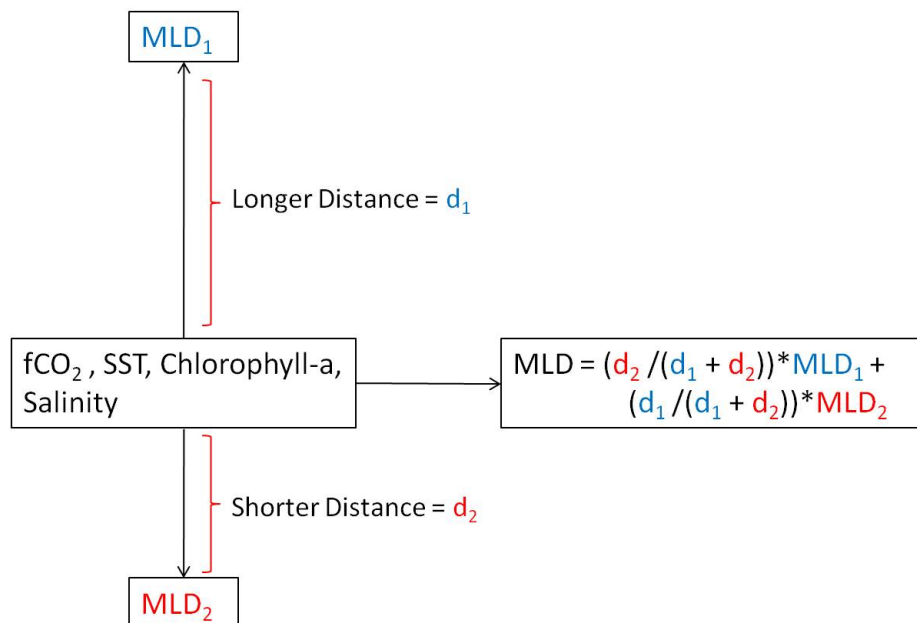


Figure 2.5: Euclidean distance weighted interpolation of MLD values

2.5.3 Description and exploratory analysis

Once the MLD measurements in the SANAE49L6-MLD data set are interpolated to the same co-ordinate references as the SANAE49L6-EQU3 data set, the 2 data sets are combined to obtain a final, clean data set known further as SANAE49L6-final. It was also determined, at this stage, that data observations 4353, 4354, 5270 and 5271 in the combined SANAE49L6-MLD and SANAE49L6-EQU3 data sets had missing values for fCO₂ due to a failure in the measurement machine and therefore, it was decided to remove these observations before the final data set was constructed. The SANAE49L6-final data set comprises of 12 columns and 6101 rows (observations). The starting date of measurements in the SANAE49L6-final data set is 13 February 2010 at 18:07:50 GPS time and the finishing date is 21 February 2010 at 18:30:53 GPS time. The measurements are obtained between the GPS co-ordinates (69.5998°S, 5.9036°W) and (37.0004°S, 12.918°E). The only variable, namely MLD, added to the list of variables in Table 2.1 is explained in Table 2.2.

Table 2.2: Explanation of new variables in SANAE49L6-final

Variable	Explanation
MLD	Mixed Layer Depth (Meters) Observed

Figure 2.2 depicts the line plots of the variables in SANAE49L6-final versus latitude.

Once the data has been cleaned and all variables put on the same co-ordinate basis, it is possible to perform a preliminary exploratory data analysis. This section is devoted to the initial study of the data using descriptive statistics and graphical checks in order to identify significant statistical properties of the data. The plots included previously in Figures 2.4 and 2.6 also represent a graphical approach to the exploratory analysis and will be discussed in more detail here.

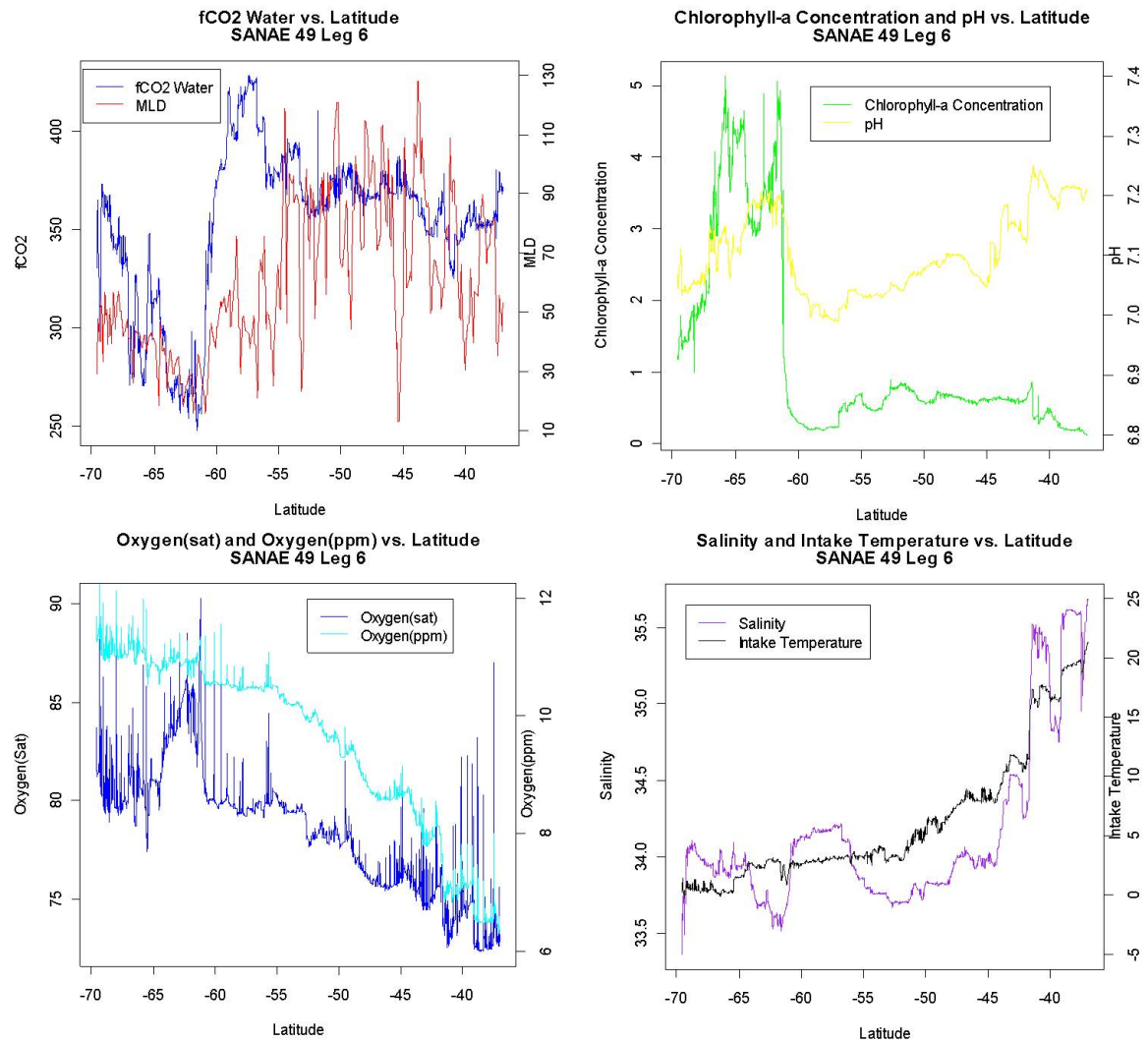


Figure 2.6: Despiked variable plots from SANAE49L6-final

In order to obtain a sense of the range and variability of the data, as well as how many observations are recorded, the descriptive statistics are generated. These are given in Table 2.3 which provides the number of observations, number of missing values, mean, standard deviation and coefficient of variation. Not all 12 variables in the SANAE49L6-final data set are included in the descriptive statistics since variables such as the GPS time, latitude and longitude have already been discussed with regards to the final data set. As can be seen from Table 2.3, there are no missing values in the final data set. The means indicate the location of the variables, while the standard deviations allow for the examination of the average spread of the variables around their respective

Table 2.3: Descriptive statistics of SANAE49L6-final

Variable	N	Missing	Mean	Standard Deviation	Coefficient of Variation
fCO ₂	6101	0	354.03	37.08	0.10
Salinity	6101	0	34.16	0.55	0.02
Oxygen Saturation	6101	0	78.28	3.05	0.04
Oxygen (ppm)	6101	0	9.53	1.46	0.15
pH	6101	0	7.10	0.07	0.01
Chlorophyll-a Concentration	6101	0	1.16	1.23	1.07
Intake Temperature	6101	0	6.29	5.68	0.90
MLD	6101	0	61.57	24.92	0.40

means. Comparing the standard deviations to one another, however, will not provide us with any additional information due to the fact that the measuring units play a role in the size of the standard deviation. We therefore rather consider the coefficient of variation (COV) which is a measure of the relative spread of the observations around the means. It is obtained by dividing the standard deviation by the respective mean value. The COV indicates that variables such as fCO₂, Salinity, Oxygen saturation and ppm and finally pH do not vary a great deal in comparison to their means. This observation is specifically interesting since from Figures 2.4 and 2.6 it seemed that Oxygen saturation and ppm had very high variance. Chlorophyll-a concentration, however has a standard deviation of more than 100% of its mean (1.07 or 107%), indicating that it varies a great deal in comparison to its mean. If we consider the top right plot of Figure 2.6 it is seen that most of this variation is observed close to Antarctica (i.e. at a lower latitude) until approximately 60°S latitude, and north of this the chlorophyll-a concentration is relatively constant. The intake temperature also varies a great deal, with a standard deviation of approximately 90% of its mean value. This, however can be explained by the fact that the intake temperature is expected to rise significantly as the ship travels further north towards Cape Town since the ocean waters tend to be

much warmer and therefore the mean temperature measurement is not an accurate indication of the central location of the observations. This can, again, be seen in the bottom right plot of Figure 2.6. Finally the COV for the MLD is also moderately high since the standard deviation is approximately 40.5% of its mean. The top left plot of Figure 2.6 validates this, particularly north of 60°S where the measurements of the MLD are much more variable. When this fact was queried with experts from the SOCCO group, it was discovered that the MLD is dependent on climactic conditions and other factors at the time of measurement. This may provide some explanation for the extreme variability apparent from the plot in Figure 2.6 and may not be indicative of truly variable MLD's, but rather variable conditions at the time of measurement.

Table 2.4 further provides the minimum, first quartile, median, third quartile and maximum observed values in order to discuss the shape and range of the data which, as will be seen, can be potentially misleading. Again, these are only provided for the same variables as were indicated in Table 2.3.

Table 2.4: Shape and range descriptive statistics

Variable	Minimum	Q1	Median	Q3	Maximum
fCO ₂	247.32	345.13	362.39	373.85	428.29
Salinity	33.36	33.82	33.98	34.18	35.69
Oxygen Saturation	72.30	75.80	78.90	80.00	90.30
Oxygen (ppm)	6.34	8.59	10.06	10.78	12.39
pH	6.99	7.05	7.09	7.15	7.25
Chlorophyll-a Concentration	0.12	0.46	0.62	1.44	5.14
Intake Temperature	-0.28	2.65	3.60	8.37	21.30
MLD	13.15	42.08	55.84	82.43	127.93

The descriptive statistics in Table 2.4 are generally used to indicate the form and range of the observed values of the variables. The response variable, fCO₂, has a wide range of approximately 180.97 μatm. This indicates that the median of the fCO₂ values (362.39) is almost in the centre of the range and is

close to the mean value in Table 2.3 of 354.03. This may provide evidence to suggest that the fCO₂ values are symmetric, but as will be seen later in the graphical approach, this is not the case. The salinity and pH do not have very wide ranges. With respect to the pH, this indicates that the pH of the SO is very close to 7, which is an indication of a neutral system (i.e. not alkaline or acidic). This can be seen in the fact that the maximum pH observed is 7.25, while the minimum is 6.99. The variability observed in the MLD and chlorophyll-a concentration in Table 2.3 is seen again here in that the range is large with the MLD ranging from approximately 13 meters to almost 130 meters, while the large range observed in the chlorophyll-a concentrations is again due mainly to the high variability near Antarctica. This is confirmed by median chlorophyll-a concentration being much closer to the minimum value than the maximum value. As expected, the intake temperature has a wide range due to the ocean temperature becoming much warmer as the ship travels further north towards Cape Town.

2.5.4 Graphical approach to exploratory analysis

This section takes a graphical approach to examining the statistical structures of the data. In Figure 2.7 we plot the histograms of the variables. From the graphs in Figure 2.7 it is observed that the fCO₂ measurements are not normally distributed. They appear to be multi-modal (i.e. the distribution has more than one mode). Since this is the response variable of interest in this study, it is important that we obtain a model which captures this distribution. The distributions of the MLD and pH observations seem to be skewed to the right, with the majority of observations coming between 40 and 50 meters or around 7.05 respectively. This indicates that, in most areas of the SO, the MLD is shallow and also neutral in terms of pH. There also seems to be an increased frequency of observations of pH levels which are slightly more alkaline (around 7.2). Finally, a large majority of the salinity levels observed

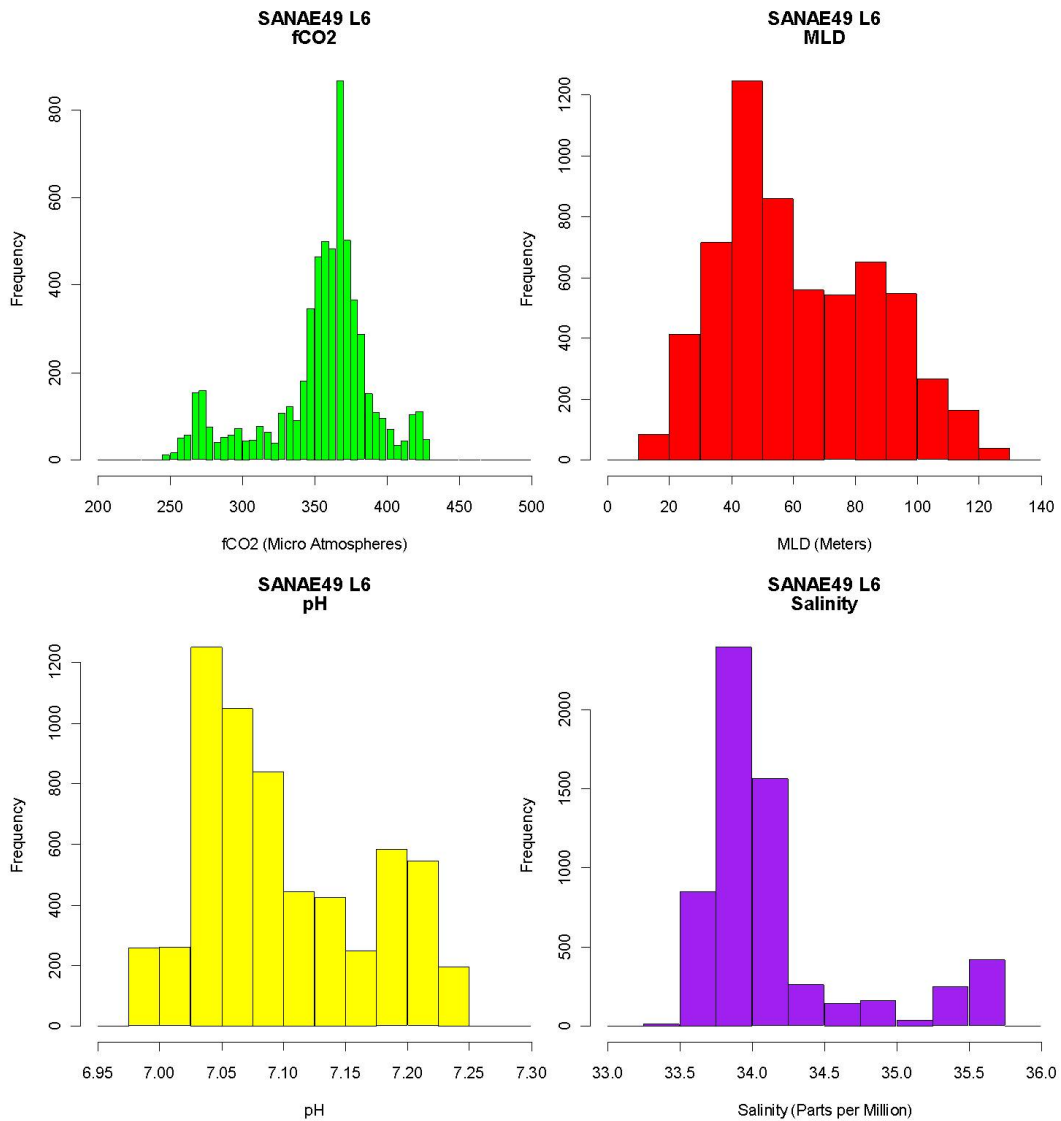


Figure 2.7: Histogram of fCO₂, MLD, pH and salinity

in the SO seem to be less than 34.5 ppm. There is, however, a slight increase in the frequency of salinity observations around 35.5 ppm.

The graphs in Figure 2.8 provide histograms for the remaining variables discussed thus far. They indicate the distribution of the observed values of intake temperature, chlorophyll-a concentration, oxygen (ppm) and oxygen (Saturation). The form of the distributions of intake temperature and chlorophyll-a concentration seem very similar. They both are skewed to the right, with the majority of observations of chlorophyll-a concentration being between 0 and 1 microgram per liter while the majority of SST measurements observed are

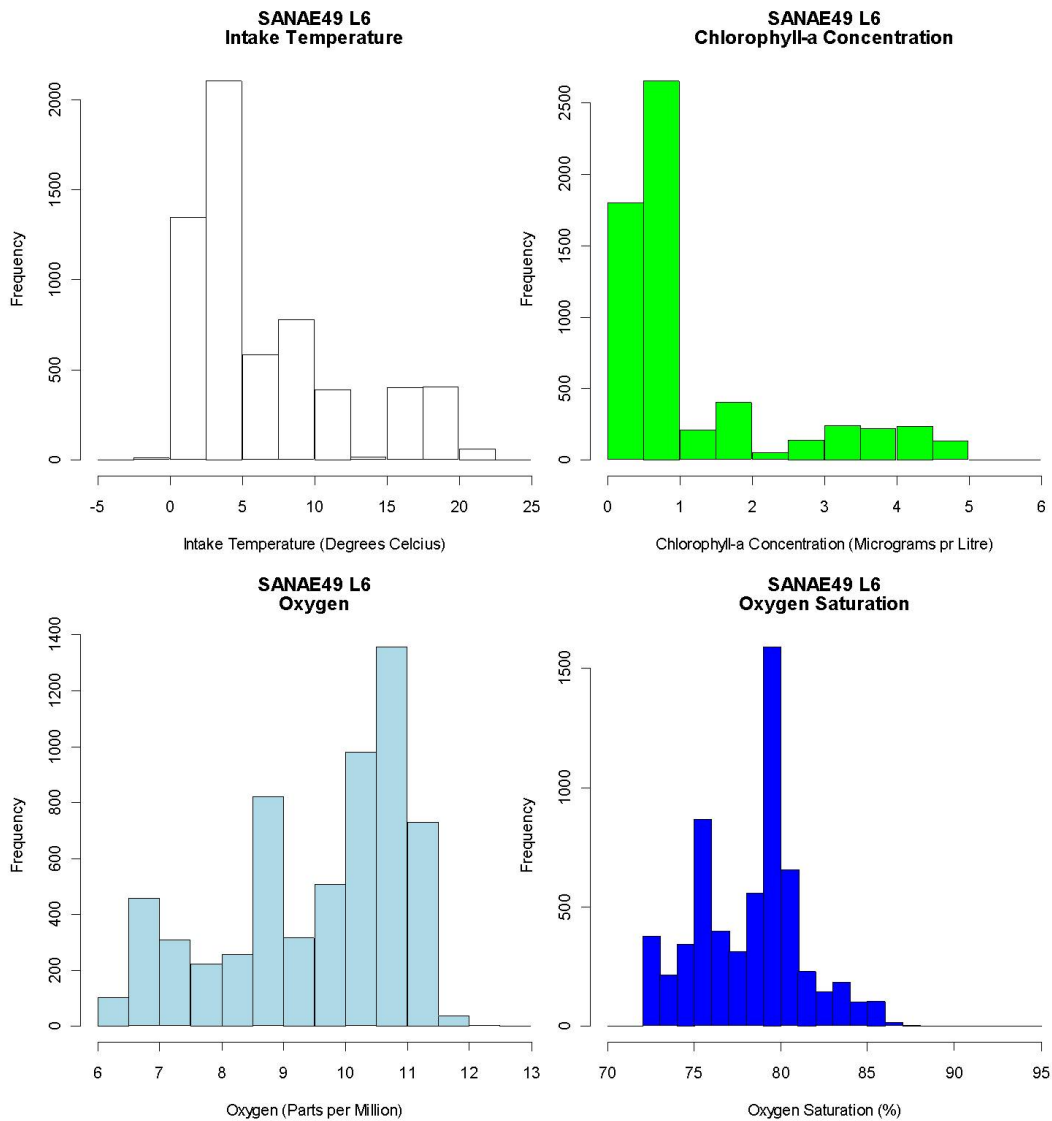


Figure 2.8: Histogram of Intake Temperature, Chlorophyll-a Concentration, Oxygen (ppm) and Oxygen (Saturation)

close to 0°C. When considering the corresponding plot for chlorophyll-a concentration in Figure 2.6 the plot indicates that for large areas of the SO the chlorophyll-a concentration observed was very close to 0 $\mu\text{g}/\text{l}$. This is due to areas further north of the Antarctic having large sections of ocean where little, or no, biological activity is present (in the form of chlorophyll-a blooms). The final 2 histograms in Figure 2.8 are of oxygen concentration and ppm seem to be more irregular, with each having multiple modes. The majority of observed levels of oxygen seem to be around 11 ppm with a saturation of around 80%.

2.6 Summary

This chapter discussed the SANAE49L6 data set as well as its reduction and combination with the MLD data set provided by the SOCCO group in order to develop a single data set for further analysis. The final, clean data set produced (i.e. SANAE49L6-final) displayed some interesting properties that will have to be considered in further discussions. Firstly, the histogram of the response variable of interest - fCO₂ - seemed to indicate a complex and specifically multi-modal structure for its distribution. This structure must be accounted for in the modelling procedure. Proposed statistical models for this problem considered in later chapters include the multiple linear regression (MLR) model, which has stringent assumptions and an inflexible form, but a simple and easily explainable formula; as well as the non-parametric kernel regression (NPKR) approach which accounts for complex data distributions by using a data dependant model whereby the observations themselves determine the form of the regression function.

Secondly, the data reveals a clear separation in the observed form of the plots for some of the covariates such as chlorophyll-a concentration, MLD, pH and even salinity. There seems to exist an area in space where the behaviour of the measurements changes. If we consider Figure 2.6, it can be seen that this area of change seems to be between 60°S and 55°S. The chlorophyll-a concentration before this is much higher and more variable, whereas after this area it seems to be very low (around 0µg/l) and does not vary much. The pH level in the ocean drops significantly in this area of the ocean, while the MLD becomes much more variable north of 55°S. Finally, salinity has a large increase in the observed values of this area. All these observations seem to suggest that the SO is a complex system and would not, thus be suited to the fitting of only a single rigid model.

Finally, all the observations are obviously not spatially independent from one another. Since measurements are made along a spatial “time-line” and

since the ocean has certain characteristics in certain areas, it is important that this spatial dependence be removed from the data. This is not the focus of this thesis, but is of importance since this may improve the generalizing ability of the model to other data sets and further to satellite data in order to predict the fCO₂ levels for the entire SO.

Chapter 3

Parametric regression model for CO₂ concentrations

3.1 Introduction

In Section 1.2.1 the objectives of this thesis outlined the need for an approach for predicting and extrapolating *in situ* predictions of fCO₂ in the SO to unsampled areas. The first of the methods proposed to achieve this regresses fCO₂ onto the set of independent predictor variables selected according to both their availability from the *in situ* measurements from the SANAE 49 ship and their usefulness in being applied to a more global database in which satellite (remote) measurements are used. Multiple linear regression (MLR) is discussed in this chapter and applied to the SANAE49L6-final data set as elaborated on in Section 2.4. In this chapter in Section 3.2 we provide a review of literature on MLR and specifically its uses for predicting or extrapolating CO₂ data; Section 3.3 discusses the MLR methodology of estimating the regression parameters. The method is then applied to the SANAE49L6-final data set and a discussion of the results follows in Section 3.6 in order to understand how these results will be further used to compare with models developed in later chapters. The objective of this chapter is thus to determine if MLR is an

appropriate approach to model fCO₂ in the SO.

3.2 Modelling ocean CO₂ with MLR

MLR is a widely used method for not only predicting concentrations of CO₂ in the ocean, but also in explaining the high variations in these values through predictions of partial pressure of CO₂, as well as dissolved inorganic carbon (DIC) (Bates *et al.*, 2006; McNeil *et al.*, 2007; Slansky *et al.*, 1997; Wallace, 1995). This method allows for the variations in CO₂ concentrations observed in the surface ocean to be explained using a set of independent variables. McNeil *et al.* (2007) proposed a MLR model for DIC in the ocean as well as for the alkalinity. These models were used to provide estimates for these values, which were then used to estimate the flux of CO₂ between the air and sea in the SO. The results of the MLR models provided by McNeil *et al.* (2007) served to confirm the results by Takahashi *et al.* (2002) which identified a CO₂ sink in the SO both below 50°S and within the sub-Antarctic zone (between 40°S and 50°S).

A study by Jamet *et al.* (2007) also makes use of the MLR method by using VOS data obtained in the North Atlantic Ocean as the *in situ* data on which to develop and assess the MLR models. The independent variables proposed in Jamet *et al.* (2007) are the SST, chlorophyll-a concentration and the MLD. These predictor variables will also be used in the models along with the latitude and salinity in order to investigate their effect on the predictive ability of the MLR models. According to Jamet *et al.* (2007), SST (usually satellite measured values) have been widely used as an independent or explanatory variables in the development of extrapolated maps of CO₂ concentrations (Boutin *et al.*, 1999; Lee *et al.*, 1998; Nelson *et al.*, 2001; Olsen *et al.*, 2003, 2004; Stephens *et al.*, 1995). This is not, however, the only variable that may provide useful information in order to produce a more accurate interpolation

of CO₂. Measures of chlorophyll-a concentration, obtained from ocean colour satellites provide an indication of biological activity in the ocean which can affect ocean CO₂. Some recent studies have focused on trying to incorporate this type of independent variable into their models as well as including a measure of satellite ocean salinity (Jamet *et al.*, 2007; Ono *et al.*, 2004; Rangama, 2005). Because satellite ocean salinity measurements may be unreliable it will not be included for further model development in the later chapters of this thesis. The importance of salinity in influencing concentrations of CO₂ has, however, been discussed in previous studies and therefore it is important to observe its effect in the models' predictive ability as well as attempt to find another variable which can capture its effect (Sarma *et al.*, 2006). A further measure of vertical mixing of carbon dioxide, known as the mixed layer depth (MLD), has been used to explain the variation in the surface ocean CO₂ flux. Lüger *et al.* (2004) identified that the air-sea gas exchange of CO₂ is dependent on the vertical mixing. This casted new light on previous results which indicated a lack of correlation between the MLD and the levels of CO₂ concentration (Dandonneau, 1995). However, both Lüger *et al.* (2004) and Dandonneau (1995) focused on small regions of ocean (in most cases identified by the biogeochemical provinces as proposed by Longhust (1995)). Finally, the latitude at which the CO₂ concentration is measured is included. This method has been used in previous applications and it seems to have improved the fit of the models in previous studies (at least for small areas of ocean waters) (Stephens *et al.*, 1995; Lefèvre and Taylor, 2002; Jamet *et al.*, 2007). The focus of this thesis, however is to identify and model the relationship between fCO₂ and the physical and bio-geochemical processes in the ocean. Latitude does not fit into this framework and therefore will be excluded from the final model.

This provides evidence for the use of an MLR model to predict fCO₂ in the SO which is discussed in the subsequent sections. The MLR models seemed to produce positive results in previous studies (mainly in the North Atlantic)

and therefore this approach will also be taken in the SO.

3.3 Theoretical background of linear regression

Simple linear regression (SLR), or straight-line regression is concerned with using the information obtained from a single (independent) variable, henceforth referred to as x , in order to estimate or predict another variable (dependent variable) henceforth referred to as y . The notation used in this section corresponds to the notation used by Sheather (2009) . This is done by estimating the function (in this case a straight line) which “best” fits the.

3.3.1 Simple linear regression

In this section, and in later chapters, model development will be performed on a subset of the total data set. This subset will be referred to as the training data set, while the observations not included in the training data subset will be known as the test data subset. This test data subset will be used to assess the predictive ability of the developed models by using them to predict the responses for the test data independent variables and observing the error in prediction from the observed response values. For SLR, let the training data set consists of n independent observations of the response variable Y , which we denote as y_1, y_2, \dots, y_n , corresponding to n observed values of the independent (or explanatory) variable X , denoted x_1, x_2, \dots, x_n . The regression function of these values can then be written as

$$y_i = E(Y_i|X_i = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \dots, n \quad (3.3.1)$$

In (3.3.1), SLR attempts to estimate the expected response for a given value of the independent variable. The remaining part of the SLR function (i.e. the e_i values) represent a random or unexplained error in each of the observed responses. Included in the SLR approach are certain, inherent assumptions that must be made regarding the data:

- The response variables in the data should be mutually independent of one another. (This tends not to be the case in both the *in situ* and satellite data, however for simplicity purposes it will be assumed to hold for all data in this thesis);
- The random errors should follow a normal distribution;
- The conditional expected value of the random errors given the independent variable in the regression model must be 0, i.e. $E(e_i|X_i = x_i) = 0$ for all $i = 1, 2, \dots, n$;
- The conditional variance of the random errors given the independent variable discussed in Section 3.3.1 should be a constant value, which is denoted as σ^2 ;
- The β_0 and β_1 values are referred to as the regression parameters. These values describe the mean response value as well as the relationship between independent and response variables respectively.

3.3.2 Multiple linear regression

The multiple linear regression (MLR) model is simply an extension of the SLR model defined in (3.3.1) with the inclusion of more than one independent variable. The training data set for MLR still consists of the n independent observations of the response or dependent variable, Y , however now each observed set of independent variables consists of a vector of p realisations of the independent variables. These vectors are denoted as $\underline{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$, where the first value in each subscript refers to the variable number and the second refers to the observation number. The MLR function, therefore, can be written in a similar form to (3.3.1) as

$$Y_i = E(Y_i|\underline{X}_i = \underline{x}_i) + e_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i \quad (3.3.2)$$

(3.3.2) has the same assumptions as the SLR method discussed in Section 3.3.1 with regards to the response variable and the error terms in the model. The term “linear” in this model, however, no longer refers to a straight line function, but rather that the relationship between the response variable Y and the independent variables X_1, X_2, \dots, X_p is linear in terms of the parameters (i.e. the β_i parameters). The models may include interactions between the variables. As in Section 3.3.1, the goal of MLR is to estimate the function $E(Y_i|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$.

3.3.3 Least Squares Minimisation

Consider the least squares estimate of the regression parameters $\beta_0, \beta_1, \dots, \beta_p$ to be b_0, b_1, \dots, b_p such that the estimated regression function used to predict the response variable for a given set of input variables is given by

$$\hat{y}_i = E(Y_i|X_i = \underline{x}_i) = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_px_{pi} \quad i = 1, 2, \dots, n \quad (3.3.3)$$

The true responses can be represented by adding an estimated random error value to (3.3.3) denoted by \hat{e}_i . The residual sum of squares (RSS) can therefore be represented as

$$\begin{aligned} RSS &= \sum_{i=1}^n e_i^2 \\ &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1x_{1i} - \dots - \beta_px_{pi})^2 \end{aligned} \quad (3.3.4)$$

In order to estimate the regression parameters $(\beta_0, \beta_1, \dots, \beta_p)$, it is required to minimise the RSS with respect to $\beta_0, \beta_1, \dots, \beta_p$ as follows

$$\begin{aligned} \frac{\partial RSS}{\partial \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1x_{1i} - \beta_2x_{2i} - \dots - \beta_px_{pi}) = 0 \\ \frac{\partial RSS}{\partial \beta_1} &= -2 \sum_{i=1}^n x_{1i}(y_i - \beta_0 - \beta_1x_{1i} - \beta_2x_{2i} - \dots - \beta_px_{pi}) = 0 \\ &\vdots \\ \frac{\partial RSS}{\partial \beta_p} &= -2 \sum_{i=1}^n x_{pi}(y_i - \beta_0 - \beta_1x_{1i} - \beta_2x_{2i} - \dots - \beta_px_{pi}) = 0 \end{aligned} \quad (3.3.5)$$

3.3.4 Matrix notation of the least squares minimisation

Let $\underline{Y}_{(1 \times n)}$ denote the n -dimensional vector of responses and let $\mathbf{X}_{(p \times [n+1])}$ denote the matrix of n observed values of the p independent variables preceded by a column of 1's. Further let the 2 column vectors, $\underline{\beta}$ (of size $(1 \times p)$) and \underline{e} (of size $(1 \times n)$) represent the vector of regression parameters and error terms in the MLR functions respectively. In matrix notation, (3.3.2) can be written as

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{e} \quad (3.3.6)$$

The constraints on the model discussed in Section 3.3.1 still hold and therefore since $E(\underline{e}|\underline{X} = \underline{x}) = 0$, the function that will be estimated using least squares can be written as

$$E(Y|\underline{X} = \underline{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (3.3.7)$$

$$= \underline{x}'\underline{\beta} \quad (3.3.8)$$

In (3.3.7), the function estimates each response value (Y) for a given set of p input variables (\underline{X}) using $p + 1$ parameters which must be estimated. As described in Section 3.3.3, these parameters will be estimated by minimising the RSS as given below

$$\begin{aligned} RSS &= \underline{e}'\underline{e} \\ &= (\underline{Y} - \mathbf{X}\underline{\beta})'(\underline{Y} - \mathbf{X}\underline{\beta}) \\ &= \underline{Y}'\underline{Y} + \underline{\beta}'(\mathbf{X}'\mathbf{X})\underline{\beta} - 2\underline{Y}'\mathbf{X}\underline{\beta} \end{aligned} \quad (3.3.9)$$

In order to minimise the RSS with respect to the model parameters, $\underline{e}'\underline{e}$ must be differentiated with respect to $\underline{\beta}$ and the resulting equation set equal to 0. This differentiation requires a knowledge of matrix differentiation which can be found in Golub and van Loan (1996). Differentiating (3.3.9) and setting it equal to 0 results in

$$2\underline{\beta}'(\mathbf{X}'\mathbf{X}) - 2\underline{Y}'\mathbf{X} = 0 \quad (3.3.10)$$

Solving for the column vector $\underline{\beta}$ in (3.3.10) the least squares estimate for the regression parameters is given by

$$\hat{\underline{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y} \quad (3.3.11)$$

This implies that the predicted response values for a given set of input vectors (each of length p) is given by

$$\hat{\underline{Y}} = \mathbf{X}\hat{\underline{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y} \quad (3.3.12)$$

The predicted responses for a given regression problem can be written simply as a linear combination of the observed response values in the column vector \underline{Y} .

3.4 Linear models to predict fCO₂

This section is concerned with the MLR model which will be used to predict fCO₂ in the SO based on the *in situ* observations taken from the SANAE49 ship travelling on Leg 6 of the journey (from Antarctica to Cape Town). This data is discussed in Section 2.5.3 as the SANAE49L6-final data set. The models are all developed using least squares regression as discussed in Section 3.3 and the variables considered for modelling purposes are SST, chlorophyll-a concentration, MLD, salinity and latitude. Jamet *et al.* (2007) included the Longitude co-ordinate of the measurements in the models as well in order to compare to previous analyses which have included the geographic co-ordinates as independent variables in the models (Stephens *et al.*, 1995; Lefèvre and Taylor, 2002). Due to the *in situ* nature of the data, longitude co-ordinates were not included in the models as an independent variable since the longitude measurements observed only covered a small area in comparison to the entire SO.

It is important that the test data subset be kept completely separate from the model development procedure so that the model can be evaluated by quantifying the error of the predictions. The errors were recorded as the difference

between the predicted response (\hat{y}) and the true observed response (y) in the test data subset. This gives a measure of how well the models can generalise to an “unseen” test data set.

Due to observed chlorophyll-a concentration values having several orders of magnitude as identified by Jamet *et al.* (2007), it was preferred to not use the observed values directly, but rather to transform the values by taking the \log_{10} of the observed values before including them in the model. Later studies using international databases (such as the SOCAT database) require the models developed on *in situ* data to only include independent variables which are available through remote sensing such as SST, chlorophyll-a concentrations, MLD and Latitude. This section, includes SSS in order to understand whether, in future studies, salinity could provide helpful information in estimating oceanic fCO₂. As identified by previous studies, including the spatial position of the measurements as an independent variable in the models generally produces a better fit model for small regions, therefore the effect of including latitude will be examined, but not used for the extrapolation of fCO₂ values to the rest of the SO (Stephens *et al.*, 1995; Lefèvre and Taylor, 2002; Jamet *et al.*, 2007). The models developed in this thesis, containing the variables discussed above are listed in Table 3.1

Table 3.1: MLR Models Investigated

Model	Variables
M1	SST, Log(chlorophyll-a) and MLD
M2	SST, Log(chlorophyll-a), MLD and Latitude
M3	SST, Log(chlorophyll-a), MLD and Salinity

An important characteristic of the model to study is the impact of the size of the training data subset on the predictive ability of the model developed. For this reason, the model including SST, \log_{10} (chlorophyll-a concentration) and MLD are developed on 80%, 70%, 60%, 50%, 40% and 30% training data subset of the SANAE49L6-final data set and then assessed on the remaining

test data. The resulting mean square error (MSE), mean absolute error (MAE) and root mean square error (RMSE) values will provide an indication of how the amount of data in the training data set affects the ability of the various models in predicting fCO₂ for an unseen data set.

Finally the variables (both dependent and independent), in the same model, are standardised to have a mean of 0 and a standard deviation of 1. This is done by applying a transformation to each observation in the training data set

$$x_i^N = \frac{x_i - \bar{x}_i}{\sigma(x_i)} \quad (3.4.1)$$

where x_i refers to the observed value of each independent variable, the \bar{x}_i and $\sigma(x_i)$ refer to the mean and standard deviation of that variables observations respectively. In order to care to the results of Jamet *et al.* (2007), the MLR model was then developed without an intercept term. The fCO₂ predictions obtained from this standardised model must then be converted back to it's original units by multiplying them by the standard deviation and adding the mean value of the responses in the training data subset. The model's predictive ability can then assessed by standardising the test data set in 2 different ways:

1. The observed independent variable values in the test data set can be standardised using the respective means and standard deviations for each independent variable calculated from the training data subset;
2. The observed independent variable values in the original test data subset can be standardised using the means and standard deviations for each independent variable calculated from the test data subset.

Models M1, M2 and M3 were all developed using 70% of the original data for the training data set and the remaining 30% for a test data set. Further models (M4, M5, M6, M7, M8 and M9) are developed using different divisions of test and training data, but the same predictor variables as M1 in order to investigate the effect of the training data set size on the accuracy of the

model predictions in an independent test data set (these divisions were 80%-20%, 60%-40%, 50%-50%, 40%-60%, 30%-70% and 20%-80% respectively). Finally, model M10 contains the same variables as model M1, however, as was discussed earlier, the training data set is standardised before developing the model. Model M10 is then assessed using the 2 different standardised test data sets mentioned above. Since no definite rule could be identified as to which of the two methods to use consistently, both were applied. When the models were assessed, M10_Train_Stats refers to the assessment of model M10 using the test data set standardised with the training data set means and standard deviations as set out in point 1 above. M10_Test_Stats refers to the assessment of model M10 using the test data set standardised using the test data means and standard deviations as set out in point 2 above.

3.5 Multiple linear regression results to predict fCO₂

The MLR models discussed in Section 3.4 are fit to the respective subsets of the SANAE49L6-final data set using the R function *lm* with fCO₂ as the response variable (R Development Core Team, 2011). The models developed using the non-standardised training data sets are fit with an intercept in the model, while the standardised training set models do not have an intercept since the fCO₂ values are transformed to have a mean of 0. This section provides the results of models M1 to M10, with a discussion of these results to follow.

3.5.1 Optimising the regression model

The critical objective of the MLR approach (as discussed in Section 3.3) is to produce optimum estimates of the parameters $\beta_0, \beta_1, \dots, \beta_p$ in the MLR regression function given by (3.3.2). Table 3.2 provides the least square parameter estimates for each of the variables of the 10 MLR models discussed in Sec-

Table 3.2: MLR model parameter estimates

Models	Intercept	SST	log Chlorophyll-a Concentration	MLD	Salinity	Latitude
M1	345.562	-3.326	-93.108	0.282		
M2	418.112	-4.661	-89.123	0.184		1.080
M3	580.602	-2.739	-94.253	0.249	-6.935	
M4	345.196	-3.329	-93.303	0.286		
M5	345.630	-3.337	-93.335	0.280		
M6	345.523	-3.353	-93.339	0.281		
M7	346.184	-3.325	-92.688	0.274		
M8	345.863	-3.272	-92.043	0.278		
M9	347.012	-3.337	-93.336	0.267		
M10		-0.513	-0.991	0.190		

tion 3.4. The p-values for each of these parameter estimates corresponding to the null hypothesis test $H_0: \beta_i = 0$ versus the alternative hypothesis $H_A: \beta_i \neq 0$ for $i = 0, 1, \dots, p$ (where p is 4 for models M2 and M3 and 3 for the rest) in each of the models were less than 0.0001 and therefore deemed to be highly significantly different from zero. This is, however, expected due to the large size of the data sets used to estimate the model parameters. The estimated parameters give an indication of the relationship between the independent variables and the response. A positive parameter estimate indicates a positive (direct) relationship, while a negative parameter estimate indicates a negative (inverse) relationship. These models will further be assessed using their respective test data sets in order to determine their predictive abilities based on “unseen” observations.

3.5.2 Assessing the regression model

The model assessment procedure is performed by using the models developed on the training data sets to predict fCO₂ in the test data sets. This procedure is divided into four aspects, which identify specific qualities in the models. The first aspect identifies the variables to be used in the model, as well as investigating the effect that variables such as latitude and salinity have on the

predictive ability by comparing the test error rates of models M1, M2 and M3. The second aspect indicates the effect of a decaying amount of data in the training data set on the predictive ability of the model by comparing the test error rates of the models M1, M4, M5, M6, M7, M8 and M9. This allows for the identification of how the model responds to varying amounts of data in the training data sets. The third aspect assesses the effect of standardising the data subsets on the predictive ability of model M1, developed without an intercept on a subset where each of the observed values of the variables has been transformed to have a mean of 0 and standard deviation of 1. The final aspect assesses the distribution of the error rates of model M1 by performing 100 simulations of the random subset divisions and then developing and calculating the test error rates for model M1 on each of these subset divisions.

Table 3.3 provides the error rates for the models described in Table 3.1 of Section 3.4. These error rates are all based on a random 70% – 30% division of the SANAE49L6-final data set to a training and test data set respectively. The error rates are all calculated on the test data set which is considered to be “unseen” during the model development stage.

Table 3.3: Multiple linear regression model error rates

Model	Mean Square Error	Mean Absolute Error	Root Mean Square Error
M1	328.789	14.150	18.133
M2	317.283	13.631	17.812
M3	325.125	13.903	18.031

Figures 3.1, 3.2 and 3.3 plot the observed and predicted fCO₂ values from the SANAE49L6-final test data set against latitude in order to identify how well models M1, M2 and M3 are able to predict fCO₂. The observed fCO₂ values are plotted as the blue dots, while the red line indicates the predicted values from the respective models.

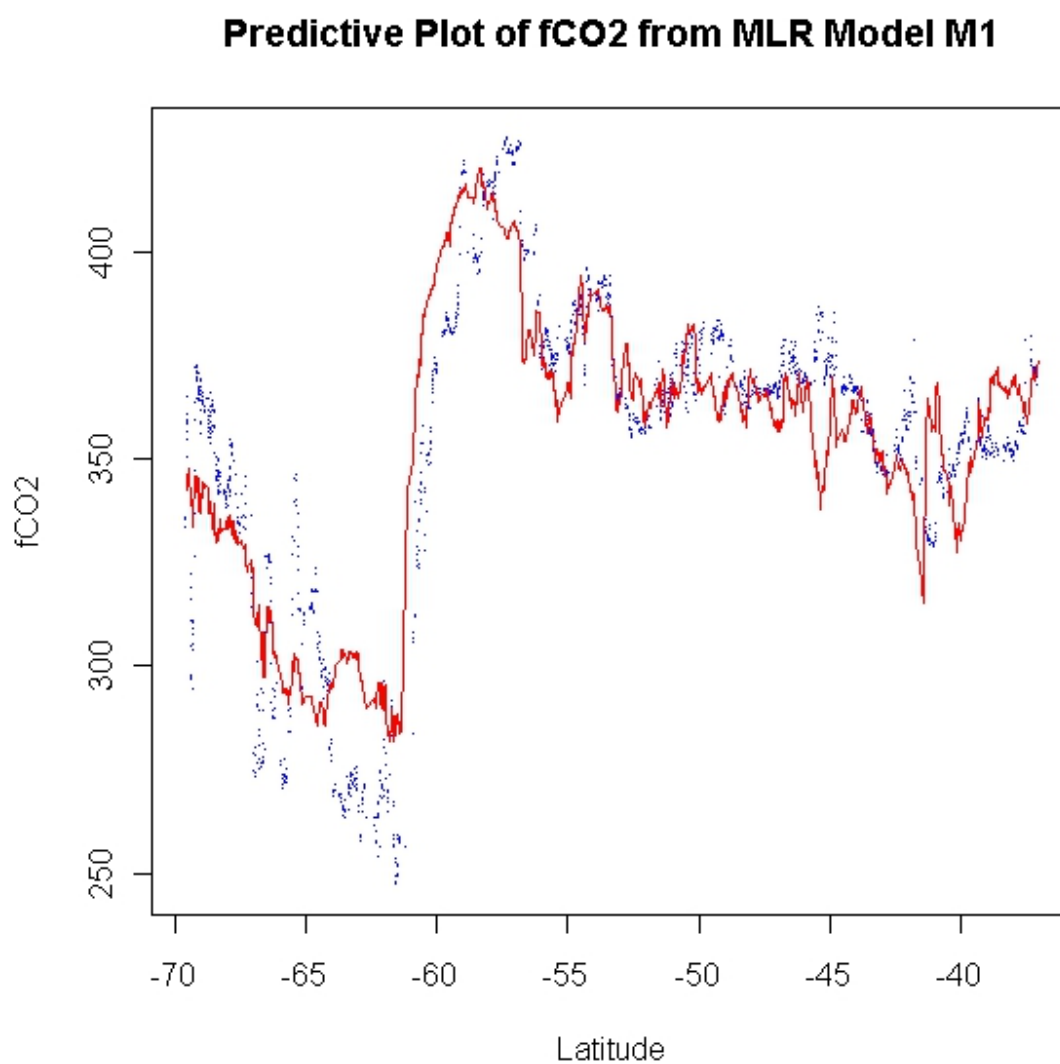


Figure 3.1: Multiple linear regression observed and predicted fCO₂ for model M1 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

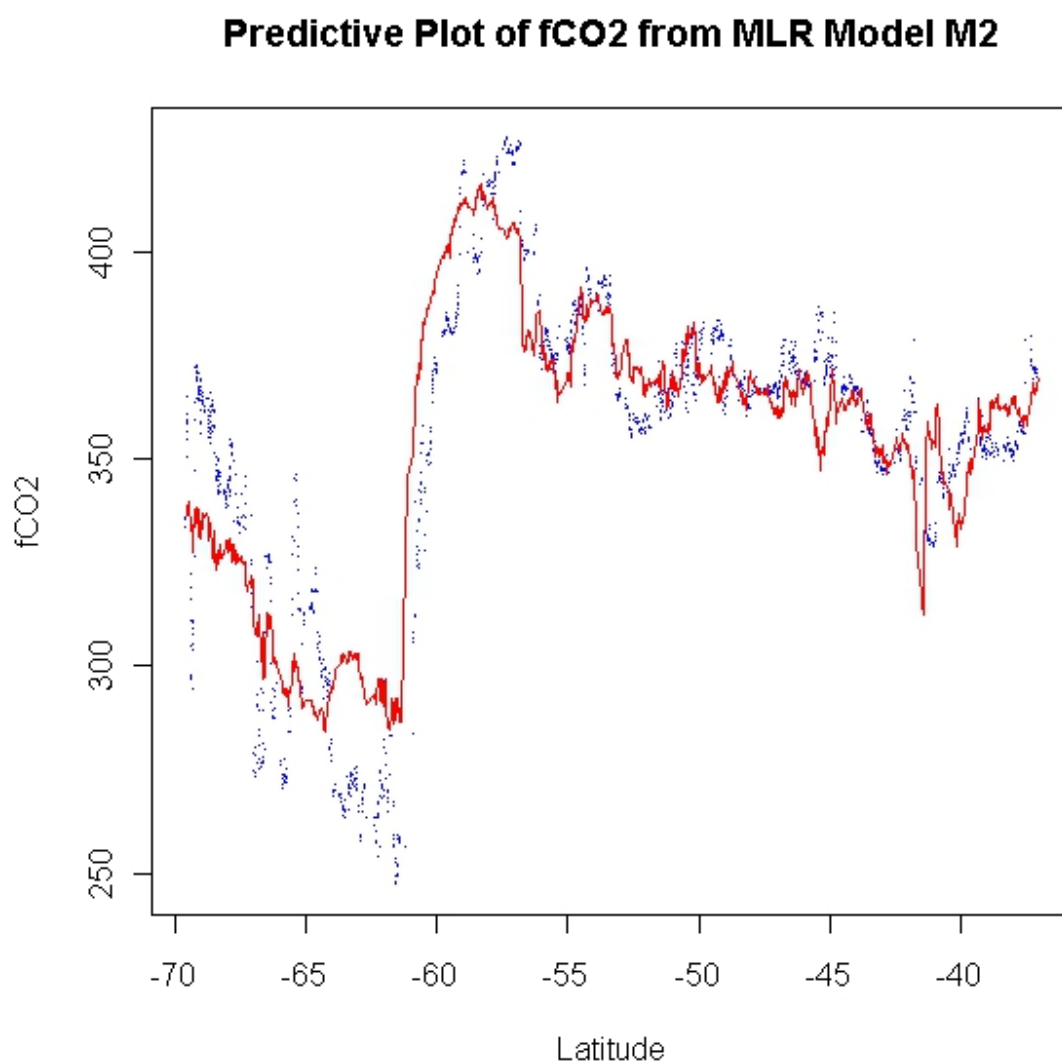


Figure 3.2: Multiple linear regression observed and predicted fCO₂ for model M2 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

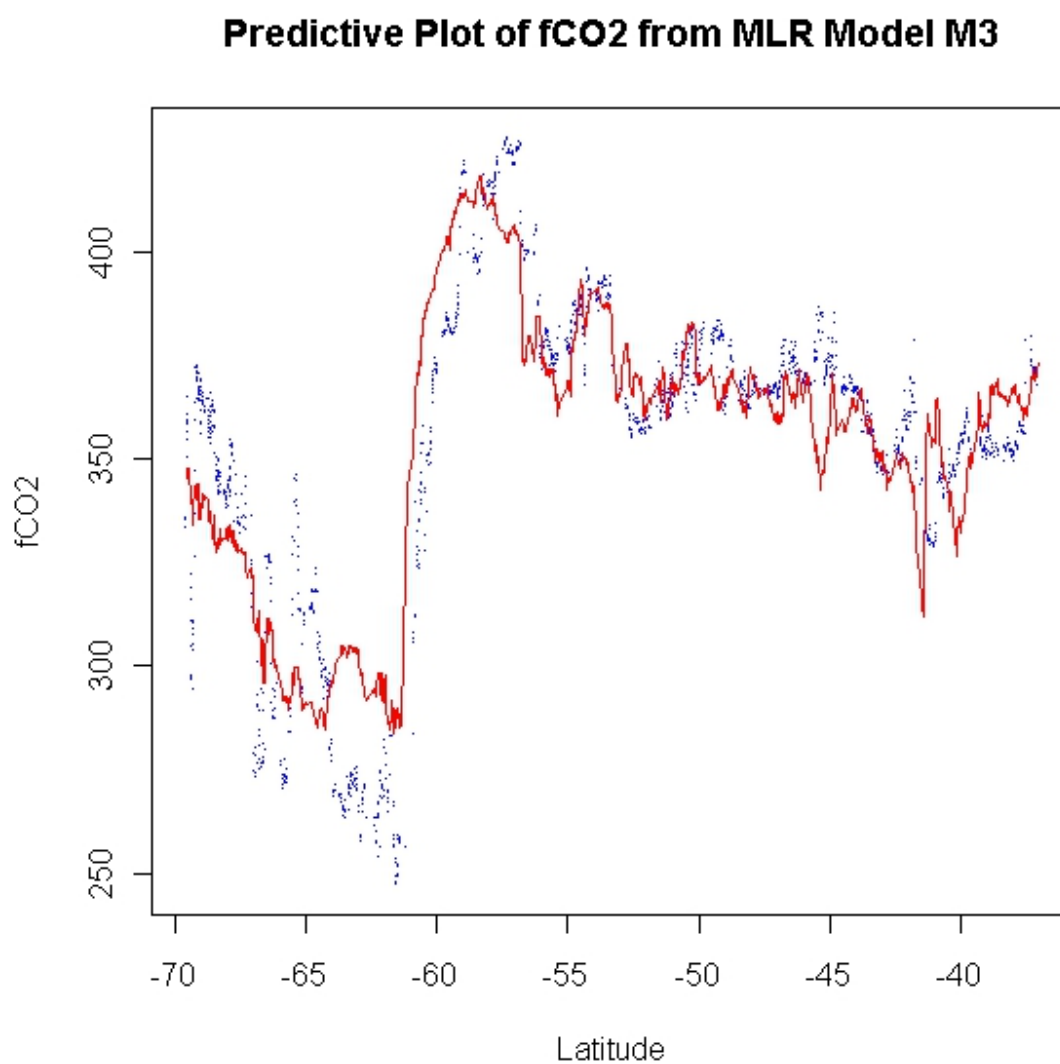


Figure 3.3: Multiple linear regression observed and predicted fCO₂ for model M3 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

As discussed previously, all three of these models were estimated using a random subset of the data set containing 70% of the data. The remaining 30% is used for assessment and is plotted in these figures. Figure 3.1 uses only SST, Log(chlorophyll-a) and MLD as predictor variables, while Figures 3.2 and 3.3 include latitude and salinity respectively in order to study the effect of their inclusions.

3.5.2.1 Training-Test Data Splits

As discussed in Section 3.4, models M1, M4, M5, M6, M7, M8 and M9 all make use of the same set of independent variables. The only difference between them being the random division of training and test data splits. Table 3.4 provides the error rates for these models. Columns 2, 3 and 4 present the MSE, MAE and RMSE respectively, while column 5 indicates the percentage division between the training and test data subsets.

Table 3.4: Multiple Linear Regression Subset Division Error Rates

Model	Mean Square Error	Mean Absolute Error	Root Mean Square Error	Training-Test Division
M1	328.789	14.150	18.133	70% - 30%
M4	332.922	14.322	18.246	80% - 20%
M5	331.883	14.152	18.218	60% - 40%
M6	330.823	14.115	18.189	50% - 50%
M7	330.030	14.051	18.167	40% - 60%
M8	329.130	14.034	18.142	30% - 70%
M9	330.390	13.957	18.177	20% - 80%

Table 3.4 quantifies the effect of the training data size on the models developed using the MLR method.

3.5.2.2 Standardised regression model to predict fCO₂

Along with the models discussed in Section 3.5.2.1, a standardised model, as discussed in Section 3.4, was also developed on the same 70% training data subset as was used for models M1, M2 and M3. The model was then assessed

using the two different standardisation versions of the remaining 30% of the entire SANAE49L6-final data set discussed in Section 3.4. Table 3.5 provides the MSE, MAE and RMSE for these two standardised test data subsets. As discussed in Section 3.4, the error rates for the model assessed on the test data set standardised using the training data means and standard deviations is denoted as M10_Training_Stats, while the error rates for the model assessed on the test data set standardised using the test data means and standard deviations is denoted as M10_Test_Stats.

Table 3.5: Standardised model error rates for Multiple Linear Regression Models

Model	Mean Square Error	Mean Absolute Error	Root Mean Square Error
M10_Training_Stats	612.697	20.186	24.753
M10_Test_Stats	618.224	20.170	24.864

3.5.2.3 Simulating the regression error

The final area of assessment for the MLR models involves simulating the division of test and training subsets in order to assess the variability of the error rates obtained from the models developed on the training data sets. This allows for comparison with further methods used in the later chapters of this thesis. Figures 3.4, 3.5 and 3.6 illustrate the spread of the test MSE, MAE and RMSE respectively for the MLR method of predicting fCO₂ using model M1 using histograms of 100 random divisions from which 100 MLR models are developed and assessed.

Table 3.6 provides information regarding the distribution of the error rates for the MLR models developed using 100 repetitions of random divisions into training and test subsets from the SANAE49L6-final data set. The table provides the mean, standard deviation, COV, minimum, median and maximum values of the MSE, MAE and RMSE from the 100 repetitions of subset divi-

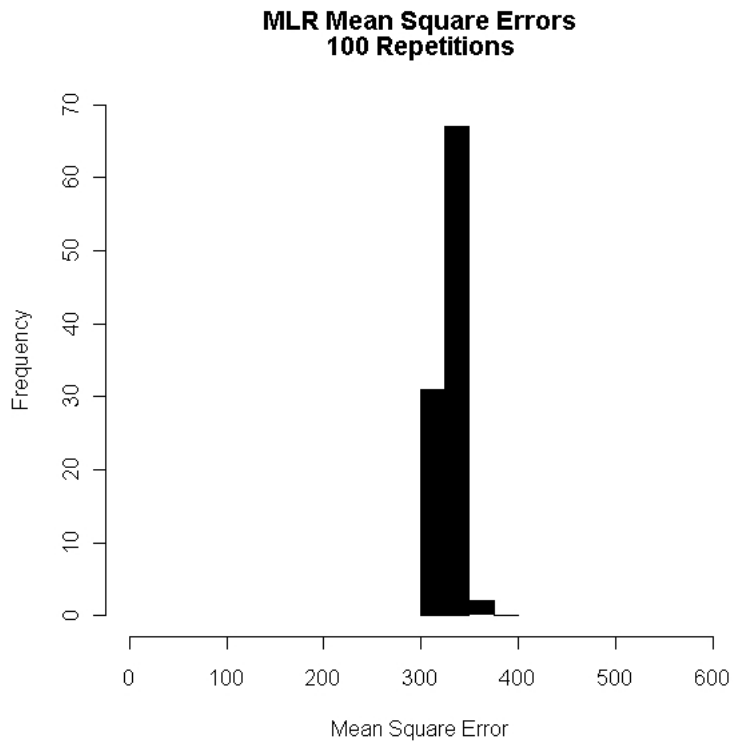


Figure 3.4: Histogram of 100 MLR model MSEs

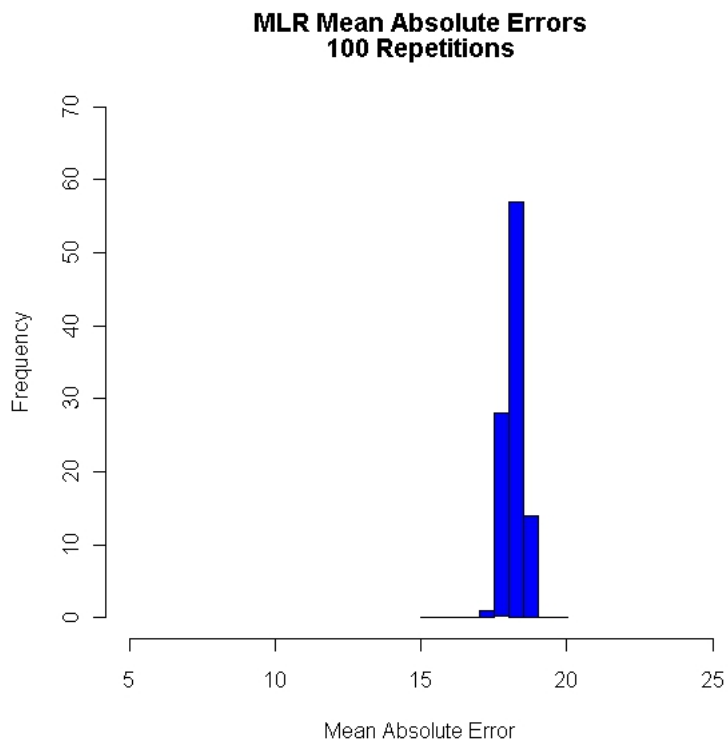


Figure 3.5: Histogram of 100 MLR model MAEs

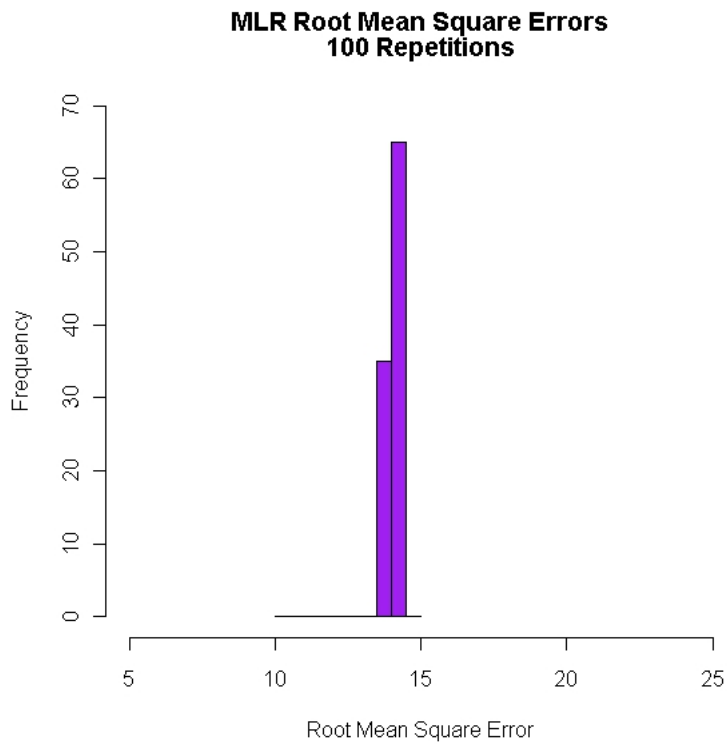


Figure 3.6: Histogram of 100 MLR model RMSEs

sions. This allows for the quantitative comparison of this method to further methods proposed.

Table 3.6: MLR error rate statistics for 100 subset divisions

	Mean Square Error	Mean Absolute Error	Root Mean Square Error
Mean	330.590	18.180	14.089
Standard Deviation	10.088	0.278	0.215
Coefficient of Variation	0.031	0.015	0.015
Minimum	305.703	17.484	13.596
Median	331.795	18.215	14.086
Maximum	350.631	18.725	14.461

3.6 Discussion of the linear regression results

This section provides a discussion of the results provided in Section 3.5. This allows for inference regarding the relationship between fCO₂ and the predictor

variables in the models as well as to draw conclusions regarding the ability of the MLR approach to describe this relationship and predict an unseen, independent test data set. This allows for an assessment into how well the model may generalise to unsampled parts of the SO.

3.6.1 Model parameter interpretation

The parameter estimates in Table 3.2 provide insight into the relationship identified by the MLR model to exist between the independent variable in question and the response variable, in this case fCO₂. The intercept parameter estimate provides an estimate for the fCO₂ when all the other independent variables observed at that point are 0. The intercept estimates, however, do not provide much insight into the relationship between fCO₂ and any of the independent variables. The other regression parameter estimates (namely SST, Log Chlorophyll, MLD, Salinity and Latitude), however, provide an estimate of the marginal linear relationship between each independent variable and the response. The negative parameter estimates observed for SST in each of the 10 models suggests a inverse linear relationship between fCO₂ and the SST. This implies that for each degree the SST increases (all other independent variables remaining constant), the fCO₂ estimates will decrease by between 2.7 μ atm and 4.6 μ atm as indicated in Table 3.2. For the Log chlorophyll, the parameter estimates seem to have a much larger absolute magnitude which implies that a unit increase in the Log chlorophyll concentration) would result in a decrease of between 89.1 μ atm and 93.3 μ atm in the fCO₂ as shown in Table 3.2. The MLD, on the other hand, produces a positive parameter estimate which indicates a direct relationship between fCO₂ and the MLD. Specifically, this implies that for every one meter the MLD increases, the fCO₂ estimate from the MLR model increases by between 0.18 μ atm and 0.29 μ atm. The parameter estimate for latitude (in model M2) indicates a direct relationship to fCO₂ such that for each degree further South travelled, the fCO₂ estimate

will decrease by $1.08\mu\text{atm}$ since degrees South are represented by negative values in SANAE49L6-final. Finally, the parameter estimate for Salinity (in model M3) indicate an inverse relationship with $f\text{CO}_2$ where for each unit increase in salinity, the $f\text{CO}_2$ estimate will decrease by $6.935\mu\text{atm}$.

3.6.2 Discussion of error statistics

This section discusses the predictive errors of models M1–M3 to identify any benefit provided by including salinity or latitude into the model and therefore what the effect will be, in an MLR model, of not being able to use these variables to extrapolate to the area beyond the SANAE49L6 path.

Table 3.3 provides the error rates (MSE, MAE and RMSE) of models M1–M3. There is a decrease in the model test error rates when latitude and salinity are included in the model individually. Model M2 produces a decrease in MSE of approximately $11.506\mu\text{atm}$, while the MAE decreases by approximately $0.519\mu\text{atm}$. This indicates that by including the latitude position of the measurements into the MLR model, the MSE and MAE rates are reduced by 3.5% and 3.67% respectively. This decrease indicates that the use of latitude in the models does have an effect. Model M3 indicates a much smaller decrease in the MSE and MAE compared to model M1. The test MSE and MAE obtained presents a decrease from model M1 of $3.664\mu\text{atm}$ and $0.247\mu\text{atm}$ respectively. This is a reduction of approximately 1.11% and 1.75% in MSE and MAE respectively. This suggests that the inclusion of salinity as an independent variable in the MLR models produces an improved MLR model.

The error rates presented in Table 3.3 seem to suggest that the use of the MLR model in the SO is not as accurate as was seen in the paper of Jamet *et al.* of (2007) which predicted $p\text{CO}_2$ in the North Atlantic for the Summer period of 1994–1995 using 3 MLR models. The first model made use of only SST as an independent variable, while the second model made use of SST as well as the latitude and longitude measurement. The final model replaced latitude and

longitude with the log chlorophyll-a concentration and MLD. The MLR model based on the data collected in the summer months (predicting pCO₂) produced MSEs of 203.633 μ atm, 178.49 μ atm and 130.874 μ atm respectively. These are lower than all 3 models discussed thus far in this thesis and therefore suggest that the MLR model does not perform as well in the SO.

Figures 3.1, 3.2 and 3.3 present a graphical representation of the predictive abilities of models M1, M2 and M3. The plots indicate that, although all three models seem to capture the general shape of the observed fCO₂ values, the models seem to not accurately capture the fCO₂ values south of 57°. The plots seem to indicate that all three models struggle in this regard, however model M2 does seem to provide a closer fit to the test fCO₂ values than the other two, as was also suggested by the error rates in Table 3.3. This conforms to the fact that this area is known to be an area of deep water up-welling of CO₂ and it is therefore imperative that the model be able to accurately predict the surface water fCO₂ in these areas. The MLR model does not seem to do this, but the positive aspect that can be taken from these figures is that the MLR models do seem to capture the general form of the *in situ* data.

3.6.2.1 Training-Test Data splits

In order to better understand what effect the division of the data set into training and test data subsets has on the predictive ability of the model. This will provide an indication as to how the model reacts to a diminishing amount of data in the training subset and how this affects the ability of the model to predict the response in “unseen” test data.

In Table 3.4 the error rates for models M1, M4, M5, M6, M7, M8 and M9 show that the size of the training data subset seems to have little effect on the prediction ability of the model. The difference in MSE between the model containing the most data points in the training subset (model M4) and the model with the least (model M9) is 2.532 μ atm or an approximate 0.76%

decrease in MSE. The MAE decrease between models M4 and M9 is $0.365\mu\text{atm}$ or 2.55%. These decreases are small and suggest that the size of the training data subset has little effect on the predictive ability of the MLR model.

The error rates do not seem to vary much with the decreasing size of the training data subset. The MSEs for all the subset divisions remain above $300\mu\text{atm}$. The same lack of trend can be seen in the MAEs and RMSEs where the errors remain slightly less than $15\mu\text{atm}$ and close to $18\mu\text{atm}$ respectively. These errors are, however, all still much larger than those seen in the analysis of Jamet *et al.* (2007).

3.6.2.2 Standardised regression models

The MLR models M1 to M9 all are non-standardised models fit with an intercept. Jamet *et al.* (2007), however suggest a model developed on variables standardised to have a common mean and standard deviation without an intercept, and thus model M10 was developed using this approach. Table 3.5 provides the error rates for this model tested on the two standardised test data sets discussed in Section 3.4.

The error rates obtained from the standardised models indicate a 86.4% increase in MSE and a 42.7% increase in MAE from model M1 for the model tested on the subset standardised using the training subset means and standard deviations and an 88% increase in MSE and 42.5% increase in MAE from model M1 for the model standardised using the test subset means and standard deviations. These results seem to suggest that a model developed on standardised training data set does not have the same predictive ability as the models developed on the original data as was the case in model M1. These results warrant further investigation to determine exactly why the observed error rates were so poor. This is, however, outside the scope of this thesis and will not be discussed.

3.6.2.3 Simulating the regression error

The model simulation results displayed in Figures 3.4, 3.5 and 3.6 indicate the advantages and disadvantages of the MLR method for predicting fCO₂ using *in situ* measurements of SST, log chlorophyll-a concentration and MLD. The 100 repetitions of the MSE of model M1 are displayed in Figure 3.4 and indicate how closely the MSE values are concentrated around the mean value of 330.590 μatm obtained from Table 3.6. This is further observed from the standard deviation and COV. The standard deviation of the MSEs seems high at 10.088 μatm , however when considering the COV for the MSEs (0.031) it is seen that this standard deviation represents only 3.1% of the mean fCO₂ value observed. The range of the 100 MSEs is only 44.928 μatm , which is only 13.6% of the average observed MSE. This further provides evidence of the low variability of the error rates when using the MLR models. The MAE and RMSE produced similar results with co-efficient of variations of 1.5% for both, while the ranges of these observed error rates were 1.241 μatm and 0.865 μatm respectively. These represent less than 7% of the mean for the respective observed error rates.

The disadvantage of the MLR models is the bias in the predicted fCO₂ values with respect to the observed values. The average RMSE is almost 3 μatm more than the RMSE observed by Jamet *et al.* (2007) in the North Atlantic Ocean.

3.7 Summary

The MLR approach in the SO (compared to the North Atlantic MLR model by Jamet *et al.* (2007)) produces error rates which do not satisfy the objectives discussed in Section 1.2.1 which aims at reducing uncertainty in the estimates to within 10% of the average CO₂ concentration. The bias in the model is large due to the linear functional form of the MLR model, which is unable to fully

capture the complex structure of the data, however, the model does seem to capture the general shape of the data as displayed in Figures 3.1, 3.2 and 3.3. The MLR models are robust towards diminishing amounts of data in the training data set as evidenced from Table 3.4 and also produce very consistent error rates for differing subset divisions of the same size. The standardised model, however, did not perform as well as the regular models. A method which allows the data to define its own distribution rather than imposing a normal distribution (as is done in MLR) may provide better (more accurate) results. On its own, therefore, the MLR approach to model fCO₂ is not satisfactory and therefore a more flexible approach needs to be investigated.

Chapter 4

Non-parametric Kernel Regression

4.1 Introduction

In this chapter we first begin with a brief review of a previously used non-parametric approach, specifically neural networks, to model the concentration of CO₂ in ocean waters. In Chapter 3, we provided results and a discussion surrounding a parametric approach to this problem, which has been used in various previous studies (Bates *et al.*, 2006; McNeil *et al.*, 2001; Slansky *et al.*, 1997; Wallace, 1995). The parametric approach, however, makes strong assumptions about the form of the regression function and therefore, in this chapter, a non-parametric approach that allows the data to define the form of the regression function is proposed. This non-parametric kernel regression (NPKR) approach is discussed and results from the SANAE49L6-final data set (as discussed in Section 2.4) is compared with the MLR models developed in Chapter 3. The non-parametric approach is used in an attempt to improve on the predictions of the MLR models. The objective of this chapter is thus to identify whether the NPKR method can improve on the predictive ability of fCO₂ from *in situ* data using remotely available independent variables.

4.2 Review on non-parametric research of CO₂ data

The relationship of fCO₂ to the predictor variables in the SO discussed in Chapter 3 is more complex than can be captured by a simple linear function, as was seen in the initial analyses of Section 2.5.3. The relationship may, in fact, be non-linear and therefore a MLR model fails to capture this relationship accurately. A less restrictive model is thus required in order to predict fCO₂ better.

A neural network (NN) is a method, which is specifically useful in obtaining generalisable estimates of responses, because of its ability to identify and exploit complex relationships. Also, the NN models do not require to be expressed in terms of an explicit, predefined function. In this way NNs are especially useful in estimating relationships, which are typically non-linear and need to be defined empirically. This is, of course, dependent on the availability of enough data for the models to be trained (Lefèvre *et al.*, 2005). Within the NN subset of models, the self organising map (SOM), introduced by Kohonen (2001), have received the most focus from a geosciences point of view due to their ability to identify and make use of relationships (not necessarily linear relationships) between predictor variables (Kohonen, 2001). These SOM techniques only use the observations to obtain the estimated model using an unsupervised learning algorithm (Telszewski *et al.*, 2009). Models suggested by Lefèvre *et al.* (2005) and Friedrich and Oeschies (2009) used the spatial position and time of the measurements in the SOM approach. This, however, provided relatively artificial and unrealistic results, as the use of spatial position (latitude and longitude) created locally clustered values of similar pCO₂ values, where this type of large clustering was not guaranteed to occur, while including the time of the measurements increased the effect of seasonality on the estimated pCO₂ maps. As such, these results may be applicable and ac-

curate for small areas or regions in the ocean, but cannot be used for large stretches (Telszewski *et al.*, 2009). The application of the SOM is a three step algorithm: First, an unsupervised analysis is performed (i.e. excluding the pCO₂); second, the observed (*in situ*) pCO₂ values are then used to label the neurons of the predictor variables; and third, the trained neurons, labelled with the *in situ* pCO₂, assign pCO₂ values to all the geographical grid points on the map (Telszewski *et al.*, 2009). The SOM is regarded as a “black box” type of method, which produces difficulties in explaining and interpreting the method and the resulting models.

4.3 Using non-parametric kernel methods for predicting fCO₂

In the SO, the interaction between fCO₂ and its driving factors is complex, because in certain regions the main drivers can be different (Telszewski *et al.*, 2009). This is particularly evident from the exploratory plots given in Figure 2.6, which indicate that certain predictors, such as chlorophyll-a concentration, show different patterns in certain areas of the ocean. This may indicate that the relationships between oceanic fCO₂ and the factors affecting it are complex and non-linear. The reasoning for this spatial variability of the relationship may be due to the presence of bio-geochemical provinces in the SO, such as the polar front. These are known areas of the ocean (although their spatial position varies according to the season), which display different characteristics with regards to the physical and bio-geochemical factors and may affect the oceanic fCO₂ levels differently. More information on the location of the Antarctic polar front can be obtained from the papers of Moore *et al.* (1999) and Dong *et al.* (2006).

The focus of this thesis is to propose an effective method to predict fCO₂ in terms of an input vector of predictor variables. The method proposed is

a non-parametric kernel regression (NPKR) approach due to its ability to identify and incorporate non-linear relationships in complex data structures, which may follow no known parametric function.

A non-parametric model does not specify a functional form for the regression function applied, but instead allows the data to determine a non-specific form. This is in contrast to parametric methods such as the MLR method, which specifically defines the regression function (and therefore the conditional response) as being a linear function of the explanatory variables in the data. This strict specification of the regression function limits the model, especially when dealing with relationships that are typically non-linear. The unspecified regression function approach taken by the NPKR model allows the method to identify interactions and relationships in the data that would otherwise not be recognised or incorporated into a parametric model (Fox, 2005). This ability of the non-parametric model makes it an attractive candidate for the task at hand since the relationship between $f\text{CO}_2$ and its physical and bio-geochemical drivers in the ocean is typically non-linear. Non-parametric regression techniques also have a wide spectrum of options available and, while this thesis focuses on the application of NPKR, other approaches such as local polynomial regression and regression splines are available for use within the non-parametric pool.

The non-parametric regression approach, however, is not without its disadvantages. One disadvantage that is often criticised relates to the so called problem of the curse of dimensionality. Methods that are heavily dependent on data in the local neighbourhood require large data sets in order to produce accurate and reliable estimates. In data sets where there are a large number of input variables, this is particularly true, since the number of observations required to produce estimates as accurate as (or more accurate than) a parametric model increases exponentially as the number of predictor variables in the data set increases. The reason for this increase in data required

is due to the decreasing amount of observations occurring in a fixed neighbourhood width around each possible target input vector (Fox, 2005). In the SANAE49L6-final data set, however, the curse of dimensionality does not play a large role due to the large number of observations and the small number of independent variables.

4.3.1 Theoretical overview of non-parametric kernel regression

As discussed in Chapter 3, a regression model is concerned with the estimation of the average (conditional average) of a real valued response variable y_i given a set of continuous and/or discrete input (predictor) variables \underline{x}_i (where $\underline{x}' = (x_1, x_2, \dots, x_p)$). This estimate is given as a function of these predictor variables and will be denoted as $g(x_1, x_2, \dots, x_p)$ such that

$$\mu_y | x_1, x_2, \dots, x_p = g(x_1, x_2, \dots, x_p) \quad (4.3.1)$$

The only limiting assumption made in the NPKR approach is that the function that relates the mean conditional response to the input variables should be smooth and should exist (implying that there must be some sort of relationship between the conditional response and the explanatory variables) (Fox, 2005; Racine and Li, 2004). This less stringent assumption, however, is not without a cost. Specifically it is heavily data dependent and computationally intensive (especially for large data sets), and does not have the same, easily interpretable output that regular parametric models, such as MLR, may have. In spite of its limitations, the advantage of NPKR is its flexibility and potentially more accurate predictability compared to the parametric approach, and hence provides a strong case for its use in this thesis.

The NPKR method allows the data itself to define a model to summarise the information effectively and accurately. Intuitively, the process underlying NPKR is simple and provides a prediction based on a method which is un-

derstandable. The regression function $g(\underline{x})$ is simply defined as an empirically weighted average of the observed responses which correspond to observed input vectors falling within a defined “close” neighbourhood of the target input vector \underline{x} . The defined neighbourhood is determined by the smoothing parameters (bandwidths) denoted by $\underline{h} = (h_1, h_2, \dots, h_p)'$. Larger values of the individual h_i for $i = 1, 2, \dots, p$ provides a smoother (and therefore less variable) function $g(\underline{x})$. This also increases the bias of the function and can therefore result in an underfit model with a potentially large MSE. Larger bandwidths also create a model which is no longer local with regards to that variable, but rather global, which defeats the purpose of a local regression model. In contrast, smaller bandwidth values provide a less biased, but more variable function for the training data. This can provide an overfit of the model, which may have a very small error rate in the training data set, but may not generalise well to “unseen” data and therefore may not be useful outside of the training data.

In the notations of Racine and Li (2004), let Y_i denote a univariate response, and let the p -dimensional input vector be denoted by $\underline{X}_i' = (X_{i1}, X_{i2}, \dots, X_{ip})$, $i = 1, 2, \dots, n$. The NPKR model is defined as follows

$$Y_i = g(\underline{X}_i) + \epsilon_i. \quad (4.3.2)$$

The independent error terms (ϵ_i) are assumed to be normally distributed with a mean of 0 and a constant variance, which is denoted by σ^2 ¹. Furthermore, the NPKR approach does not explicitly define the functional form of $g(\underline{X}_i)$. Let $f(\underline{x})$ denote the joint density function of the input variables and let $m(y, \underline{x})$ denote the joint density function of the response y and the input variables $\underline{x} = (x_1, x_2, \dots, x_p)'$. These multivariate density functions are estimated using product kernels, by averaging the n product kernels to obtain the estimate as

$$\hat{f}(\underline{x}) = \frac{1}{n} \sum_{i=1}^n K_{h,i}(\underline{x}), \quad (4.3.3)$$

¹This assumption of normality and constant variance of the error terms is also made in the MLR model.

where the product kernel $K_{h,i}(\underline{x})$ is simply the product of p univariate kernels such that

$$K_{h,i}(\underline{x}) = \frac{1}{h_1 h_2 \dots h_p} \prod_{j=1}^p k\left(\frac{X_{ij} - x_j}{h_j}\right). \quad (4.3.4)$$

In (4.3.4), $k(\cdot)$ denotes an appropriate univariate kernel function, which should be symmetric and a decreasing function of the distance from the target input vector \underline{x} (i.e. decreasing as $\|\underline{X}_i - \underline{x}\|$ increases) (Racine and Li, 2004). The joint density function of the response and the input variables is similarly defined by the average of the product kernels and is given by

$$\hat{m}(y, \underline{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_y} k\left(\frac{Y_i - y}{h_y}\right) K_{h,i}(\underline{x}). \quad (4.3.5)$$

Using these two estimates of the joint density functions, we can now estimate the conditional expected value $g(\underline{x}) = E[Y_i | \underline{X}_i = \underline{x}]$ as

$$\begin{aligned} \hat{g}(\underline{x}) &= \int y \hat{f}(y | \underline{X} = \underline{x}) dy \\ &= \frac{\int y \hat{m}(y, \underline{x}) dy}{\hat{f}(\underline{x})}. \end{aligned} \quad (4.3.6)$$

By substituting the empirical estimates of the joint density functions (4.3.3) and (4.3.5) into (4.3.6) the estimate of the regression function in terms of the product kernels $K_{h,i}(\underline{x})$ is obtained as

$$\hat{g}(\underline{x}) = \frac{\int y \frac{1}{n} \sum_{i=1}^n \frac{1}{h_y} k\left(\frac{Y_i - y}{h_y}\right) K_{h,i}(\underline{x}) dy}{\frac{1}{n} \sum_{i=1}^n K_{h,i}(\underline{x})}. \quad (4.3.7)$$

Making the variable change to $v = \frac{Y_i - y}{h_y}$ and noting that $dv = \frac{1}{h_y} dy$ it follows that (4.3.7) can be written as

$$\begin{aligned} \hat{g}(\underline{x}) &= \frac{\frac{1}{n} \sum_{i=1}^n \int (Y_i - v h_y) k(v) K_{h,i}(\underline{x}) dv}{\frac{1}{n} \sum_{i=1}^n K_{h,i}(\underline{x})} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n [Y_i K_{h,i}(\underline{x}) \int k(v) dv - h_y K_{h,i}(\underline{x}) \int v k(v) dv]}{\frac{1}{n} \sum_{i=1}^n K_{h,i}(\underline{x})}. \end{aligned} \quad (4.3.8)$$

Since $\int k(v) dv = 1$ and $\int v k(v) dv = 0$, (4.3.8) simplifies to

$$\hat{g}(\underline{x}) = \frac{\sum_{i=1}^n Y_i K_{h,i}(\underline{x})}{\sum_{i=1}^n K_{h,i}(\underline{x})}. \quad (4.3.9)$$

An intuitive interpretation of (4.3.9) is that it is a weighted average of the observed response values, which correspond to sets of input variables surrounding the target input vector \underline{x} . The weights in this average are defined by the n product kernels $K_{h,i}(\underline{x})$, each consisting of functions that are symmetric and decreasing with the distance between the target input vector \underline{x} and the observed input vector \underline{X}_i (Racine and Li, 2004). For certain choices of the kernel function $k(\cdot)$, these product kernel weights will be 0, such as the Epanechnikov kernel, in which if the observed input variables falls outside the bandwidth of that specific target input variable, the weight applied is 0.

4.3.2 Specifying the kernel and bandwidth optimisation

When developing a NPKR model, there are two main components of the model that must be determined, namely the bandwidths for each of the independent variables and the kernel function used. A very important and largely criticised area of the non-parametric regression procedure is the choice of an optimal bandwidth. In essence the bandwidth selection procedure is a trade-off between bias and variance, since larger bandwidth values will provide less variable predictions for the model, but may produce a larger bias in terms of prediction error due to observations further away from the target, test observation being included in the neighbourhood; while smaller bandwidths will produce less biased models (since only observations very close to the test observation will be included in the neighbourhood), but will not only allow for more variable predictions, but may also worsen the curse of dimensionality.

In this chapter, we use the R function `npregbw` to determine optimal bandwidths, while the function `npreg` is used to fit the NPKR model to the data and to obtain fit statistics and predictions of the responses in the test data set. These functions (found in the R package `np`) make use of leave-one-out cross-validation (CV) in order to determine the optimal value of the bandwidths for each of the variables in the model (Racine and Li, 2004; R Development Core

Team, 2011; Li and Racine, 2004). The leave-one-out estimate of the joint density function of the input variables using the kernel estimate is defined as

$$\hat{f}_{-i}(\underline{X}_i) = \frac{1}{n} \sum_{j \neq i} K_{h,j}(\underline{X}_i) \quad (4.3.10)$$

and is used to obtain an estimated response for the i^{th} input vector using the non-parametric kernel estimate, which is based on a data set which omits the i^{th} observation. The i^{th} leave-one-out estimated response is denoted by $\hat{g}_{-i}(\underline{X}_i)$ and is written as

$$\begin{aligned} \hat{g}_{-i}(\underline{X}_i) &= \frac{\frac{1}{n} \sum_{j \neq i} Y_j K_{h,j}(\underline{X}_i)}{\hat{f}_{-i}(\underline{X}_i)} \\ &= \frac{\sum_{j \neq i} Y_j K_{h,j}(\underline{X}_i)}{\sum_{j \neq i} K_{h,j}(\underline{X}_i)}. \end{aligned} \quad (4.3.11)$$

This is again intuitively explainable as a weighted average of the responses corresponding to input vectors, which are close in input space to the target input vector for which a prediction of the response is required. To select optimal values of the bandwidths, a set of p bandwidth values are chosen to minimise the leave-one-out CV MSE, which is calculated as

$$MSE(CV) = \min_{h_1, h_2, \dots, h_p} \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{g}_{-i}(\underline{X}_i))^2 \right). \quad (4.3.12)$$

This minimisation is done by applying an iterative procedure to this function which, if the kernel function and training data set is specified, depends only on the bandwidths. The combination of optimal bandwidths is chosen to minimise the cross-validation MSE. This method for selecting an optimised set of bandwidths is extremely computationally intensive, specifically for large data sets, as in the case of this thesis, since the model has to be refitted at each observation for the data excluding that specific data point. The price paid through computational intensity is, however, balanced by the potential of developing a model that provides more accurate predictions. Further derivations, assumptions and explanations of the method and formulae used in this analysis can be found in papers by Racine and Li (2004) and Li and Racine

(2004) as well as in the book by Li and Racine (2007). Further information on the application of the method can be obtained from the help files in R for the *np* package (R Development Core Team, 2011).

Another challenge in the application of the NPKR method is the long computational time in the determination of the optimal bandwidths in the model. The length of the computational time is due to the “trial-and-error” nature of the bandwidth optimisation. The *npregbw* function continues selecting bandwidth combinations until some stopping rule is met also known as the “tolerance”. Once a set of bandwidths is found, the search is restarted and run again to ensure that the minimum CV MSE obtained is not simply a local minimum. Usually the number of resets (multi-starts) is determined as the minimum of the number of independent variables and 5. To reduce computational time, a two-stage approach is adopted in this thesis: first, a higher tolerance value is used to obtain rough estimates of the optimal bandwidth values; second, these rough estimates obtained, are entered as starting values for the search and the function is restarted with the tolerance values reset to their defaults.

The NPKR method does have certain limitations which can prevent it from providing an accurate and effective model for data. The first limitation would be its lack of ability to deal with missing values in the data. Unlike other statistical methods (such as regression trees), the function used to develop the NPKR models used in this thesis does not have an automatic method for dealing with missing observation points. Observations which contain a missing value in any of the variables are simply omitted from the data set before fitting the model. This reduces the size of the data set and, as discussed in Section 4.3.1, this may result in an inaccurate model for smaller data sets (as non-parametric models are very data dependent and therefore require as much data as possible in order to produce an accurate model).

A second limitation of the NPKR method is the difficulties experienced

in predicting (forecasting) responses for input vectors of explanatory variables which fall well outside the range of input vectors in the training data set. This is especially evident in the method used in this analysis of multivariate data due to the use of product kernels in determining the weights for the responses within the neighbourhood of the target input vector.

Given that the kernel function chosen in the analysis applies larger weights to those observed values closer to the target input vector \underline{x} and that the weights decrease symmetrically and smoothly around \underline{x} , the specific choice of kernel function used in the analysis is not critical (Fox, 2005). The *np* function in R allows for the use of either the uniform, Gaussian, or Epanechnikov kernel functions. The kernel function used in the analysis of the SANAE49L6-final data is the Epanechnikov kernel represented by

$$k(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1). \quad (4.3.13)$$

From the calculation of the kernel weight in the product kernel seen in (4.3.4), and from the Epanechnikov kernel (4.3.13), it is seen that all observed input vectors which have a single variable falling further than the bandwidths distance away from the target value for that variable obtain a weight of 0. The use of kernel functions such as the Epanechnikov kernel may further complicate matters as these kernel structures place a weight of 0 on all responses corresponding to input vectors, where one of the explanatory variables for in the training data set falls outside of the neighbourhood of the target explanatory variable (i.e. the observed explanatory variable in the training data set is more than 1 bandwidth distance above or below the desired value in the test data set for that specific variable). The uniform kernel is a square function defined by

$$k(u) = \frac{1}{2}I(|u| \leq 1), \quad (4.3.14)$$

which simply provides equal weights to all observations within a bandwidths distance from the target input vector, like a histogram. This is not an appropriate method since it does not take into account a decreasing reliance in

responses for input vectors further away from the target. The Gaussian kernel, on the other hand, relies too heavily on observations far away (in input space) from the target. The Gaussian kernel has the form

$$k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, \quad (4.3.15)$$

which is defined for all real values of u and therefore all observations are weighted for each prediction, with observations far away (in input space) having a very small weight. For large data sets, the calculation of these weights can be very time consuming with distant observations (from the target) playing a minute role in the prediction. For the analysis of SANAE49L6-final it is chosen to use the Epanechnikov kernel function as it is a smooth, decreasing function of the distance, in input space, from the target vector \underline{x} and it has cutoff points, after which the weighting function is 0. For forecasting, the Epanechnikov kernel may be a problem, but the focus of this analysis is to provide reliable measures of the fCO₂ in the SO where no *in situ* values are available using a model designed on areas where ship movement and *in situ* measurements of fCO₂ are readily available rather than to forecast future fCO₂ values.

Once the optimal bandwidths have been determined, estimates of the response using the kernel regression method are obtained as described above. The results and a discussion of the results are provided in the subsequent sections for this method being applied to the SANAE49L6-final data set and also indicates the models ability to generalise to “unseen” data sets.

4.4 Non-parametric results to predict fCO₂

The NPKR model, as discussed in Section 4.3.1, is fit to the SANAE49L6-final data set (as described in Section 2.4) and the results were then compared to those obtained from the multiple linear regression model. The NPKR model is developed using the functions *npregbw*, to optimise the bandwidths, and

npreg, to fit the model while using the Epanechnikov bandwidth function to provide the weights. Where applicable, identical divisions of training and test data sets were made to those used when fitting the MLR model in order to ensure that the results obtained were comparable. These results are provided in Sections 4.4.1 and 4.4.2 followed by a discussion of the results in Sections 4.5.1 and 4.5.2.

4.4.1 Optimising the non-parametric regression model

The NPKR modelling procedure focuses on initially determining optimal bandwidths for the model based on leave-one-out CV applied to the training data set as described in Section 4.3.1. Models M1 to M10 in Table 4.1 represent the same 10 models as discussed in Section 3.4. Table 4.1 provides the optimal bandwidths determined by the leave-one-out CV performed on the same training and test data subsets as were used for the MLR approach in the previous chapter. The optimal bandwidth estimates provided in Table 4.1 give an

Table 4.1: Non-parametric kernel regression bandwidth estimates

Model	SST	log Chlorophyll-a Concentration	MLD	Salinity	Latitude
M1	0.162	0.043	7.721		
M2	0.251	0.102	1.198		0.493
M3	0.165	0.102	3.389	0.070	
M4	0.162	0.043	7.721		
M5	0.172	0.043	7.721		
M6	0.172	0.043	7.721		
M7	0.197	0.021	7.721		
M8	0.172	0.043	6.120		
M9	0.209	0.015	7.209		
M10	0.032	0.017	0.091		

indication of the neighbourhood of each variable considered to be near to the value of that variable in the observation for which the predicted $f\text{CO}_2$ is required. Larger values of the bandwidth indicate a wider local neighbourhood for that variable and therefore a less local and more global neighbourhood.

The optimised bandwidth values indicated in Table 4.1 can only be positive values as a negative value has no meaning in defining a local neighbourhood. Since the NPKR method is non-parametric, the regression function cannot be written in a simple linear formula as was the case in Chapter 3.

4.4.2 Assessing the non-parametric regression model

The assessment of the NPKR models is again divided into 4 distinct aspects identical to those used in assessing the MLR models in Section 3.5. Initially models M1, M2 and M3 are assessed and compared to one another using the same 70% – 30% division of training and test data sets in order to identify the improvements (if any) in the predictive abilities of the models by including latitude or salinity into the model as independent variables. The second aspect assesses the effect of the size of the training data set on the predictive ability of the model. The third aspect assesses the approach of standardising the variables in the training data set to have a mean of 0 and standard deviation of 1 before developing the model, while the final aspect simulates the model M1 for varying training data subsets in order to identify the distribution of the model error rates for a fixed data set size.

The MSE, RMSE and MAE rates of the NPKR models M1, M2 and M3 are provided in Table 4.2. These error rates are all determined based on the test data subsets, considered to be “unseen” by the models developed based on the training data subsets.

Table 4.2: Non-parametric kernel regression model error rates

Model	Mean Square Error	Mean Absolute Error	Root Mean Square Error
M1	76.560	5.037	8.750
M2	38.091	3.600	6.172
M3	33.750	3.834	5.809

These errors provide a numerical representation of the predictive ability of

models M1, M2 and M3, as well as identifying the improvements in the models by the inclusion of latitude and salinity. Figures 4.1, 4.2 and 4.3 provide a graphical representation of the predictive ability of these models by plotting the observed and predicted values of $f\text{CO}_2$ from the test data set against their latitude measurements.

Figure 4.1 corresponds to model M1 that uses only SST, log chlorophyll and MLD as predictor variables to describe the distribution of $f\text{CO}_2$ in the SO. The models used to generate Figures 4.2 and 4.3 add latitude and salinity respectively as predictor variables to model M1 and provide important information as to how these variables may further improve the predictive ability of the model.

4.4.2.1 Training-Test Data Splits

As described in Table 3.4, models M1, M4, M5, M6, M7, M8 and M9 all represent models using the same set of independent variables to describe the distribution of $f\text{CO}_2$ in the SO. These models differ only in the divisions of the training and test subsets where the training subsets decrease from 80% of the total data set for model M4 to 20% of the total data set for model M9. Table 4.3 provides the MSE, MAE and RMSE for these models and indicates the respective percentage divisions of the entire data set placed in the training and test data subsets.

Table 4.3: Non-parametric kernel regression subset division error rates

Model	Mean Square Error	Mean Absolute Error	Root Mean Square Error	Training-Test Division
M1	76.560	5.037	8.750	70% - 30%
M4	69.238	4.972	8.321	80% - 20%
M5	80.053	5.073	8.947	60% - 40%
M6	78.268	5.084	8.847	50% - 50%
M7	193.333	4.928	13.904	40% - 60%
M8	102.999	5.107	10.149	30% - 70%
M9	277.424	4.991	16.656	20% - 80%

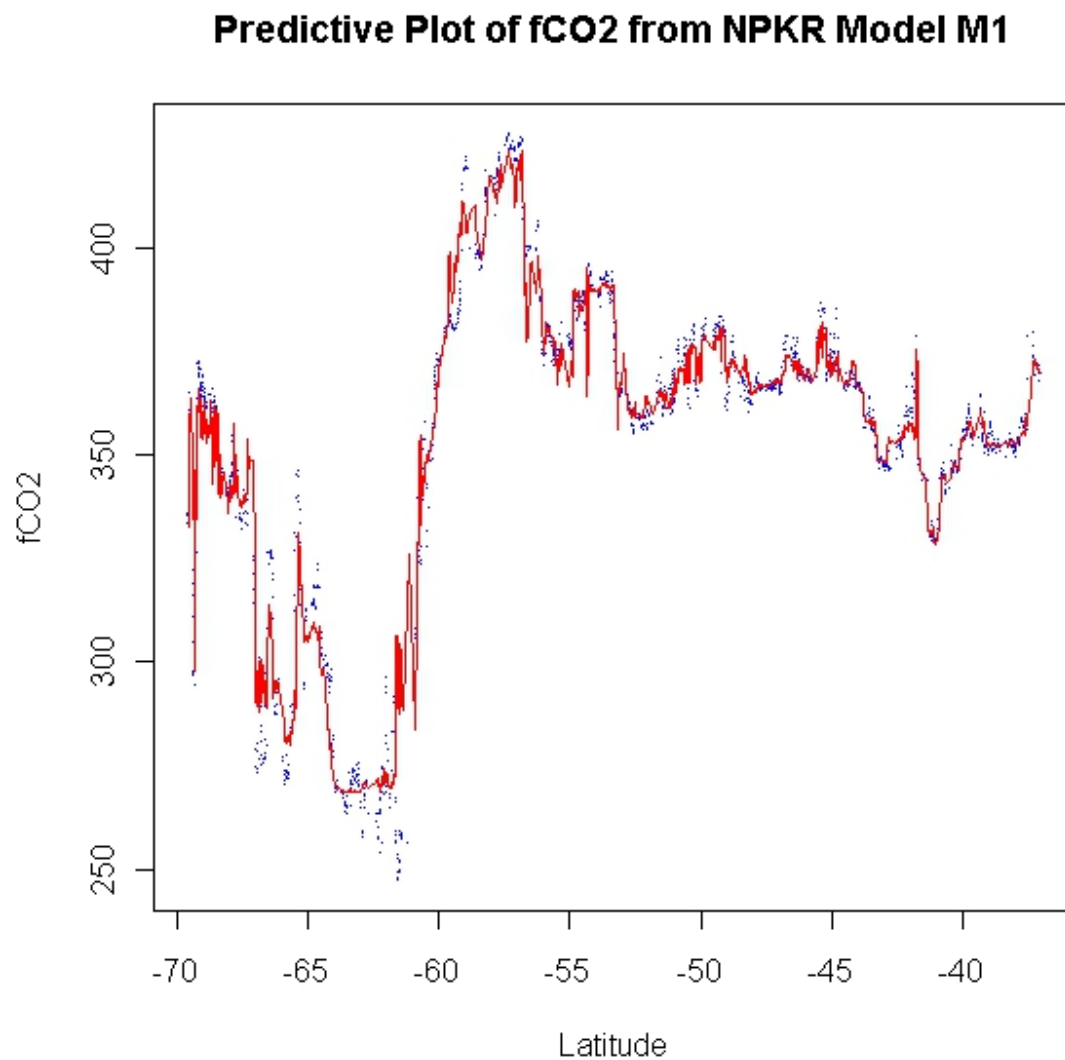


Figure 4.1: Non-parametric kernel regression observed and predicted fCO₂ for model M1 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

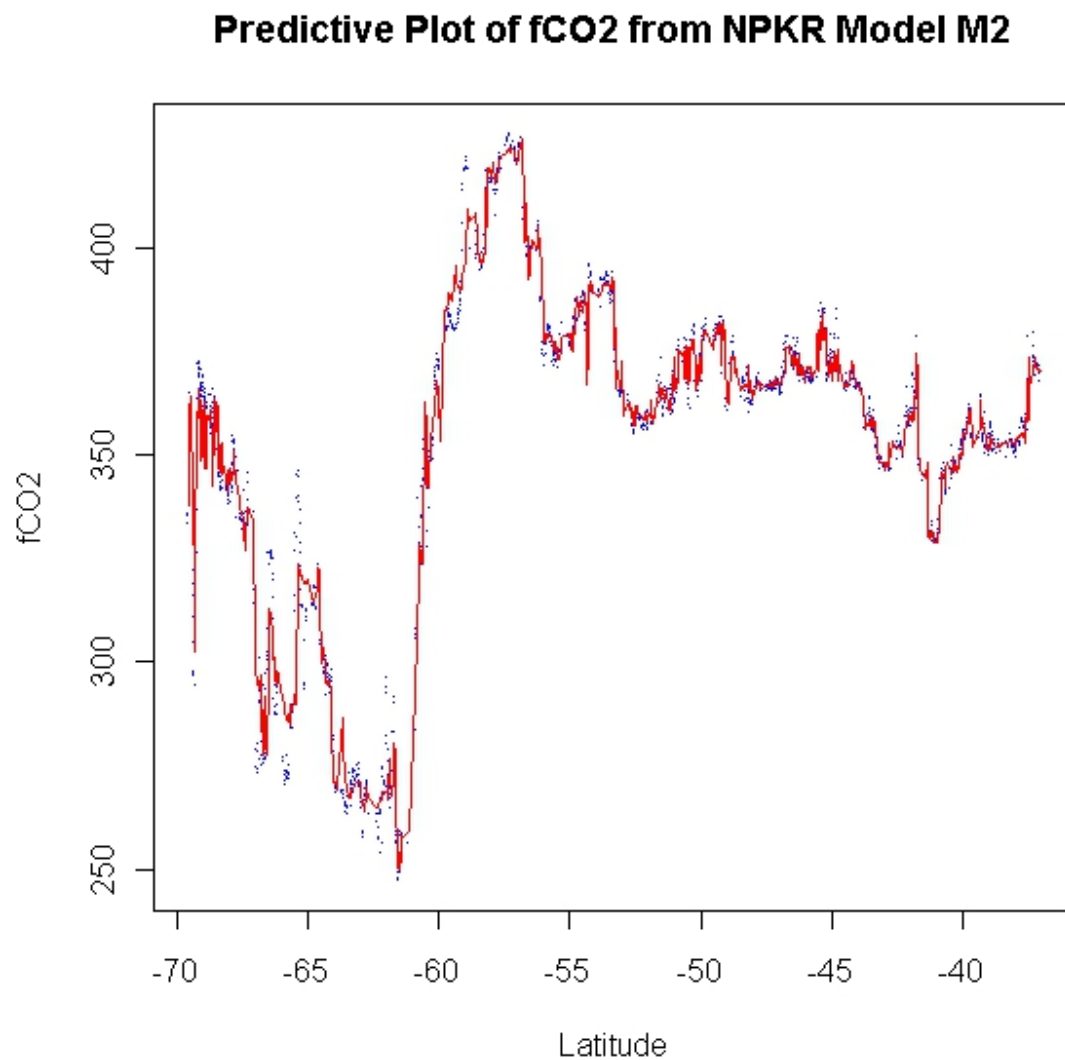


Figure 4.2: Non-parametric kernel regression observed and predicted fCO₂ for model M2 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

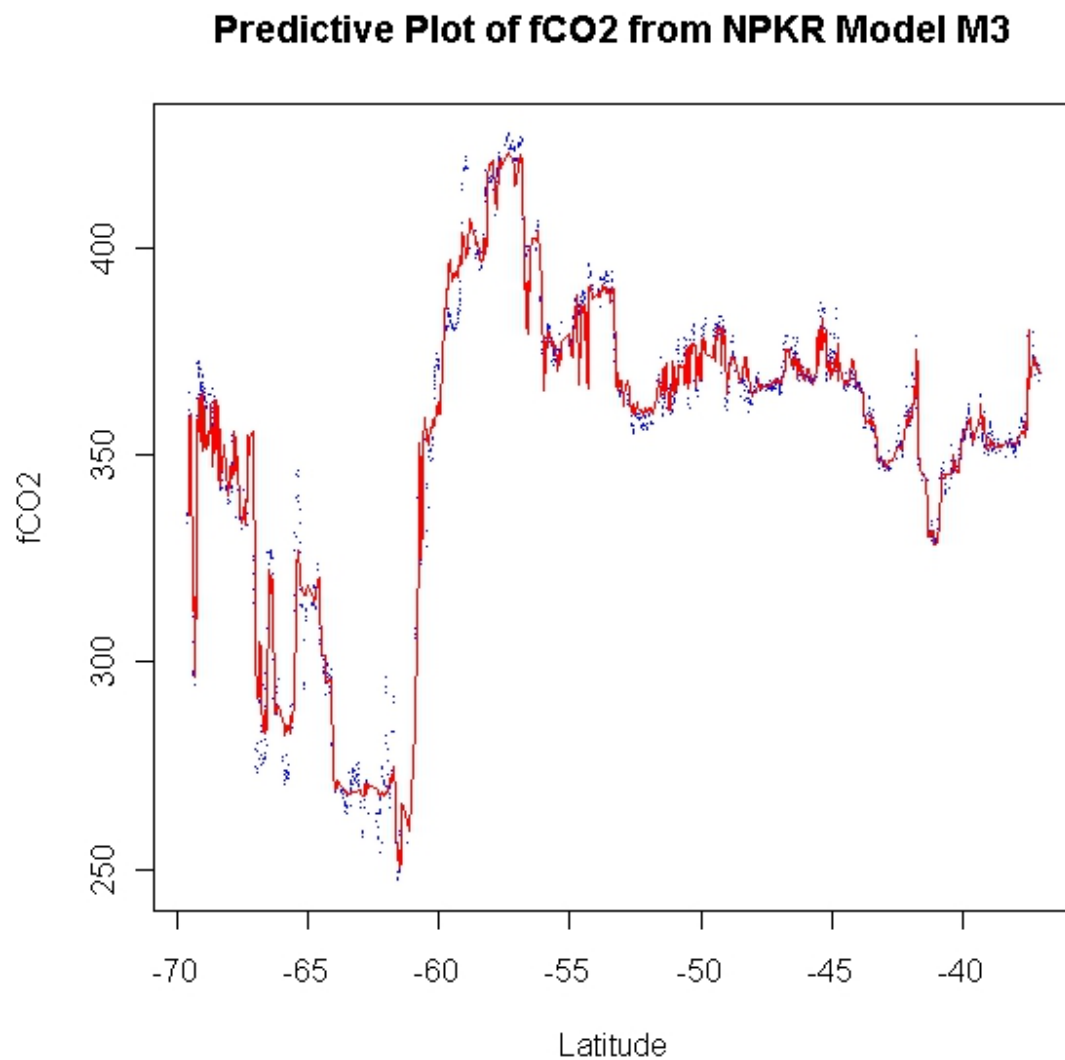


Figure 4.3: Non-parametric kernel regression observed and predicted fCO₂ for model M3 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

Figures 4.4 and 4.5 present a plot of the observed and predicted values of fCO_2 over latitude for models M8 and M9 which both are identical to model M1 in terms of the independent variables, however these model are developed and assessed on a 30%-70% and 20%-80% division between training and test data subsets respectively. This provides a visual representation as to why the MSE and RMSE values increase as much as they do in Table 4.3, while the MAE values seem to remain relatively constant.

4.4.2.2 Standardised non-parametric regression models

As discussed in Section 3.3 and as was performed using the MLR method, an NPKR model is also developed on a standardised version of the 70% training data subset used in the development of models M1, M2 and M3. The developed model is once again assessed using two separate standardised test data sets derived from the same 30% of the original SANAE49L6-final data set used to assess models M1, M2 and M3. These standardised sets are discussed in Section 3.3. Table 4.4 provides the MSE, MAE and RMSEs of this model when assessed on these standardised test data sets.

Table 4.4: Standardised model error rates for non-parametric kernel regression models

Model	Mean Square Error	Mean Absolute Error	Root Mean Square Error
M10_Training_Stats)	1494.527	27.500	38.659
M10_Test_Stats)	1483.087	27.723	38.511

4.4.2.3 Simulating the non-parametric regression error

Finally, the training and test data subset division used for developing and assessing model M1 is performed 100 times in order to obtain a sample of 100 training and test data subsets, each of which is used to develop a NPKR model using the same independent variables as model M1 (namely SST, log

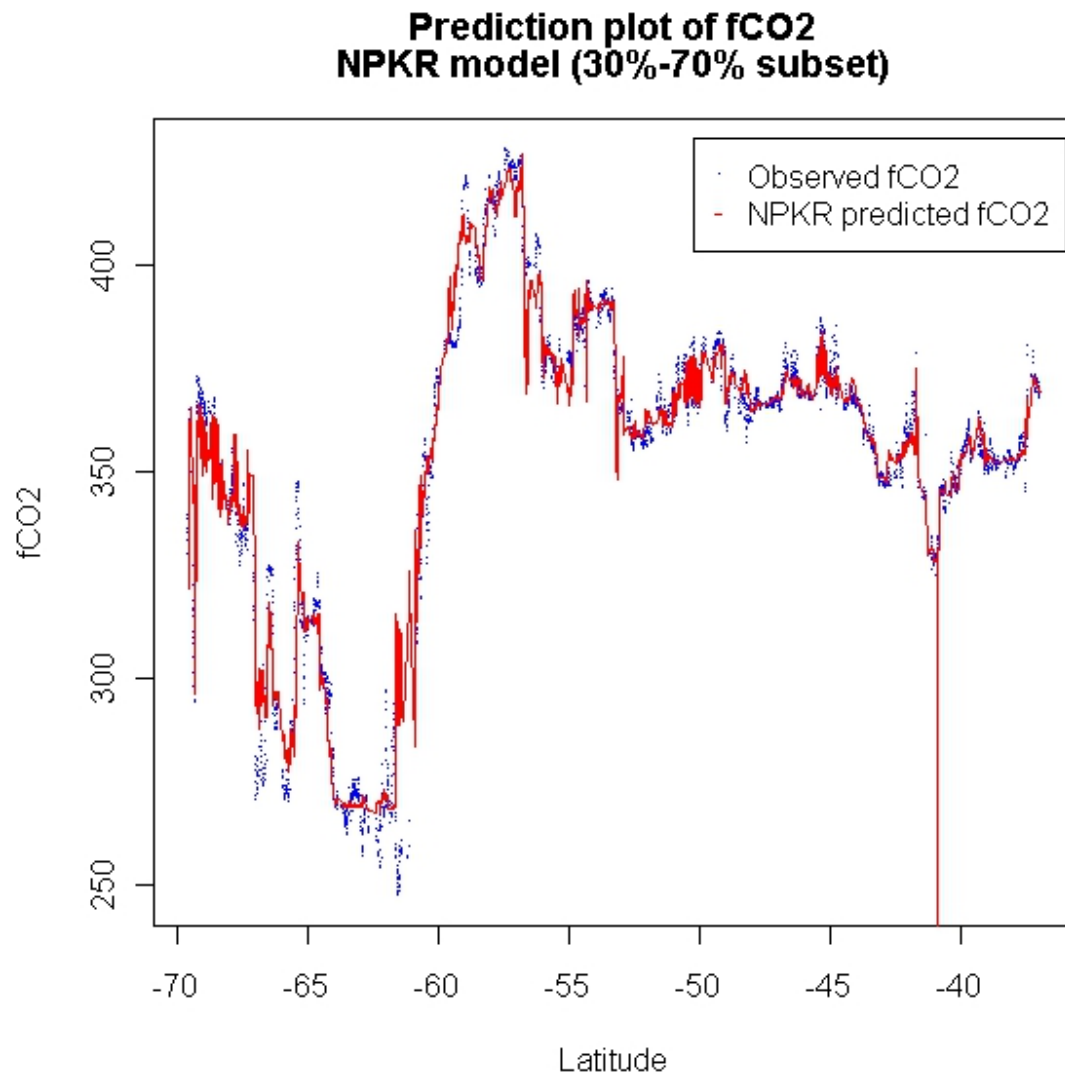


Figure 4.4: Non-parametric kernel regression observed and predicted fCO₂ for model M8 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

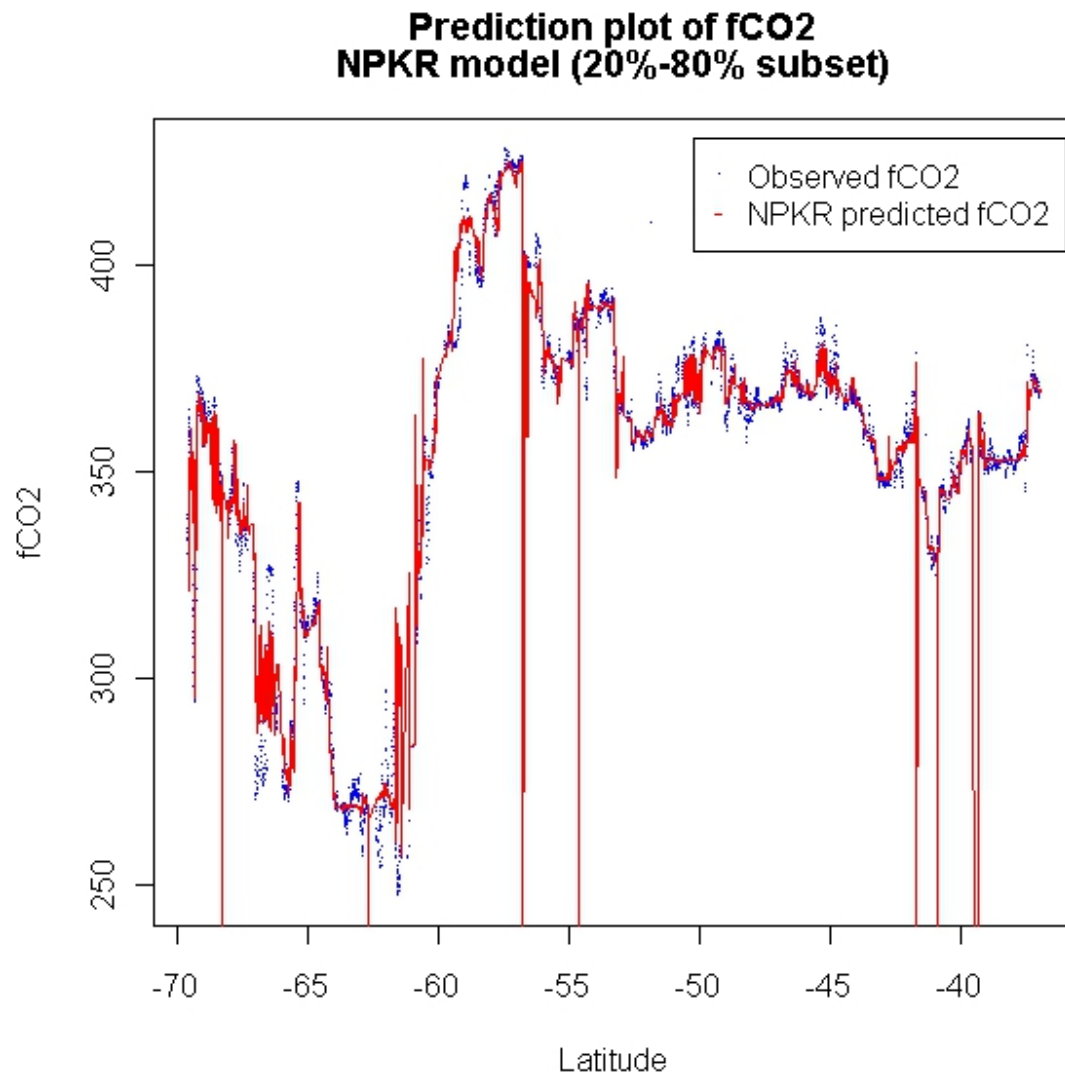


Figure 4.5: Non-parametric kernel regression observed and predicted fCO₂ for model M9 (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

chlorophyll-a and MLD) with a 70%-30% division between training and test data subsets. The models were also assessed using their respective test data sets and the results were plotted in the histograms shown in Figures 4.6, 4.7 and 4.8. These figures display the observed frequencies of MSE, MAE and RMSEs respectively. These plots allow for the comparison of the distribution of the error rates obtained from the NPKR models to those obtained from the MLR models in Chapter 3.

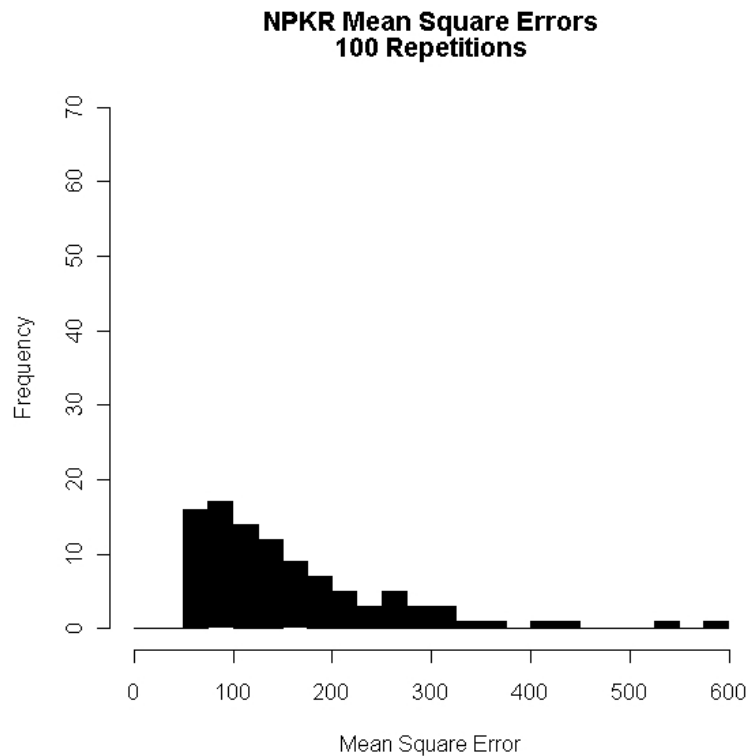


Figure 4.6: Histogram of 100 non-parametric kernel regression model MSEs

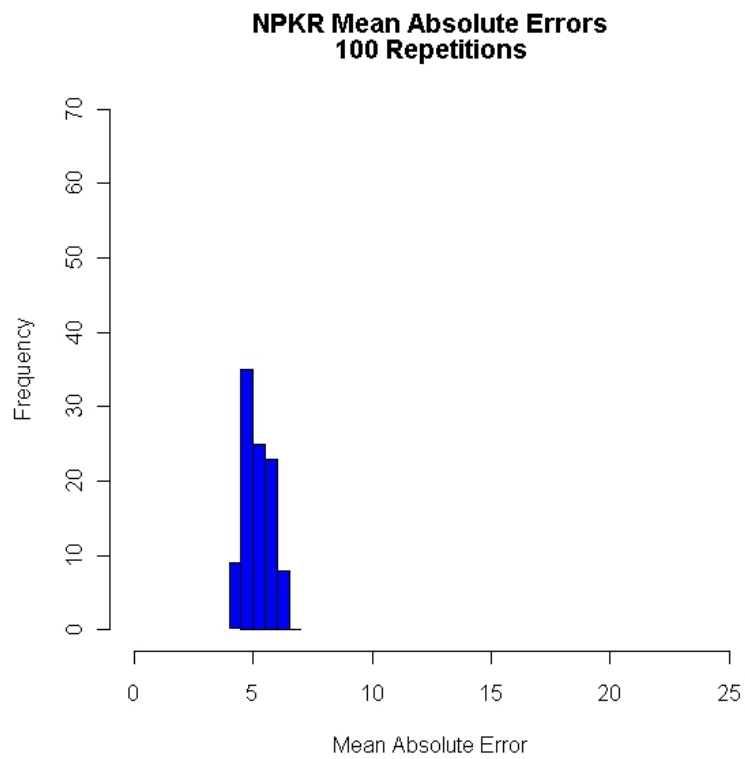


Figure 4.7: Histogram of 100 non-parametric kernel regression model MAEs

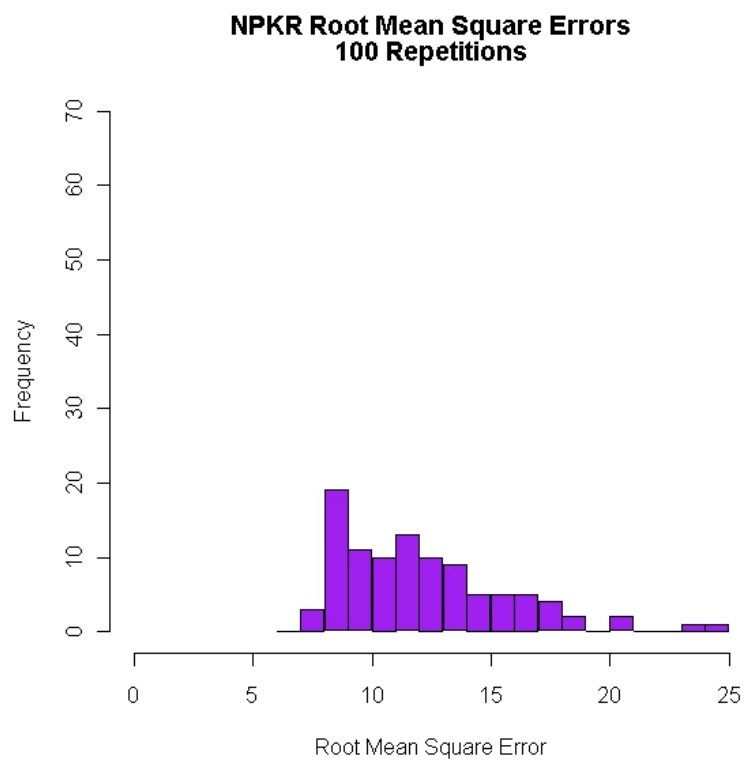


Figure 4.8: Histogram of 100 non-parametric kernel regression model RMSEs

The distribution of the 100 error rates obtained from the models developed is further described in Table 4.5. This table provides the mean, standard deviation, COV, minimum, median and maximum of the MSEs, MAEs and RMSEs obtained from the models. This allows for a comparison of the distribution of these error rates with those obtained from the MLR models.

Table 4.5: Non-parametric kernel regression error rate statistics for 100 subset divisions

	Mean Square Error	Mean Absolute Error	Root Mean Square Error
Mean	159.997	5.183	12.147
Standard Deviation	100.479	0.528	3.546
Coefficient of Variation	0.628	0.102	0.292
Minimum	58.394	4.124	7.642
Median	131.962	5.102	11.487
Maximum	582.239	6.381	24.130

4.5 Discussion of the non-parametric regression results

This section presents a discussion of the results of the NPKR model development and assessment provided in Section 4.4. The method provides a non-parametric alternative to the multiple linear regression approach of Chapter 3, which can be compared to other non-parametric approaches used in previous studies such as the SOM models of Telszewski *et al.* (2009).

4.5.1 Model bandwidth interpretation

Table 4.1 provides the bandwidths for each of the variables in the 10 NPKR models developed as described in Section 3.4. These bandwidths identify the fixed neighbourhoods which define the observations “near” to a target input vector for which a predicted fCO₂ is required. The bandwidth values are af-

ected by the unit of measurement and therefore the magnitudes of the bandwidths in models M1 to M9 are not comparable. These bandwidths are fixed values and therefore constitute a fixed nearest neighbourhood irrespective of where the test observation occurs in input space. For the NPKR approach, in order for an estimate to be made (i.e. a non-zero estimate), at least one observation in the training data set must fall within the bandwidth distance around each of the observed independent variables of the test observation.²

4.5.2 Model error rates

The error rates (MSE, MAE and RMSE), as indicated in Table 4.2, display the improvement of the NPKR models due to the inclusion of latitude (in model M2) and salinity (in model M3) into M1 individually. Models M1, M2 and M3 were developed on identical training data subsets and assessed using the same “unseen” test subset. The improvement in error rates is much larger in these models (M1, M2 and M3) compared to the corresponding MLR models discussed in Section 3.5. In model M2 a decrease in MSE of $38.469\mu\text{atm}$ which is a 50.25% decrease. Compared to the 3.5% decrease in MSE for the MLR models, this is a notably larger amount. Similarly the MAE and RMSE display a decrease from model M1 to model M2 of 28.5% and 29.5% respectively. These decreases in the error rates due to the inclusion of latitude are expected, but cause a problem in the model when an attempt is made to extrapolate the predictions on a larger scale. This is due to the fact that certain oceanic processes do not occur at the same latitude intra- or inter-annually. Model M3 includes salinity into the NPKR model. This model displays an even greater improvement than model M2 in terms of MSE. The MSE improves from $78.560\mu\text{atm}$ in model M1 to $33.750\mu\text{atm}$ in model M3. This constitutes a 55.92% decrease in MSE and therefore salinity is observed to have a positive effect on the model. The MAE and RMSE also display large decreases from

²If no observations in the training data are found which satisfies this requirement then the model simply estimates the fCO_2 value as 0.

model M1 (23.88% and 33.61% respectively). These results all suggest that salinity could be a useful addition as an independent variable in the model. The disadvantage of salinity as an independent variable is that it is not yet accurately globally available and therefore the issue for this model becomes one of either removing salinity altogether from the model, or finding another variable which can capture the same information as salinity in the model, but still be remotely available. In this chapter the former is chosen.

Comparing these results to those of Telszewski *et al.* (2009) and Friedrich and Oschlies (2009)³ provides a good measure of how well these NPKR models are able to predict $f\text{CO}_2$ values in the SO compared to other non-parametric methods used in the North Atlantic (which has much better coverage in terms of *in situ* data). Both the aforementioned studies focused on annual predictions of $p\text{CO}_2$ rather than seasonal predictions. Friedrich and Oschlies (2009) develop SOMs using only SST and chlorophyll-a concentration, and identified two separate error rates of their NN estimates - A RMSE of $19\mu\text{atm}$ was observed along the VOS lines not in the training data and where there were gaps in the remote sensing data; and where coverage by the satellite is not optimal, the RMSE was slightly higher (around $21.1\mu\text{atm}$). Telszewski *et al.* (2009) used another SOM approach to develop a model which displayed a RMSE of $8.1\mu\text{atm}$, $12.6\mu\text{atm}$ and $12.5\mu\text{atm}$ for the years 2004, 2005 and 2006, respectively. This presented a large improvement on the approach by Friedrich and Oschlies (2009) and also compares well to the RMSE values obtained from the *in situ* data used in the analysis of SANAE49L6-final. These errors, however, are only with respect to remote sensing along the VOS lines. The corresponding RMSE obtained by Friedrich and Oschlies (2009) (for those areas corresponding to the VOS lines used in training the models) was $6.3\mu\text{atm}$, which is less than a third of the basin-wide error rate ($21.1\mu\text{atm}$). This suggests that this error rate does not necessarily represent the basin-wide error

³both use $p\text{CO}_2$ as response instead of $f\text{CO}_2$

rate, however Telszewski *et al.* (2009) suggested that an improved training scheme (comprising of larger data sets) produce much closer results between the errors on the VOS lines and the basin-wide errors.

4.5.2.1 Training-Test Data splits

In the MLR model assessment described in Table 3.4, the error rates associated with decreasing sizes of the training data subset seemed to be fairly constant which indicated that the division of training and test data sets did not have an effect on the predictive ability of the MLR model. Table 4.3, however indicates a different trend for the NPKR models. Although it is not a smooth increase, there is definitely an increasing tendency in the MSEs of the models when the independent variables are kept the same, with only the size partition of the training and test data sets changing. The difference in MSE and RMSE between the model with the highest percentage data in the training data subset (M4) and the model with the lowest (M9) is $208.186\mu\text{atm}$ and $8.335\mu\text{atm}$ respectively. These values present a 300.68% and a 100.17% increase in MSE and RMSE respectively for the model with only SST, log chlorophyll and MLD as predictor variables. These increases suggest that the NPKR model performs better as the size of the training data subset decreases. Although this seems to be true in the MSEs and RMSEs, the same can not be said of the MAEs.

The MAE values in Table 4.3 do not display any pattern in terms of an increasing or decreasing error rate. This is unexpected, but inspection of the plot of predicted values of fCO_2 versus the observed values in Figures 4.4 and 4.5 indicate several individual predicted values of fCO_2 , which deviate away from the observed values while the rest of the predictions seem to be accurate. These “bad” estimates are caused by the observations in the test subset having independent variables which have no observations in the training subset that fall within their local neighbourhood defined by the bandwidths. This results in a zero prediction by the NPKR model, which produces a large error (and

therefore a very large squared error) for that test observation. These “bad” errors inflate the MSE (and thus the RMSE), however the MAE is more robust towards large individual errors. These zero estimates produced by the models containing a smaller amount of data in the training data subset may also happen when the model (developed on the SANAE49L6-final *in situ* data) is extrapolated to the entire SO due to the complex nature of the SO in different regions.

Although it is important to identify that the models seem to still perform well (ignoring the zero estimates) regardless of the size of the training data subset, the issue of the “bad” estimates must still be dealt with in order to prevent such errors occurring when the model is used to predict fCO₂ estimates for the entire SO.

4.5.2.2 Standardising non-parametric regression models

The model developed and assessed using the standardised variables produced results comparable to those obtained using the MLR method described in Section 3.5.2.2. In both approaches, the standardised model (developed and assessed on the training and test data subsets) produced error rates much higher than the non-standardised models. The NPKR standardised model’s error rates are given in Table 4.4. The MSEs here are $1494.527\mu\text{atm}$ for the model assessed on the test subset standardised using the training subset’s variable means and standard deviations and $1483.087\mu\text{atm}$ for subset standardised using the test subset’s variable means and standard deviations. This is in excess of 18.5 times more than the MSE obtained from the NPKR model M1.

Similarly the MAE and RMSE values are increased to a large extent from model M1. The MAE increased by more than 445% while the RMSE displayed an increase of over 340%. These indicate that the standardised models (whether under the MLR method or under the NPKR method) in the SO seem to produce much worse models than the non-standardised models. This

confirms that the standardisation procedure suggested by Jamet *et al.* (2007) in the North Atlantic is not applicable in the SO.

4.5.2.3 Simulating non-parametric regression error

From the 100 repetitions of model M1 using randomised divisions of the training and test data subsets, the MSE values in Figure 4.6 seem to be centered around $100\mu\text{atm}$ (the RMSE rates are centered around $10\mu\text{atm}$ which is approximately 2,82% of the mean observed fCO_2 value), but are much more spread out than those observed in the MLR models from Figure 3.4. This identifies the disadvantage of the NPKR method in that the model produces very accurate estimates, but the results are much more variable than those from the MLR models with test error rates close to $600\mu\text{atm}$ also being observed from the NPKR models. This suggests that the NPKR models are very data dependent, relying heavily, not only on the amount of data in the training data subset (from Section 4.4.2.1), but also on which observations are included in the subset. This is, however, not entirely true. Figure 4.7 provides further evidence that the higher MSE and RMSE rates presented in Figures 4.6 and 4.8 are mainly due to the “zero” or “bad” estimates discussed earlier. These “zero” estimates inflate the MSE and RMSE and therefore produce inflated error rates. The MAE, on the other hand, is much more robust towards outlying values and therefore provides evidence that the NPKR models do produce more accurate predictions. Figure 4.7 shows this in that the MAE rates are centered around $5\mu\text{atm}$ which is about 1,41% of the mean observed fCO_2 . The MAE rates displayed in Figure 4.7 do not indicate the same variability as the MSE and RMSE values and are much lower than the MAE values observed in the MLR models of Section 3.5.2.3

Table 4.5 provides some descriptive statistics of the error rates obtained from the 100 subset repetitions. The average MSE ($159.997\mu\text{atm}$) and MAE ($5.183\mu\text{atm}$) obtained are much lower than those seen in the MLR method

of Chapter 3. A decrease of approximately 52% and 71.5% is observed in the average MSE and MAE respectively. This indicates that the error rates for the NPKR models seem to be centralised around much lower values than the error rates for the corresponding MLR models. The higher standard deviation in the error rates observed in Figures 4.6, 4.7 and 4.8 could further be identified by both the standard deviations of 100 repetitions of the error rates as well as the COV. The COV for the MSE, MAE and RMSE rates in the NPKR models is 0.628 (62.8%), 0.102 (10.2%) and 0.292 (29.2%). These values are much larger than the COVs of the MLR models. The NPKR models, however, did produce much lower minimum error rates ($58.394\mu\text{atm}$, $4.124\mu\text{atm}$ and $7.642\mu\text{atm}$ for the MSE, MAE and RMSE respectively) than the MLR models. The maximum MSE and RMSE for the NPKR models, however, are much larger than those from the MLR models. This suggests a much wider spread of the error rates for the NPKR models. The median MSE, MAE and RMSE values in the NPKR models, however, are still much smaller than the median values for the MLR models which provides further evidence that the error rates for the NPKR models are centralised around smaller values than the error rates for the MLR models.

4.6 Summary

The NPKR model approach seems to provide a more accurate approximation of the relationship between the explanatory variables discussed and $f\text{CO}_2$ in the SO for SANAE49L6-final than the MLR models. The inclusion of latitude and salinity in the NPKR model improves the model accuracy. A variable such as sea surface topography (which is remotely available) may also be able to capture the effect of salinity in the SO and thereby improve the NPKR model's predictive ability when used in unsampled area of the SO where only remote sensing is available. The lower error rates for all models identifies the reduced

bias of the NPKR models as opposed to the MLR models. This decrease in the prediction bias of the NPKR models is, however, coupled with an increase in the variability of the error rates seen in Figures 4.6 and 4.8 as well as from Table 4.5. The MAE rates seem to not be adversely affected by the size of the training subset, or which observations are included in it. This is due to the fact that (as seen in Figures 4.4 and 4.5) the inflated MSE is caused by single “zero” estimates, which result in large errors, further inflated by the MSE. These “zero” estimates are caused by test observations being predicted for which the neighbourhood (as defined by the bandwidths) contains none of the training observations.

The NPKR approach on its own, therefore, may estimate $f\text{CO}_2$ well in areas of the SO where data is readily available, but not in unsampled areas. Although the NPKR approach seems to be a step in the right direction, a solution to the “zero” estimates is needed. Since the MLR approach captured the general form of the test subsets (as seen in Section 3.5), it seems appropriate to incorporate the MLR model predictions where the “zero” estimates occur in the NPKR model, generating an improved, semi-parametric approach.

Chapter 5

Sea surface topography and the mixed regression model

5.1 Introduction

The previous models, while capturing the relationship between *in situ* $f\text{CO}_2$ and its drivers, however, also displayed short-falls in the ability to predict $f\text{CO}_2$. Two specific attributes observed in the non-parametric kernel regression (NPKR) models are the improvement of the predictive ability of the model on inclusion of salinity and latitude into the model, as well as the presence of “zero” estimates which produce large errors in the predictions (further inflated by the mean square error (MSE) and root mean square error (RMSE)). The multiple linear regression (MLR) models, on the other hand, displayed a larger predictive bias than the NPKR models. These specific short-falls are addressed in this chapter which proposes the inclusion of a new independent variable - altimetry - in an attempt to capture the effect of up-welling of nutrient and CO_2 rich deep-waters on the $f\text{CO}_2$ values. Also, although models including salinity as an independent variable produced better results, salinity is not yet reliably available via remote sensing and, since it is believed that regions of the ocean with similar altimetry measurements have similar salinity levels, altimetry also

acts as a proxy for salinity. A new, mixed model is proposed, which combines the estimates of the MLR and NPKR models in an attempt to eliminate the “zero” estimates in the NPKR model. The objective in this chapter is thus to identify if this mixed model approach can be used to improve the predictability of $f\text{CO}_2$ in an unseen data set not used in the model development stage. This will provide insight into the ability of the model to generalise to unsampled areas of the SO.

5.2 Sea surface topography

5.2.1 Background on sea surface topography

In the SO, it is believed that no single physical or bio-geochemical factor is solely responsible for all the past variability in CO_2 (Sigman and Boyle, 2000; Archer *et al.*, 2000). The sea surface topography represents the sea surface height (SSH) relative to the earth’s geoid. This geoid can be described as the surface at which Earth’s gravity acts the same at all points (i.e. is constant). Simply, this implies that the geoid represents the shape the sea surface would be if no movement occurred in the ocean (Fu *et al.*, 2010). The method of measuring this SSH (topography) relies on an altimeter and hence the sea surface altimetry (SSA) variable discussed in this chapter attempts to incorporate this topography into the model.

The altimetry of the ocean, and hence the sea surface topography, is influenced by many dynamic factors including wind speeds, air pressure and ocean currents. The effect of wind on the ocean surface can cause a phenomenon known as deep water upwelling (or downwelling depending on the direction of the wind). This is when surface waters in the ocean are driven away from a certain area (or towards a certain area). This results in deeper, more nutrient rich, waters being drawn up to replace the surface water (or surface water being driven down in the case of downwelling) (Gaines and Airame, 2012; Lindstrom,

2012). In the SO, the nutrient rich deep waters are also CO₂ rich due to CO₂ absorbing algae which die and sink to the ocean depths where the absorbed CO₂ is then stored. Deep water upwelling therefore results in this stored CO₂ being brought to the surface.

Figure 5.1 shows the sea surface topography measured over the entire globe using satellite altimetry, near-surface drifters, National Centers for Environmental Prediction (NCEP) wind and Gravity Recovery and Climate Experiment (GRACE) measurements. The 1992-2002 mean ocean dynamic topography data has been obtained from Maximenko and Niiler (2009). The data used and a discussion thereof is provided in the paper by Maximenko *et al.* (2009). The enlarged section on the right of Figure 5.1 displays the sea surface topography in the area of ocean covered by the SANAE49 ship.

The SSH in this area seems to be divided into zones. The ocean area close to Antarctica has a topography well below the geoid since the dark blue colour indicates sea surface heights of up to 200cm below the geoid. A slightly lighter blue band nearer to the 50°S latitude mark indicates a rapid ascension to near 100cm below the geoid. A further increase in topography around 45°S moves it up to just below the geoid. Further north, the altimetry seems to be more variable than was seen in the south. The SSH increases to above the geoid around 40°S, but at latitudes closer to Cape Town, the SSH decreases to just below the geoid again.

As far as is known, the use of sea surface topography is limited in statistical models. The inclusion of the SSH (described as the altimetry in this data set) is done in order to capture the effect of the nutrient and CO₂ rich deep water upwelling as well as attempting to capture the positive effect salinity had on prediction errors in the MLR and NPKR models. The sea surface topography also introduces flexibility into the model as compared to latitude as is used in the MLR and NPKR approaches of Chapters 3 and 4. The altimetry also allows the non-parametric model to identify certain zonal boundaries in the

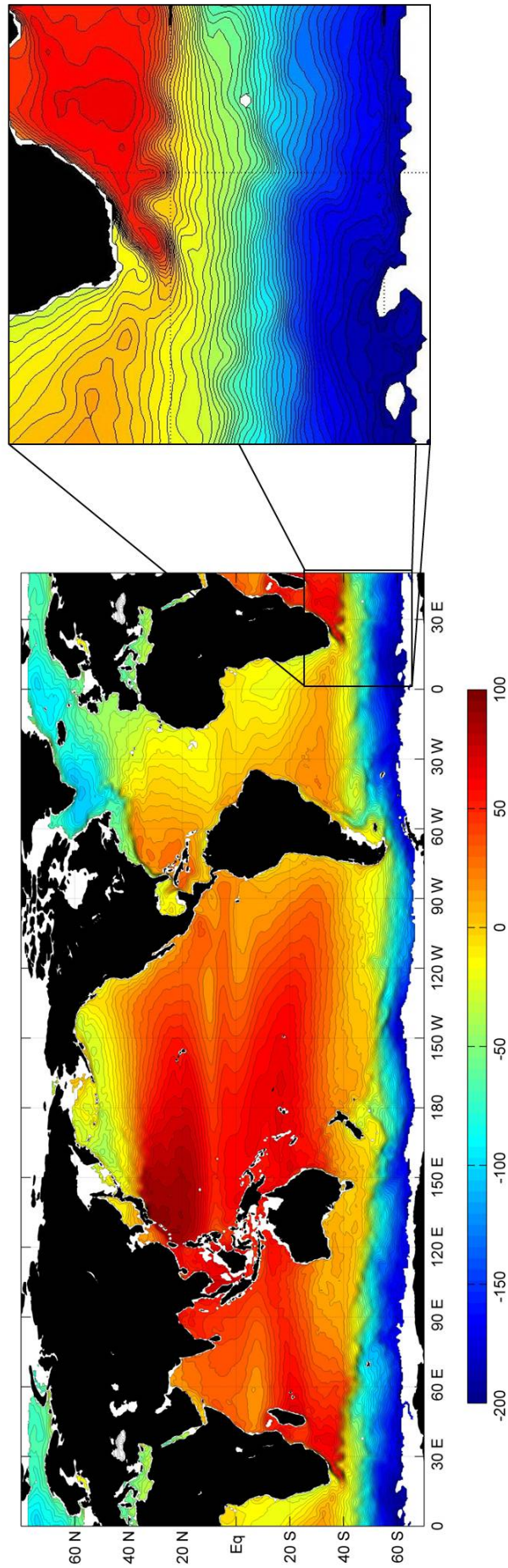


Figure 5.1: 1992-2002 Mean dynamic ocean topography on a 0.5° grid Maximenko and Niiler (2011)

SO, which will be seen in the plot of the altimetry data.

5.2.2 Altimetry data collection

The altimetry data, used to describe the sea surface topography, is obtained from the SOCCO group of the council for industrial and scientific research (CSIR). The altimetry readings are obtained to correspond exactly to the latitude and longitude observations in the SANAE49L6-final data set. Once the data was collected and added to the SANAE49L6-final data set, a new data set consisting of 13 variables and 6101 observations was created, henceforth referred to as SANAE49L6-Alt. Since the other variables and observations in this data set remain the same as the SANAE49L6-final data set described in Section 2.4, the only variable which has not been discussed is the altimetry. Table 5.1 displays the descriptive statistics of the altimetry data used. The table provides the number of observations, number of missing values, mean, standard deviation and COV of the altimetry data.

Table 5.1: Descriptive statistics of altimetry data

Variable	N	Missing	Mean	Standard Deviation	Coefficient of Variation
Altimetry (in meters)	6101	0	-0.703	0.705	-1.003

The first observation made regarding the descriptive statistics is the negative mean (-0.703 meters) value. This indicates that the SSH (topography) seems to be centered below the geoid (SSHs below the geoid are represented by negative altimetry values). The standard deviation seems small (0.705m), but upon further inspection using the COV, it is seen that the absolute value (1.003) indicates the standard deviation is as large as the mean (slightly larger). This suggests that the altimetry data has large variability relative to the mean.

Table 5.2 provides further statistics to describe the altimetry data in terms of the shape and range, and provides the minimum, first quartile, median,

third quartile and maximum of the observed altimetry values. These statistics give an indication of the shape, location and range of the data.

Table 5.2: Shape and range statistics of altimetry data

Variable	Minimum	Q1	Median	Q3	Maximum
Altimetry	-1.442	-1.341	-1.011	-0.07617	0.6733

From Table 5.2, a definite location of the data can be seen. The minimum, first quartile, median and third quartile all are less than 0, indicating that for the most part, the SSH is below the geoid. The maximum value is above the geoid, however not to the same magnitude as the negative values. The median altimetry measurement is 0.308m less than the mean, suggesting that the altimetry data is right skewed, further strengthening the initial observation that the majority of the data is negative (i.e. below the geoid).

Figure 5.2 presents a line plot of the altimetry measurements used in this analysis versus the latitude at which the measurements are taken. The plot allows for the identification of sudden changes in the altimetry, which indicate a movement from one oceanic zone into another.

The first abrupt change in Figure 5.2 occurs between 60°S and 55°S (nearer to 55°S). Before this the altimetry has little variability. Near 50°S, another (much larger) sudden change in the altimetry is observed. Then finally at 45°S, after a short period of little variability in altimetry between 50°S and 45°S, the last sudden change occurs. Above 45°S, the altimetry seems to be more variable than before.

5.3 Regression models to include altimetry

This section describes the models developed which include altimetry as a predictor variable, as well as the development of the mixed models that will be used in this chapter to improve on the estimation of $f\text{CO}_2$.

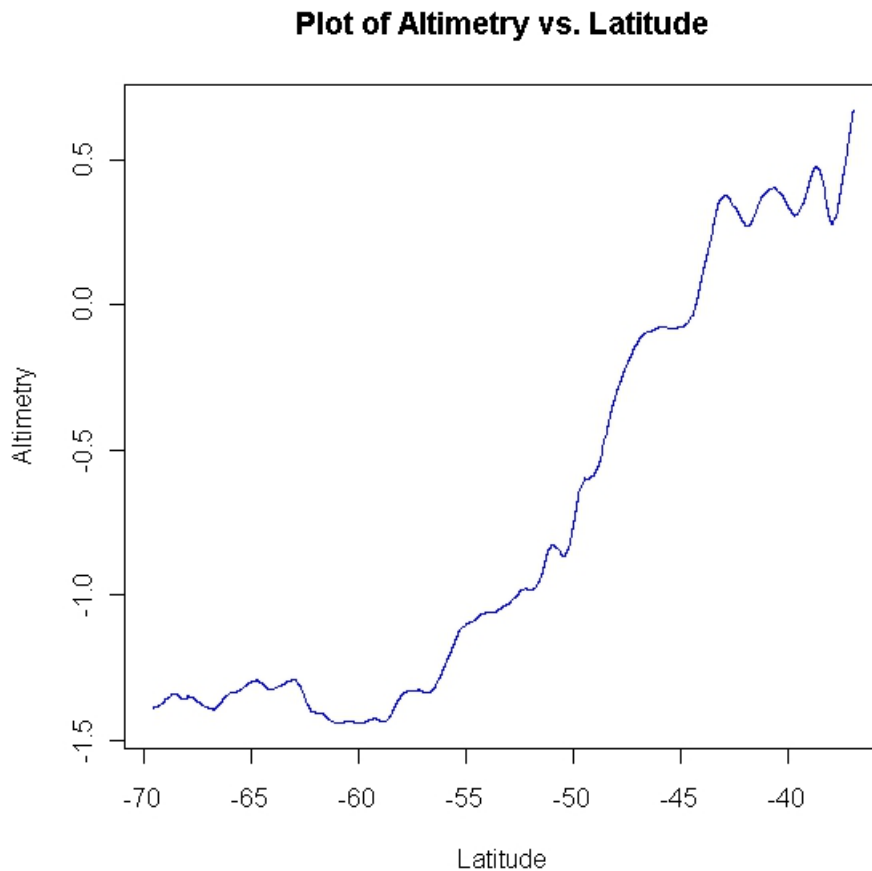


Figure 5.2: Line plot of altimetry versus latitude

5.3.1 Developing the regression model

Altimetry, as described in Section 5.2, is a variable which can also be remotely measured for the entire SO. It was also suggested, by domain experts in the SOCCO group, that regions of the ocean which have similar altimetry measurements (i.e. topography), have similar salinity values and therefore altimetry may act as a proxy for salinity. For this reason it was included in the models as an independent variable. The independent variables in the model will now consist of the SST, log chlorophyll-a, MLD and altimetry (all these variables are available by remote measurement). The models used to compare the inclusion of altimetry into the model to the other MLR and NPKR models as discussed in Chapters 3 and 4, is now denoted as model M11. Both the MLR and NPKR models for M11 are developed and compared to the results

of models M1, M2 and M3 in Chapters 3 and 4.

The inclusion of another variable in the MLR model should not have an adverse effect on the prediction error of the model. Including a new variable may, however, over parameterise the model. The same is not true for the NPKR models. By including a new variable in the model, the so-called “curse of dimensionality” becomes worse. This means that the local neighbourhood around each of the observations in the test data subset required for an estimate to be made becomes less local due to the dimensions of the input space being increased. This results in a method which may no longer be local around the observation being estimated. Also, if a constant neighbourhood size is used for all estimates, this generates a higher chance for “zero” or “bad” estimates to occur since smaller amounts of data fall within these neighbourhoods. Although this may not always occur for all divisions of the training and test data set, it has been seen in Section 4.4 that even for 3 input variables, this occurs when the amount of training data is decreased as well as for some divisions of the training and test subsets when this division is repeated a number of times. For this reason, the mixed model is used for estimating model M11 as a possible solution to the problem of “zero” estimates from the pure NPKR model.

5.3.2 Mixture of parametric and non-parametric regression models

The two model approaches, namely MLR and NPKR, discussed thus far have had both positive and negative aspects with respect to the bias of the model in predicting *in situ* $f\text{CO}_2$ from other *in situ* independent variables. The MLR model produces a larger bias due to the dynamic nature of the relationship between $f\text{CO}_2$ and the predictor variables, however on the positive side, did capture the general pattern of the response as well as produced less variable error rates due to a rigid model structure. This was seen in the results ob-

tained in Chapter 3. The NPKR model produced less biased predictions for an “unseen” test data subset which is an improvement on the errors of the MLR models. Although this improvement seemed to indicate a more accurate model, the variability of the test error rates suggested that the model presented problems. The biggest problem is that of the “zero” estimates produced due to there being no observations falling within the neighbourhood of the test observation estimated. The result is large individual errors that adversely affect the MSE and RMSE. This is particularly seen in the difference between the MSE (RMSE) and MAE values for the differing subset division models in Table 4.3. The MSEs indicate an increasing trend in the error rate as the amount of data in the training subset decreases, however the MAEs suggest that the error rate is unaffected by the size of the training data subset. For this reason a model is required to provide accurate estimates (i.e. with a low bias) with a low variability in the error rates.

A mixed model, which combines the predictions from both the NPKR and MLR approaches, is proposed to solve the issues described above, since the MLR model was shown to capture the general form of the test data. It is decided to define the mixed model $f\text{CO}_2$ estimates using (5.3.1)

$$\hat{y}_{Mixed}|\underline{X} = \underline{x} = \begin{cases} (\hat{y}_{MLR}|\underline{X} = \underline{x}), & \text{if } (\hat{y}_{NPKR}|\underline{X} = \underline{x}) = 0. \\ (\hat{y}_{NPKR}|\underline{X} = \underline{x}), & \text{otherwise.} \end{cases} \quad (5.3.1)$$

An alternative way of indicating the mixed model function would be to use indicator functions. The estimated $f\text{CO}_2$ will then be represented as

$$\begin{aligned} (\hat{y}_{Mixed}|\underline{X} = \underline{x}) = & I((\hat{y}_{NPKR}|\underline{X} = \underline{x}) = 0) \cdot (\hat{y}_{MLR}|\underline{X} = \underline{x}) \\ & + I((\hat{y}_{NPKR}|\underline{X} = \underline{x}) \neq 0) \cdot (\hat{y}_{NPKR}|\underline{X} = \underline{x}). \end{aligned} \quad (5.3.2)$$

The indicator function $I(\cdot)$ produces a 1, if the condition within the function is met and a zero otherwise. Essentially this describes a function which produces an estimate of the response ($f\text{CO}_2$) in one of two ways. Initially the estimated value of $f\text{CO}_2$ using the non-parametric method is assessed. If the estimate is

a zero, the first indicator function is 1, while the second is 0 and the result is that the model estimate of $f\text{CO}_2$ is obtained using the MLR approach. If the NPKR estimate is not 0, the first indicator function will return a 0, while the second returns a 1, therefore producing an estimate using only the NPKR approach.

5.4 Regression results to predict $f\text{CO}_2$

This section provides the results of both the NPKR and MLR models including altimetry (M11) as well as the mixed models described by (5.3.1) and (5.3.2). These results compare the addition of altimetry as an independent variable to the MLR and NPKR approach to model M1, which does not include altimetry as an independent variable, as well as to model M3 that includes salinity as an independent variable along with the independent variables of model M1. After the pure NPKR and MLR models are presented, the results of the mixed model are provided for models M1, M3 and M11 in order to compare the prediction results for the different models based on a common subset division (70% training subset, 30% test subset). The same training and test subset divisions were done for the mixed models as was done in Sections 3.5.2.1 and 4.4.2.1 for the MLR and NPKR models M1 and M3, so that the error rates obtained from the test data subsets can be compared to those of the NPKR and MLR models. Finally, the mixed model of M11 is estimated for 100 repetitions of training and test random divisions of 70%-30% in order to plot error rates in a histogram similar to Figures 3.4, 3.5 and 3.6 in Chapter 3 and Figures 4.6, 4.7 and 4.8 of Chapter 4. This allows for a comparison of the mixed model with the MLR and NPKR approach. Results reveal that the mixed model provides more accurate estimates in terms of lower errors than the MLR models, while reducing the variability of the errors for different subset selections from the NPKR models.

5.4.1 Estimating the pure parametric regression model and non-parametric regression model including altimetry

Table 5.3 presents the bandwidth estimates of the optimised MLR model M11 while Table 5.4 provides the optimal bandwidth estimates for the NPKR model M11. The methods for optimising these models is provided in Sections 4.3.1 (for the NPKR model) and 3.3.3 (for the MLR model).

Table 5.3: Optimised MLR parameter estimates

	Intercept	SST	log Chlorophyll-a Concentration	MLD	Altimetry
MLR	401.755	-6.989	-101.375	0.087	31.676

Table 5.4: Optimised NPKR optimal bandwidth estimates

	SST	log Chlorophyll-a Concentration	MLD	Altimetry
NPKR	0.444	0.035	7.131	0.025

5.4.2 Assessing the parametric and non-parametric regression models with altimetry

The altimetry model errors listed in Table 5.5 are those for the individual MLR and NPKR models of M11. The table includes the MSEs, MAEs and RMSEs for these model.

5.4.3 Estimating the mixed regression model

The model estimation for the mixed model consists of two parts. The first estimates the regression parameters of the MLR model, while the second part is considered with estimating the selection of bandwidths for the NPKR model.

Table 5.5: Error rates for model M11 including altimetry using MLR and NPKR approaches

	Mean Square Error	Mean Absolute Error	Root Mean Square Error
Multiple linear Regression	272.934	12.895	16.521
Non-parametric Kernel Regression	42.120	3.803	6.490

The estimated MLR parameters and NPKR bandwidths for models M1 and M3 for a 70% - 30% subset division are provided in Tables 3.2 and 4.1 respectively, while the estimated parameters and bandwidths for model M11 for subset divisions ranging from 80% - 20% to 20% - 80% (training data - test data) are provided in Tables 5.6 and 5.7 respectively.

Table 5.6: Multiple linear regression parameter estimates for differing subset divisions

Training - Test Subset Division %	Intercept	SST	log Chlorophyll-a Concentration	MLD	Altimetry
80 - 20	401.535	-6.996	-101.531	0.091	31.703
70 - 30	401.755	-6.989	-101.375	0.087	31.676
60 - 40	401.873	-7.008	-101.605	0.087	31.683
50 - 50	402.179	-7.053	-101.870	0.087	31.917
40 - 60	401.879	-6.979	-101.047	0.084	31.421
30 - 70	401.384	-6.947	-100.755	0.092	31.411
20 - 80	402.462	-6.984	-101.747	0.080	31.413

Table 5.7: Non-parametric kernel regression bandwidth estimates for differing subset divisions

Training - Test Subset Division %	SST	log Chlorophyll-a Concentration	MLD	Altimetry
80 - 20	0.172	0.043	7.721	0.040
70 - 30	0.444	0.035	7.131	0.025
60 - 40	0.444	0.035	7.112	0.025
50 - 50	1.885	0.035	5.432	0.013
40 - 60	0.172	0.102	5.720	0.012
30 - 70	0.172	0.043	6.120	0.022
20 - 80	0.209	0.017	6.837	0.051

5.4.4 Assessing the mixed regression model

The first approach to assessing the performance of the mixed models is to compare the error rates for mixed models M1, M3 and M11. Table 5.8 presents the MSEs, MAEs and RMSEs for the mixed models M1 and M3 containing the variables as described in Table 3.1 of Section 3.4 as well as model M11.

Table 5.8: Error rates for mixed models M1, M3 and M11

	Mean Square Error	Mean Absolute Error	Root Mean Square Error
M1	76.560	5.037	8.750
M3	33.750	3.834	5.809
M11	42.120	3.803	6.490

This table indicates the error rates for the 70% - 30% division of training and test data subsets which is identical to those used for the MLR and NPKR models M1, M2 and M3 in Chapters 3 and 4.

5.4.4.1 Training-test data splits

The mixed models were specifically selected in order to overcome the shortfalls of both the MLR and NPKR models. The NPKR models present the unique problem of “zero” estimates produced due to test observations falling in an area of input space where there are little or no observations which are within a close neighbourhood of the observation as defined by the bandwidths. The subset division error rates for the mixed models illustrate how well the mixed models respond to diminishing amounts of data in the training data subset as compared to the regular NPKR models, while also indicating the improved prediction errors as compared to the MLR models.

Table 5.9 presents the error rates of the mixed models developed and assessed on differing percentage subset divisions between the training and test data subsets. The table presents the MSEs, MAEs and RMSEs of mixed models with the same independent variables as model M11. The subset divisions

used are identical to those used in the MLR and NPKR approach of Sections 3.5.2.1 and 4.4.2.1 respectively.

Table 5.9: Error rates for mixed models developed and assessed on varying subset sizes

Training - Test Subset Division %	Mean Square Error	Mean Absolute Error	Root Mean Square Error
80 - 20	36.610	3.573	6.051
70 - 30	42.120	3.803	6.490
60 - 40	36.026	3.608	6.002
50 - 50	40.631	3.670	6.374
40 - 60	41.217	3.736	6.420
30 - 70	27.345	3.224	5.229
20 - 80	42.021	3.648	6.482

The prediction plots for the models with 30% - 70% and 20% - 80% divisions between training and test data subset are shown in Figures 5.3 and 5.4. These plots compare the predicted values of $f\text{CO}_2$ using the mixed (left) and NPKR (right) models versus the observed $f\text{CO}_2$ values over latitude. These subset divisions were chosen due to the impact of the “curse of dimensionality” on the NPKR models.

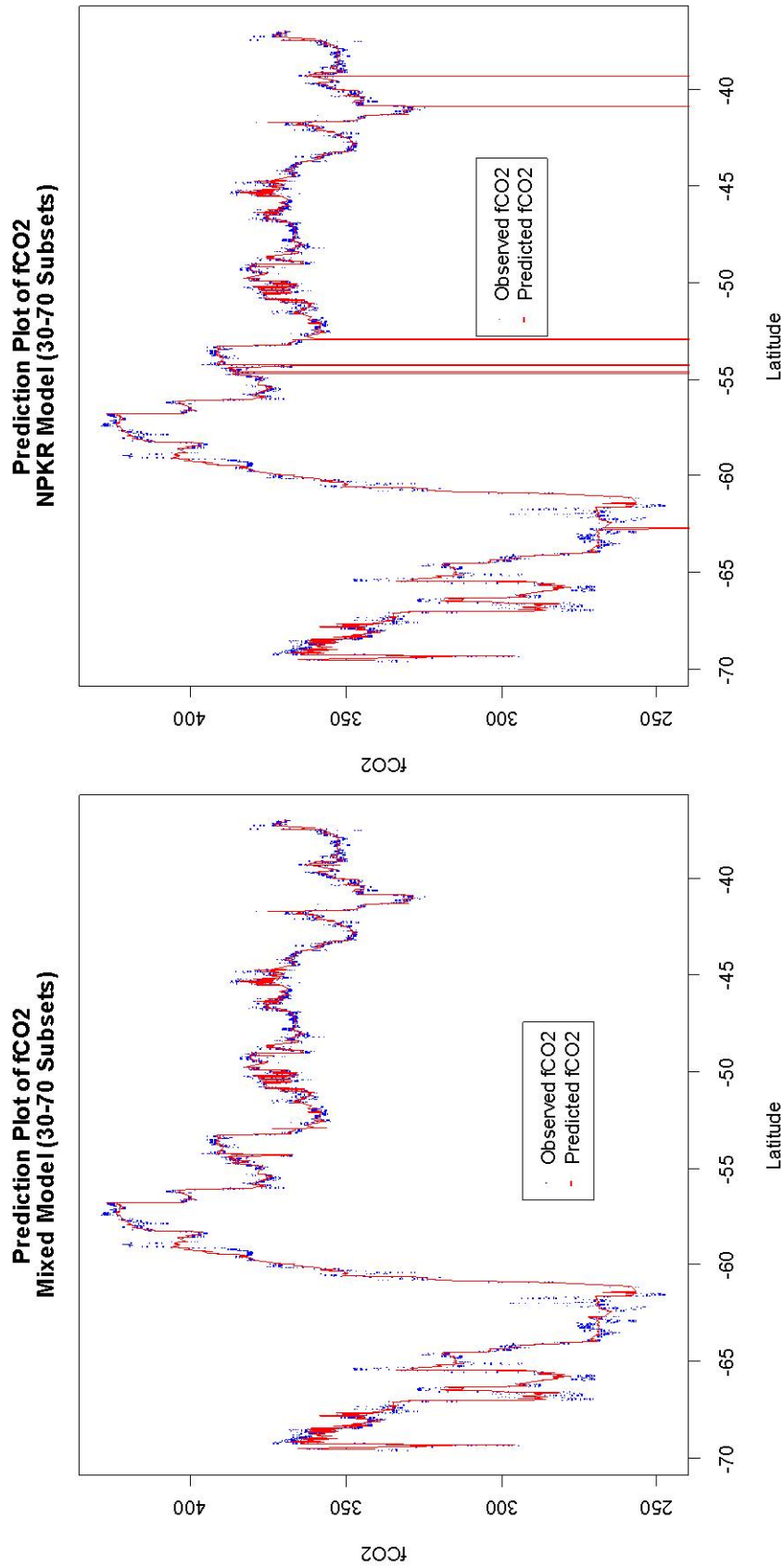


Figure 5.3: Prediction plots of fCO₂ versus latitude for the mixed model (left) and NPKR model (right) for the 30% - 70% subset division (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

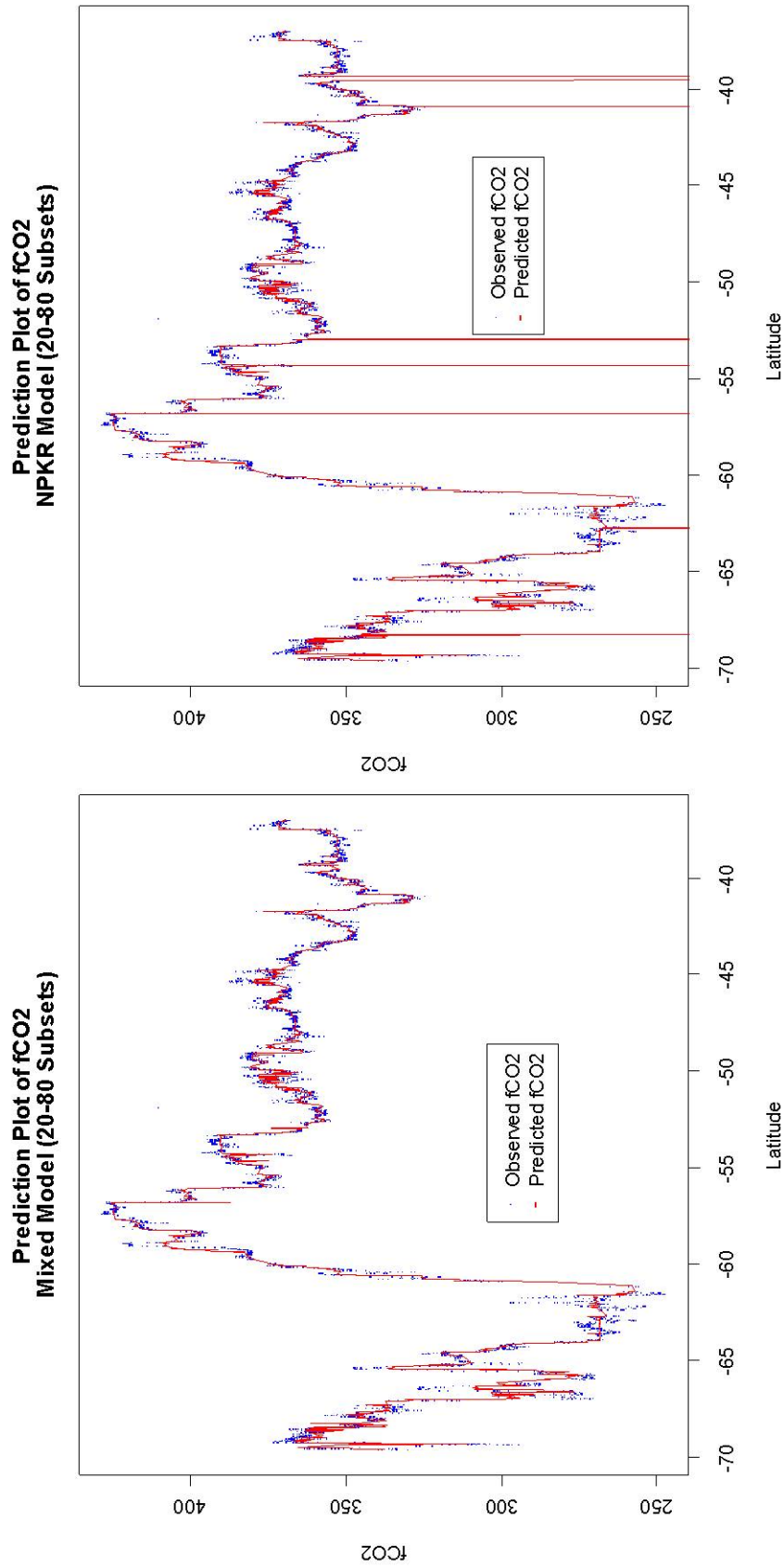


Figure 5.4: Prediction plots of fCO₂ versus latitude for the mixed model (left) and NPKR model (right) for the 20% - 80% subset division (blue dots represent observed fCO₂ while the red line represents predicted fCO₂)

5.4.4.2 Simulating the mixed regression model error

After investigating the effect of subset selection on the error rates of the mixed model approach, 100 different subset selections of the 70% - 30% division between training and test data subsets is performed using the independent variables from model M11. The histograms of the MSE, MAE and RMSE values produced by these 100 repetitions are provided in Figures 5.5, 5.6 and 5.7 respectively. These figures are compared to similar plots produced for 100 repetitions of the MLR and NPKR approaches using the same subset divisions used here and the variables from model M1. The histograms are placed on the same scale as those in Sections 3.5.2.3 and 4.4.2.3.

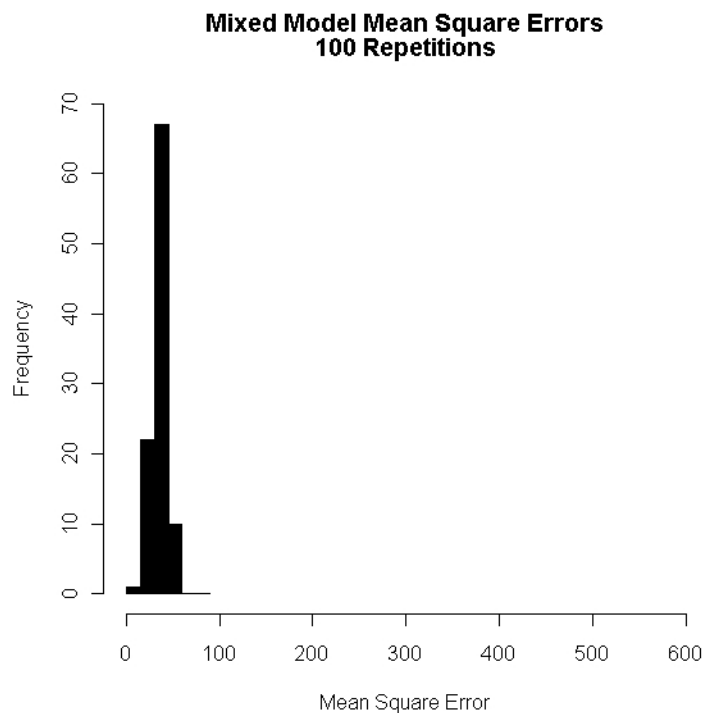


Figure 5.5: Histogram of mean square errors for 100 different subset divisions using the mixed model M11

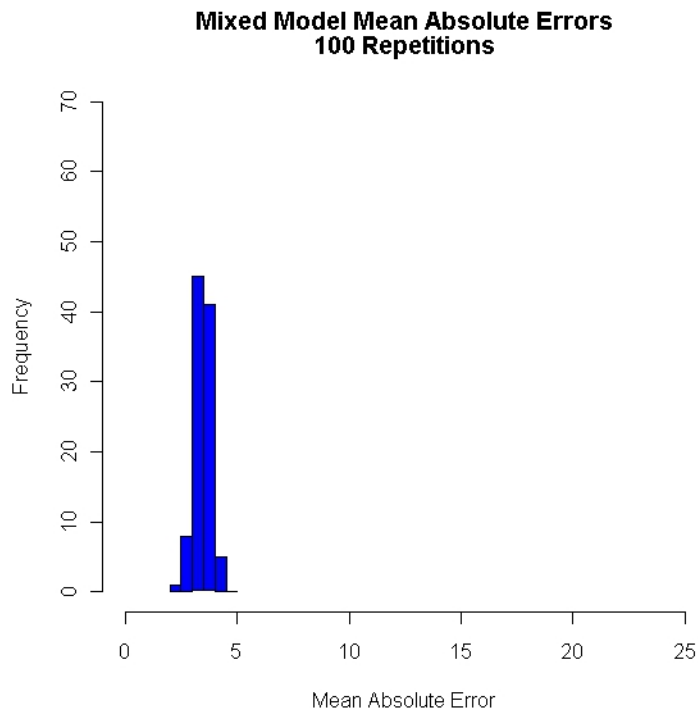


Figure 5.6: Histogram of mean absolute errors for 100 different subset divisions using the mixed model M11

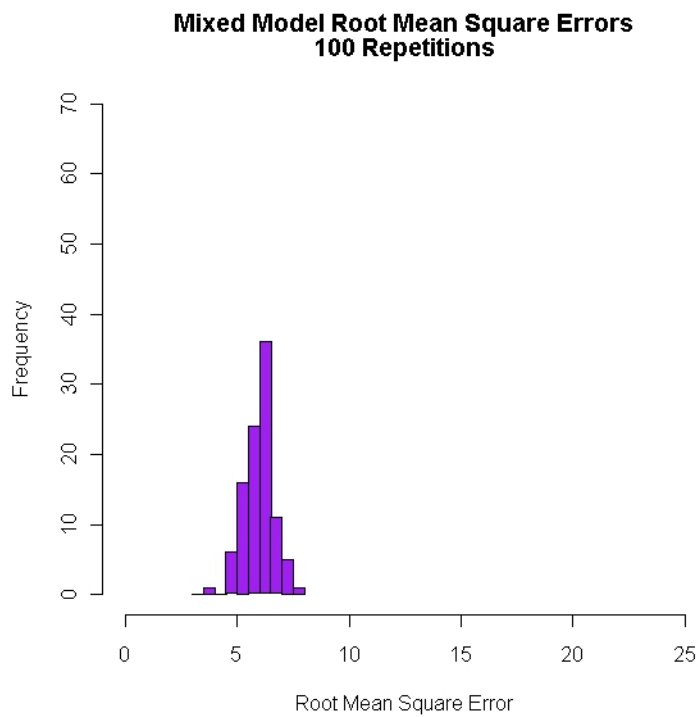


Figure 5.7: Histogram of root mean square errors for 100 different subset divisions using the mixed model M11

The histograms provide a visual representation of the distribution of the error rates for the mixed model, however a more quantifiable representation may be found in the descriptive statistics of the error rates found in Table 5.10.

Table 5.10: Descriptive statistics of the error rates of 100 repetitions of the mixed model M11

	Mean Square Error	Mean Absolute Error	Root Mean Square Error
Mean	36.030	3.465	5.968
Standard Deviation	7.719	0.343	0.650
Coefficient of Variation	0.214	0.099	0.109
Minimum	14.746	2.330	3.840
Median	36.323	3.460	6.027
Maximum	58.439	4.390	7.645

5.5 Discussion

The results in the previous section provide an indication of the effect of altimetry on the predictive ability of the MLR and NPKR models, as well as an assessment of the mixed model approach and how it solves the short-falls of the previous approaches. This section provides an interpretation and discussion of the results in order to draw conclusions surrounding the altimetry and its inclusion in both the regular and mixed models, as well as a discussion surrounding the use of a mixture of the MLR and NPKR model predictions as an alternative approach to model $f\text{CO}_2$ in the SO.

5.5.1 NPKR and MLR models including altimetry

5.5.1.1 Estimating the models

The MLR model parameter estimates and NPKR bandwidth estimates in Tables 5.3 and 5.4 respectively indicate the form of the optimised regression

functions. The MLR model parameters estimate the relationship between the independent variable and the response ($f\text{CO}_2$). Model M11 estimate indicate a negative linear relationship between SST and $f\text{CO}_2$ (-6.989) as well as between log chlorophyll-a concentration and $f\text{CO}_2$ (-101.375). The MLD, again, displays a positive linear relationship with $f\text{CO}_2$ (0.087), however as opposed to the inverse relationship that $f\text{CO}_2$ displayed with salinity, the MLR parameter estimate for altimetry is 31.676 indicating that for each meter further above the geoid the sea surface is, the $f\text{CO}_2$ will increase by $31.676\mu\text{atm}$. The NPKR estimated bandwidths do not have the same simple interpretation, since they depend on the unit of measurement for each variable, it is therefore not possible to compare the bandwidths to one another.

5.5.1.2 Assessing the models

Model M11, which includes altimetry as an independent variable, is estimated and assessed on the same random division of the training and test data subsets as models M1 and M3. The difference between these models is therefore only in the independent variables used (M1 makes use of SST, log chlorophyll-a and MLD, while M3 includes salinity in addition to these). The MLR and NPKR approaches are used in order to compare the error rates to those of models M1 and M3 in Chapters 3 and 4. The MSE from the MLR model M11 is $272.934\mu\text{atm}$, which is approximately 17% lower than the MSEs of the MLR model M1 ($328.789\mu\text{atm}$) and produces a 16% decrease from the MSE of MLR model M3 ($325.125\mu\text{atm}$) as seen in Table 3.3. Since the MSE is smaller, the same trend is also seen of the RMSEs of the MLR model M11 as compared to models M1 and M3. These decreases in the MSE of model M11 for the MLR model are also seen in the MAE of the MLR models where model M11 produces a MAE of $12.895\mu\text{atm}$. This is a decrease of 8.9% from the MAE of MLR model M1 and 7.25% from the MAE of MLR model M3. This indicates that by including the altimetry in the MLR model, the predictive

ability of the MLR models seems to have improved, however the MLR model M11 still seems to have an unsatisfactory RMSE and MAE of $16.521\mu\text{atm}$ and $12.895\mu\text{atm}$ respectively. When compared to the results of Jamet *et al.* (2007) as described in Section 3.6, it is seen that the MLR method still does not perform as well in the SO as in the wider sampled North Atlantic. The RMSE obtained in Jamet *et al.* (2007) using the MLR method and a model similar to model M1 was $11.44\mu\text{atm}$ (much lower than the $16.521\mu\text{atm}$ obtained for the SO even when altimetry is included as an independent variable) for the summer months. Even though Jamet *et al.* (2007) used pCO_2 as a response variable, the results are comparable and suggest that the MLR still does not provide an accurate model for CO_2 concentration in the SO.

The results of model M11, when using the NPKR approach, provide a similar indication in comparison to the results of models M1 and M3. One difference, however, is that model M3 seems to perform better than M11 (in terms of test error rates) when using the NPKR method. The inclusion of altimetry does, however, still seem to improve on model M1, where a decrease in MSE of approximately 45% is seen (from $76.56\mu\text{atm}$ to $42.12\mu\text{atm}$) while the MAE decreases by almost 24.5% (from $5.037\mu\text{atm}$ to $3.803\mu\text{atm}$). The MSE and MAE of model M3, however are $33.75\mu\text{atm}$ and $3,834\mu\text{atm}$ respectively, which is even lower still than model M11 and suggest that, from a non-parametric perspective, salinity seems to improve the ability of the model to predict fCO_2 in the SO more than the altimetry. However, both seem to improve the NPKR model when included as compared to model M1 which omits both salinity and altimetry as predictor variables.

5.5.2 Mixed Models

Table 5.8 presents the error rates for the 3 mixed models (M1, M3 and M11) and provides a basis which can be used to compare the altimetry model (M11) to the other models (M1 and M3) in terms of test prediction error rates when

the MLR and NPKR approaches are used together in the mixed models to predict $f\text{CO}_2$. What is interesting to note, for this specific division of the training and test data sets, is that the error rates of the mixed models are identical to those of the NPKR models as indicated in Tables 4.2 and 5.5. This is due to the fact that the mixed models only make use of the MLR model predictions if the NPKR model produces a “zero” estimate as was discussed earlier. For this reason the mixed model does not produce test error rates worse than the NPKR model and will only improve the error rates obtained if the NPKR model produces at least one “zero” estimate.

Although Table 5.8 indicates no difference in the mixed model error rates, to the NPKR models, it is important to note that for differing subset divisions of the training and test data subsets, this may not be the case. A situation where the “zero” estimates become a particular problem is when the model is generalised to the rest of the SO. Since the models are based on *in situ* data collected below South Africa and only in the summer months, when these models are used to predict observations further away or outside of where (or when) the data was collected, the problem of the new data observations having no training data points within a neighbourhood becomes a possibility. The following sections provide a better insight into how the mixed model reacts to differing divisions of the training and test subsets (i.e. how the model reacts to smaller amounts of information in the training data subset), as well as providing information on the distribution of the model test error rates for the 70% - 30% division of training and test data subsets for model M11.

5.5.2.1 Subset Division

The MLR parameter estimates and NPKR optimised bandwidths for the seven models discussed here are provided in Tables 5.6 and 5.7 respectively. As in Table 3.2, the MLR parameters display inverse relationships between $f\text{CO}_2$ and SST as well as the log chlorophyll-a concentration. The magnitude of these

negative relationships is between -6.95 and -7.05 for SST and -100.75 and -101.87 for log chlorophyll-a concentration. The positive relationship between $f\text{CO}_2$ and MLD as well as altimetry is also retained, with the magnitude of the changes in $f\text{CO}_2$ for a unit change in each of the variables being between $0.08\mu\text{atm}$ and $0.09\mu\text{atm}$ for MLD and $31.4\mu\text{atm}$ and $31.9\mu\text{atm}$ for altimetry. Once again, the optimised bandwidths do not have a similar, simple interpretation and hence are not discussed in detail.

The results for differing training and test subset divisions are provided in Section 5.4. The test error rates presented in Table 5.9 provide interesting results when compared to the MLR and NPKR approaches individually as discussed in Sections 3.6 and 4.5. The MSE figures for the mixed models which include altimetry as a predictor variable are lower for each subset division than both the MLR and NPKR approaches, which did not include altimetry. The largest MSE value ($42.12\mu\text{atm}$) is only 11.9% of the average *in situ* $f\text{CO}_2$. The MAE recorded from the mixed model, however, provides the most exciting results. The MAE ranges between $3.803\mu\text{atm}$ for the 70% - 30% subset division model and $3.224\mu\text{atm}$ for the 30% - 70% divided model. These MAE values indicate that the average absolute error produced by the mixed models is close to 1% of the average observed $f\text{CO}_2$ (provided in Table 2.3). This is very encouraging as the objective of this analysis, as laid out in Section 1.2.1 of Chapter 1, is to reduce the uncertainty of $f\text{CO}_2$ estimates in the SO to below 10% of the average $f\text{CO}_2$ values observed.

Figures 5.3 and 5.4 provide visual evidence of the benefit of the mixed models over the NPKR approach. In each figure, the prediction plot on the left presents the mixed model predictions of the test subset $f\text{CO}_2$ values overlaid on top of the observed $f\text{CO}_2$ values, while the plot on the right presents the NPKR predictions overlaid on top of the observed $f\text{CO}_2$ values. These plots represent the 70% - 30% and 80% - 20% subset divisions respectively due to the observable difference between the predictions of the NPKR and mixed

model approaches. The plots show an observable difference in the accuracy of the model predictions. The “zero” estimates produced by the NPKR approach can clearly be seen in these plots. The mixed model approach replaces these erroneous predictions with the MLR predictions due to the fact that, although the NPKR approach produces less biased estimates, the MLR method does seem to capture the trend of the $f\text{CO}_2$ values and therefore will produce more reliable estimates in these situations. The mixed model approach, therefore, improves on the NPKR approach by eliminating the “zero” estimates produced due to the curse of dimensionality, while keeping the more accurate estimates produced by the NPKR model.

The values of the MSE, MAE and RMSE for the 7 different mixed model subset divisions are provided in Table 5.9. What is interesting to note from this table is the similarity between the values and those produced from the NPKR approach of Chapter 4 (Table 4.3). The difference in the error rates between these two approaches is the consistency of the mixed model MSEs and RMSEs as opposed to the increasing errors of the pure NPKR models due to the “zero” estimates of $f\text{CO}_2$ (as seen in Figures 5.3 and 5.4). When compared to the MLR error rates in Table 3.4 of Chapter 3, the mixed model error rates are lower for each of the models than the MLR error rates. As was indicated in Section 5.4.4, the mixed models also produce lower error rates due to the inclusion of altimetry as a predictor variable as compared to the models used for subset division in Chapter 4 which did not include altimetry. This is seen in the MSE, MAE and RMSE values of the mixed model (for all subset divisions) is smaller than the same error rates in Table 4.3, for which the models did not include altimetry as an independent variable.

The mixed model error rates seem to be less biased than the MLR error rates, but also less variable than the NPKR error rates (as seen in the table of errors). Further investigation is, however, necessary in order to obtain an indication of the variability of these error rates and how spread out they are

in comparison to the MLR and NPKR approaches.

5.5.2.2 Model Simulation

The error rates of the 100 repetitions of subset divisions indicate interesting results from the mixed model M11. Specifically in Figure 5.5, the MSE values are much lower than those obtained from the MLR approach which are seen in Figure 3.4. This indicates the predictive superiority of the mixed model (which is a semi-parametric approach), as opposed to the fully parametric MLR approach. What is also of interest is the reduced spread of the MSE values as compared to the MSEs produced by the NPKR models. These error rates can be seen in Figure 4.6, where the spread of the MSEs seems to be much wider than the spread produced by the mixed model. This all suggests that the mixed modelling procedure improves on both the MLR and NPKR method, however it must be noted that the mixed model used here did include altimetry as an independent variable (model M11), while the MLR and NPKR models did not (model M1). This further reduced the errors of both the MLR and NPKR approaches (and hence also the mixed model) and therefore should be included in all subsequent models in future.

Figures 5.6 and 5.7 show the histograms of the MAE and RMSE respectively for 100 repetitions of subset divisions of a 70%-30% split between training and test data subsets. These histograms present similar trends to those from the histogram of the MSEs. The MAEs in Figure 5.6 all fall below $5\mu\text{atm}$, while the RMSEs are all below $10\mu\text{atm}$. These are lower than the error rates indicated by both the MLR and NPKR model approaches.

Table 5.10 provides a quantifiable method of comparing the mixed model errors to those of the MLR and NPKR approaches from Chapters 3 and 4. The average MSE, MAE and RMSE obtained from the mixed model are $36.030\mu\text{atm}$, $3.465\mu\text{atm}$ and $5.968\mu\text{atm}$ respectively. These values present a decrease of approximately 89%, 81% and 57.6% respectively from the mean

values of the MLR approach. This indicates that the mixed model approach produces a less biased model in terms of predictions of $f\text{CO}_2$ than the MLR approach. Although the MLR models provide less variable errors, as indicated by the lower standard deviations in the MAE and RMSE values obtained for the 100 repetitions of subset divisions of a 70%-30% split between training and test data subsets (Table 3.6), the increase in standard errors of the mixed model error rates is minimal. Comparing the maximum and minimum error values obtained from the models, it is seen that the maximum MSE, MAE and RMSE obtained from the mixed model is much smaller even than the minimum MSE, MAE and RMSE produced by the MLR model. This is strong evidence to suggest that the mixed model produces more accurate (i.e. less biased) predictions than the MLR approach.

The mixed models provide better predictions, in terms of a lower test error rate, than the MLR models, however this was also the case with the NPKR approach. The downside of these lower test error figures for the NPKR approach was a much higher variability in these error rates as indicated by the standard deviations ($100.479\mu\text{atm}$, $0.528\mu\text{atm}$ and $3.546\mu\text{atm}$ for the MSE, MAE and RMSE respectively) and COVs (62.8%, 10.2% and 29.2% for the MSE, MAE and RMSE respectively). These high standard deviation values and COV values were caused by the emergence of the “zero” estimates, which are the focus of the mixed models. It is therefore no surprise that the mixed model approach produces much smaller standard deviations and COVs than the NPKR models. The largest decreases are found in the standard deviations and COVs of the MSE (92.3% and 66% decrease respectively) and RMSE (81.7% and 62.7% decrease respectively), since these were the error values most heavily affected by the “zero” estimates. What is also of interest is the large decrease in maximum test error values for the MSE, MAE and RMSE of the mixed models ($58.439\mu\text{atm}$, $4.39\mu\text{atm}$ and $7.645\mu\text{atm}$ respectively) as compared to those of the NPKR models ($582.239\mu\text{atm}$, $6.381\mu\text{atm}$ and $24.130\mu\text{atm}$ respectively).

The maximum test errors from the mixed models, in fact, are only slightly larger than the minimum test error values produced by the NPKR approach. All of this once again indicates that the mixed model approach seems to not only produce more accurate results than the NPKR approach, but also less variable test error rates and therefore a better model.

Some of the reduction in the test errors indicated in Table 5.10 can be attributed to the inclusion of altimetry as an independent variable in the model, however since the curse of dimensionality is worsened by including more independent variables in the model, this cannot account for this magnitude of the decrease in the MSE and RMSE as indicated.

5.6 Conclusion

The topography of the SO is included in the MLR and NPKR models in terms of altimetry in order to further improve the predictive ability of the model. While the mixed model is used to address the problems observed in the MLR and NPKR approach by combining the predictions of $f\text{CO}_2$ from both models. The results indicate that these changes to the models improve the error rates observed for all subset divisions and, while model M11 is not able to improve on the prediction error rates of model M3, which includes salinity as a predictor variable rather than altimetry, the results are very close and indicate that model M11 is a good alternative to M3. The mixed model approach provides a better alternative to either the pure MLR or NPKR approaches. This is due to its lower observed prediction bias, as well as its solution to the “zero” estimates produced by some of the NPKR models (thereby reducing the variability of the test error rates). It is therefore suggested that the mixed model be the approach taken in further modelling (an estimation) of $f\text{CO}_2$ in the SO and that altimetry (i.e. a measure of sea surface topography) be included as an independent variable in all future models.

Chapter 6

Summary, conclusions and future research

6.1 Summary

The analysis of the *in situ* fCO₂ data from the SANAE49 ship travelling on Leg 6 of the journey back to Cape Town from Antarctica produces many interesting challenges and results. The objective of predictive model is to identify a procedure to reduce the average error rate in the predictions of fCO₂ to within 10% of the average fCO₂ observation. The multiple linear regression (MLR) and non-parametric kernel regression (NPKR) modelling approaches provide results which have both advantages and disadvantages and therefore a compromise between the two is preferred and found to outperform the individual approaches. This is the mixed model which combines the predictions of both models in order to solve the problems created by them. This chapter summarises the results of each of the previous chapters as well as draws conclusions as to what these results imply. Future research opportunities are also discussed in terms of what the impact of the work in this thesis.

6.1.1 Exploratory Analysis

The exploratory analysis provides a basis for the reasoning behind the initial ideas for a non-parametric regression approach instead of the parametric, MLR approach. This is due to the Figure 2.7 which indicates a non-normal distribution of the response variable $f\text{CO}_2$, along with the knowledge of domain experts who advised that the relationship between $f\text{CO}_2$ and its drivers is complex and varies inter- and intra-annually. The distribution of $f\text{CO}_2$, in fact, is multi-modal for this time period, which does not conform to regular, uni-modal, parametric distribution functions. The average $f\text{CO}_2$ value in the data set of SANAE49L6-final data set is $354.03\mu\text{atm}$. The error rates of the models developed are compared to this mean value in order to determine if the objective criterion stated in Section 1.2.1 is met by the method. The histogram of the chlorophyll-a concentrations indicated a wide spread of measurements with the majority of the values being between 0 and 1 micrograms per liter. As suggested in the paper by Jamet *et al.* (2007), the chlorophyll-a concentrations were transformed using a log transformation in order to remove the several orders of magnitude. The variables selected from the SANAE49L6-final data set for further models were SST, log chlorophyll-a concentration, salinity, MLD and latitude.

6.1.2 Multiple Linear Regression

The MLR approach to developing a model to predict $f\text{CO}_2$ was used as a comparison to the model developed by Jamet *et al.* (2007) for the North Atlantic oceans using MLR. The model parameters are estimated using least squares regression in order to minimise the in sample error (i.e. training sample error). This does not, however, guarantee a model which has a low out of sample error (i.e. test sample error). Models M1, M2 and M3, as described in Section 3.4, show a decrease in all 3 measures of test error (MSE, MAE and RMSE) when either of salinity or latitude is being included as an independent variable in

the model. Model M2 (which includes latitude) shows the largest reduction in prediction error. This could, however, be due to the data observations not being independent of one another especially with regards to their latitude position. Salinity, however, is not remotely available for areas of the ocean not physically sampled and therefore model M3 (which includes salinity) would not be useful for predicting $f\text{CO}_2$ for remotely sensed satellite data.

Model M1 was therefore selected to be simulated using varying divisions of the training and test data subsets in order to assess the effect of the amount of data in the training subset on the prediction ability of the model. From Table 3.4 it was seen that, for the MLR approach, the subset division seems to have little effect on the produced error rate. This implies that the MLR method is robust to small changes in the data. The downside to this is that the MLR approach assumes a strict form of the regression function, which may not be present in this, or future, data sets. Figures 3.1, 3.2 and 3.3 show that, while the model predictions do seem to fall short in certain areas, the MLR models capture the general form of data. The RMSE's obtained from these MLR models fell well within 10% of the average $f\text{CO}_2$ measurement, however they were all well above the values obtained by Jamet *et al.* (2007) in the North Atlantic.

Figures 3.4, 3.5 and 3.6 along with Table 3.6 indicated the concentration of the MLR error rates, which seem to be largely unaffected by the divisions of training and test data. The low variance of the error rates made MLR an attractive solution to the problem, however the strict assumption of the model along with its failure in certain areas of the SO demanded attention and therefore a more data dependent method was required.

6.1.3 Non-parametric kernel regression

The NPKR approach was adopted for two reasons: first, to solve the prediction bias issues of the MLR method in certain areas of the SO; and second,

to provide a model which makes less assumptions regarding the form of the regression function. The NPKR approach is a purely data driven model, as well as being an intuitively understandable and was therefore an attractive option.

The results obtained were initially optimistic. Table 4.2 as well as Figures 4.1, 4.2 and 4.3 all displayed a large reduction in the test error rates from the MLR approach, as well as much closer estimates of $f\text{CO}_2$ in all parts of the SO. These results, however, did not display the downfall of the NPKR approach. Upon inspection of Table 4.3 along with Figures 4.4 and 4.5 it became clear that the NPKR approach was very dependent, not only on how large the training data was, but also on which observations were included (as was seen in Figures 4.6, 4.7 and 4.8). This was due to the “curse of dimensionality” problem in the NPKR approach. The NPKR model requires large amounts of data to ensure that each combination of independent variables in the test data set has observations in the training data set which fall within their “near” neighbourhood as defined by the bandwidths.

Results also indicated that, although the NPKR method produced more accurate predictions of $f\text{CO}_2$ for the *in situ* data in the SO, the models developed using this method alone may not generalise well to the unsampled areas of the SO due to the curse of dimensionality. A solution was therefore sought in the form of the mixed model, which combined the predictions from both the MLR and NPKR approaches.

6.1.4 Including altimetry into the regression model

Both the MLR and NPKR approaches indicated that the inclusion of salinity as an independent variable in the model produced more accurate predictions of $f\text{CO}_2$. The issue here is that salinity is, as of yet, not reliably available via remote sensing for unsampled areas of the SO. This creates a problem for further research, since models including salinity could then not be able

to predict $f\text{CO}_2$ from remotely sensed predictor variables. For this reason, altimetry was introduced as an independent variable to capture some of the effect of salinity on $f\text{CO}_2$, since it was advised that a correlation exists between salinity and the sea surface topography (altimetry).

The results from the MLR and NPKR models M11 (which included SST, log chlorophyll-a concentration, MLD and altimetry as independent variables and was developed and assessed on a random 70%-30% division of training and test data subsets) were positive. The observed MSE, MAE and RMSE for both approaches displayed large decreases from model M1. For this reason it was decided that all future models should include altimetry as an independent variable, as it improved the predictive ability of the model. By including another variable, however, increases the chance of “zero” estimates in the NPKR approach exponentially (according to the curse of dimensionality) and therefore a method for eliminating these estimates was required.

6.1.5 Mixed regression model

The mixed regression model was proposed in order to solve both the bias seen in the MLR model predictions and the high variability in the NPKR error rates (due to the “zero” estimates). The method combined the predictions from both models by replacing any “zero” estimates produced by the NPKR model with the corresponding estimate from the MLR model. The MLR predictions were used as correction for the curse of dimensionality since these predictions, although displaying a larger bias than the NPKR predictions, captured the general form of the test data. From this definition of the mixed model, the error rates cannot be larger than those obtained from the NPKR model.

The mixed regression model results in Section 5.4.4 indicate the improvement of this method over both the MLR and NPKR approaches. Table 5.9 and Figures 5.3 and 5.4 show the effect of replacing the “zero” $f\text{CO}_2$ estimates with MLR predictions. The MSE, MAE and RMSE for the decreasing sizes

of training data subsets do not show the same increasing trend as was seen in Table 4.3 for the NPKR method. The histograms in Figures 5.5, 5.6 and 5.7 indicate the reduced bias in the model predictions as compared to the errors from the MLR approach. The MSEs, MAEs and RMSEs are all much lower than those obtained in Section 3.5.2.3. Also the error rates of the simulated subset divisions for the mixed regression model displays a smaller spread of the error rates as indicated by the standard deviations and COVs of the errors in Table 5.10.

These results indicate the improvement provided by the mixed model approach over the pure MLR or NPKR methods in predicting $f\text{CO}_2$ in the SO. It is therefore suggested that this be the method used in predicting the $f\text{CO}_2$ values for remotely sensed independent variables in unsampled areas of the SO.

6.2 Conclusion

Due to the reduced errors as well as the low variance of the simulated error rates observed, the mixed model is proposed as the method to be used in future to obtain $f\text{CO}_2$ estimates in the SO where *in situ* observations are not available. The independent variables that should be used are the SST, log transformed chlorophyll-a concentration, MLD and altimetry (sea surface topography). Further, a caveat remains that the data used to develop the model should not be used to predict $f\text{CO}_2$ data inter-annually or between seasons in a specific year. This is due to the complex and changing relationships between $f\text{CO}_2$ and its drivers inter- and intra-annually. The model should rather be refitted using *in situ* data available as close (in time and space) to the observations being predicted as possible.

6.3 Future research

The objective of this project was to identify a statistical method which could reliably predict $f\text{CO}_2$ from independent variables which can be remotely measured where *in situ* measurements are unavailable. Although a method has been identified (the mixed regression model), much future research must still be conducted in order to assess the feasibility of this method being applied to the entire SO.

6.3.1 Removal of spatial dependency

The regression models discussed in this thesis all assume complete independence of the observations in the data set (i.e. each observation is mutually independent of all the other observations). This is not strictly true for SANAE49L6-final, since there is a spatial dependence of each of the variables on all observations near to it. This implies that measurements of $f\text{CO}_2$ close to one another in space tend to be close to one another in value. In future, this spatial dependency should be incorporated in the model development stage. A method is, therefore, required to do this.

6.3.2 Small area modelling

Since the SO is a dynamic system, it seems logical that a single model may not be applicable for the entire ocean region. For this reason it may be helpful to model smaller areas of the ocean by creating boundaries, rather than using one model to predict $f\text{CO}_2$ for all unsampled areas. The difficulty in doing this lies in the definition of the boundaries since it is not straightforward to decide if the boundaries should be geographic (i.e. latitude and longitude) or if another, more flexible method should be used (e.g. frontal boundaries).

6.3.3 Expanding the model to remote sensing data

The ultimate objective of this study is to feed into the development of ocean carbon flux maps which will be able to identify possible sinks and sources of CO₂ in the SO and what their respective strengths are. This requires that the model developed on *in situ* data be used to predict fCO₂ in areas of the ocean where it is not available. For this, the predictive ability of the model using satellite measured (remote sensed) independent variables must be estimated. To do this, the satellite measured values of the independent variables corresponding in time and space to SANAE49 Leg 6 is required in order to compare the predicted fCO₂ values to the known values in SANAE49L6-final data set. This will give an estimate of the error rate along VOS lines. To expand from this, the global data base known as SOCAT of underway fCO₂ measurements could be used to compare the predicted fCO₂ values to. This is, however, only possible in certain areas where underway fCO₂ observations are available.

Appendix

R Code

Data cleaning

```
function (b = 5, combined.data.file)
{
  # Calling Data Set

  mld.data <- read.table(file = "mld_S49_all_leg6.csv", header = TRUE, sep = ",")
  nre.data.june2011 <- read.table(file = "SANAE49 Leg 6_pCO2 V2 13Var TypeEQU.csv",
  header = TRUE, sep = ",")
  nre.data.june2011[nre.data.june2011["pCO2W.H2OSST."]==(-9),"pCO2W.H2OSST."] <- NA

  # Deleting Rogue spike values - Refer to Notes

  nre.data.june2011 <- nre.data.june2011[-c(3001:3014,4604,6646:6653),]

  # Interpolating MLD values

  latlon <- mld.data[,c("Lat", "Lon")]
  colnames(latlon) <- c("latitude", "longitude")
  mld <- NULL

  ltmin <- nre.data.june2011[nre.data.june2011["latitude"]<min(latlon["latitude"]),]
  gtmax <- nre.data.june2011[nre.data.june2011["latitude"]>max(latlon["latitude"]),]

  reduced.pco2 <- nre.data.june2011[nre.data.june2011["latitude"]>min(latlon["latitude"])&
  nre.data.june2011["latitude"]<(-37)|nre.data.june2011["latitude"]==min(latlon["latitude"])|
  nre.data.june2011["latitude"]==(-37),]

  for(i in 1:nrow(reduced.pco2)){
```

```

loc <- reduced.pco2[i,c("latitude", "longitude")]
mld.pco2 <- rbind(loc, latlon)
dist.mat <- as.matrix(dist(mld.pco2))
distances <- dist.mat[-c(1),1]
min.dist <- which(distances == min(distances))
min.dist2 <- which(distances == min(distances[-min.dist]))
total.dist <- distances[min.dist] + distances[min.dist2]

mld[i] <- (distances[min.dist2]/total.dist)*mld.data[min.dist,"MLD.m."] +
(distances[min.dist]/total.dist)*mld.data[min.dist2,"MLD.m."]
}

full.data <- cbind(reduced.pco2,mld)
attach(full.data)

detach(full.data)
write.csv(cbind(reduced.pco2,mld), file=combined.data.file)
}

```

Exploratory Analysis

```

function (comb.fco2.mld.data, b = 2, c = 12)
{

final.data <- comb.fco2.mld.data

# Plotting covariates

plot(final.data[, "latitude"], final.data[, "fCO2.Water"], type = "l", col = "blue",
main = c("SANA E 49 L6", "Plot of fCO2(Water) and MLD"), xlab = "Latitude",
ylab = "fCO2(Water)", ylim = c(240, 450))
par(new = T)
plot(final.data[, "latitude"], final.data[, "MLD"], type = "l", col = "red", main = "", xlab = "",
ylab = "", axes = F, ylim = c(15,125))
mtext("MLD",side=4)
axis(4, ylim=c(15,125))
legend(locator(1), legend = c("fCO2(Water)", "MLD"), lty = 1, col = c("blue", "red"))

windows()

```

APPENDIX
R CODE

125

```

plot(final.data[,"latitude"], final.data[,"Ch.conc"], type = "l", col = "green",
main = c("SANAE 49 L6", "Plot of Chlorophyll Concentration and pH"), xlab = "Latitude",
ylab = "Chlorophyll Concentration", ylim = c(0, 5.5))
par(new = T)
plot(final.data[,"latitude"], final.data[,"pH"], type = "l", col = "orange", main = "",
xlab = "", ylab = "", axes = F, ylim = c(6.9,7.3))
mtext("pH",side=4)
axis(4, ylim=c(6.9,7.3))
legend(locator(1), legend = c("Chlorophyll Concentration", "pH"), lty = 1,
col = c("green", "orange"))

windows()

plot(final.data[,"latitude"], final.data[,"O2.....sat."], type = "l", col = "blue",
main = c("SANAE 49 L6", "Plot of O2(sat) and O2(ppm)"), xlab = "Latitude",
ylab = "O2(sat)", ylim = c(70, 95))
par(new = T)
plot(final.data[,"latitude"], final.data[,"O2.ppm."], type = "l", col = "cyan", main = "",
xlab = "", ylab = "", axes = F, ylim = c(5.5,12.5))
mtext("O2(ppm)",side=4)
axis(4, ylim=c(5.5,12.5))
legend(locator(1), legend = c("O2(sat)", "O2(ppm)"), lty = 1, col = c("blue", "cyan"))

windows()

plot(final.data[,"latitude"], final.data[,"Salinity"], type = "l", col = "purple",
main = c("SANAE 49 L6", "Plot of Salinity and Intake Temperature"), xlab = "Latitude",
ylab = "Salinity", ylim = c(33, 36))
par(new = T)
plot(final.data[,"latitude"], final.data[,"Intake.Temp"], type = "l", col = "black", main = "",
xlab = "", ylab = "", axes = F, ylim = c(-5,25))
mtext("Intake Temperature",side=4)
axis(4, ylim=c(-5,25))
legend(locator(1), legend = c("Salinity", "Intake Temperature"), lty = 1, col = c("purple", "black"))

# Covariate Histograms

hist(final.data[,"fCO2.Water"], breaks = seq(200, 450, by = b), main = c("SANAE49L6", "fCO2(Water)"),
xlab = "fCO2(Water)", ylab = "Frequency", col = "light green")

windows()

```

APPENDIX
R CODE

126

```
hist(final.data[, "Ch.conc"], main = c("SANAE49L6", "Chlorophyll Concentration"),
     xlab = "Chlorophyll Concentration", ylab = "Frequency", col = "green")

windows()

hist(final.data[, "Intake.Temp"], main = c("SANAE49L6", "Intake Temperature"),
     xlab = "Intake Temperature", ylab = "Frequency")

windows()

hist(final.data[, "latitude"], breaks = seq(-75, -30, by = 2), main = c("SANAE49L6", "Latitude"),
     xlab = "Latitude", ylab = "Frequency", col = "brown")

windows()

hist(final.data[, "MLD"], main = c("SANAE49L6", "MLD"), xlab = "MLD", ylab = "Frequency",
     col = "red")

windows()

hist(final.data[, "O2.ppm."], main = c("SANAE49L6", "O2(ppm)"), xlab = "O2(ppm)",
     ylab = "Frequency", col = "cyan")

windows()

hist(final.data[, "O2.....sat."], main = c("SANAE49L6", "O2(sat)"), xlab = "O2(sat)",
     ylab = "Frequency", col = "blue")

windows()

hist(final.data[, "pH"], main = c("SANAE49L6", "pH"), xlab = "pH", ylab = "Frequency",
     col = "yellow")

windows()

hist(final.data[, "Salinity"], main = c("SANAE49L6", "Salinity"), xlab = "Salinity",
     ylab = "Frequency", col = "purple")

# Calculation of discriptive statistics

discr.stat <- final.data[,c("fCO2.Water", "Salinity", "O2.....sat.", "O2.ppm.", "pH",
"Ch.conc", "Intake.Temp", "MLD")]
```



```

n.func <- function(vec){length(vec[is.na(vec) == FALSE])}
n.missing.func <- function(vec){length(vec[vec=="NA"])}

n <- apply(discr.stat, 2, n.func)
n.missing <- apply(discr.stat, 2, n.missing.func)
Means <- apply(discr.stat, 2, mean, na.rm = TRUE)
SD <- apply(discr.stat, 2, sd, na.rm = TRUE)
Mins <- apply(discr.stat, 2, min, na.rm = TRUE)
Maxs <- apply(discr.stat, 2, max, na.rm = TRUE)
Q1 <- apply(discr.stat, 2, quantile, prob = 0.25, na.rm = TRUE)
Median <- apply(discr.stat, 2, quantile, prob = 0.5, na.rm = TRUE)
Q3 <- apply(discr.stat, 2, quantile, prob = 0.75, na.rm = TRUE)

expl.data <- cbind(n, n.missing, Means, SD, Mins, Maxs, Q1, Median, Q3)

list("Descriptive Statistics" = expl.data)
}

```

Multiple linear regression

Models M1 to M10

```

function (comb.fco2.mld.data, seed = 5000)
{
  set.seed(seed)

  # Reading Data in

  co2.data <- comb.fco2.mld.data
  co2.data <- co2.data[-c(4353, 4354), c("latitude", "longitude", "Salinity", "Ch.conc",
  "Intake.Temp", "pCO2W.H2OSST.", "MLD", "fCO2.Water")]
  log.chl <- log10(co2.data[, "Ch.conc"])
  co2.data <- cbind(co2.data, log.chl)

  #Dividing data into Training and Test sets

  split <- runif(nrow(co2.data),0,1)
  training80 <- which(split > 0.2)
  training70 <- which(split > 0.3)

```

APPENDIX
R CODE

128

```
training60 <- which(split > 0.4)
training50 <- which(split > 0.5)
training40 <- which(split > 0.6)
training30 <- which(split > 0.7)
training20 <- which(split > 0.8)

test80 <- which(split < 0.2 | split == 0.2)
test70 <- which(split < 0.3 | split == 0.3)
test60 <- which(split < 0.4 | split == 0.4)
test50 <- which(split < 0.5 | split == 0.5)
test40 <- which(split < 0.6 | split == 0.6)
test30 <- which(split < 0.7 | split == 0.7)
test20 <- which(split < 0.8 | split == 0.8)

train80.data <- co2.data[training80,]
train70.data <- co2.data[training70,]
train60.data <- co2.data[training60,]
train50.data <- co2.data[training50,]
train40.data <- co2.data[training40,]
train30.data <- co2.data[training30,]
train20.data <- co2.data[training20,]

test80.data <- co2.data[test80,]
test70.data <- co2.data[test70,]
test60.data <- co2.data[test60,]
test50.data <- co2.data[test50,]
test40.data <- co2.data[test40,]
test30.data <- co2.data[test30,]
test20.data <- co2.data[test20,]

# Standardizing the training data

means.train <- apply(train70.data, 2, mean)
sd.train <- apply(train70.data, 2, sd)
means.test <- apply(test70.data, 2, mean)
sd.test <- apply(test70.data, 2, sd)

train.standard <- matrix(0, ncol = ncol(train70.data), nrow = nrow(train70.data))

for(i in 1:ncol(train70.data)){
  train.standard[,i] <- (train70.data[,i] - means.train[i])/sd.train[i]}
colnames(train.standard) <- c("latitude", "longitude", "Salinity", "Ch.conc", "Intake.Temp",
```

APPENDIX
R CODE

129

```

"pCO2W.H2OSSST.", "MLD", "fCO2.Water", "log.chl")
train.standard <- as.data.frame(train.standard)

# Model Building (fCO2)

mlr.M1.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train70.data)
mlr.M2.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + latitude,
data = train70.data)
mlr.M3.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Salinity,
data = train70.data)
mlr.M4.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train80.data)
mlr.M5.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train60.data)
mlr.M6.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train50.data)
mlr.M7.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train40.data)
mlr.M8.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train30.data)
mlr.M9.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train20.data)
mlr.M10.fco2 <- lm(formula = fCO2.Water ~ 0 + Intake.Temp + log.chl + MLD, data = train.standard)

# Model Parameters (fCO2)

mlr.parameter.M1 <- mlr.M1.fco2$coef
mlr.parameter.M2 <- mlr.M2.fco2$coef
mlr.parameter.M3 <- mlr.M3.fco2$coef
mlr.parameter.M4 <- mlr.M4.fco2$coef
mlr.parameter.M5 <- mlr.M5.fco2$coef
mlr.parameter.M6 <- mlr.M6.fco2$coef
mlr.parameter.M7 <- mlr.M7.fco2$coef
mlr.parameter.M8 <- mlr.M8.fco2$coef
mlr.parameter.M9 <- mlr.M9.fco2$coef
mlr.parameter.M10 <- mlr.M10.fco2$coef

# Standardizing the test data

test.standard.trainstats <- matrix(0, ncol = ncol(test70.data), nrow = nrow(test70.data))
test.standard.teststats <- matrix(0, ncol = ncol(test70.data), nrow = nrow(test70.data))

for(j in 1:ncol(test70.data)){
test.standard.trainstats[,j] <- (test70.data[,j] - means.train[j])/sd.train[j]}
for(k in 1:ncol(test70.data)){
test.standard.teststats[,j] <- (test70.data[,j] - means.test[j])/sd.test[j]}
colnames(test.standard.trainstats) <- c("latitude", "longitude", "Salinity", "Ch.conc", "Intake.Temp",
"pCO2W.H2OSSST.", "MLD", "fCO2.Water", "log.chl")

```

APPENDIX
R CODE

130

```
colnames(test.standard.teststats) <- c("latitude", "longitude", "Salinity", "Ch.conc", "Intake.Temp",
"pCO2W.H2OSS.T.", "MLD", "fCO2.Water", "log.chl")
```

```
test.standard.trainstats <- as.data.frame(test.standard.trainstats)
```

```
test.standard.teststats <- as.data.frame(test.standard.teststats)
```

```
# Model predictions fCO2
```

```
M1.predict.fCO2 <- predict(mlr.M1.fco2, newdata = test70.data)
```

```
M2.predict.fCO2 <- predict(mlr.M2.fco2, newdata = test70.data)
```

```
M3.predict.fCO2 <- predict(mlr.M3.fco2, newdata = test70.data)
```

```
M4.predict.fCO2 <- predict(mlr.M4.fco2, newdata = test80.data)
```

```
M5.predict.fCO2 <- predict(mlr.M5.fco2, newdata = test60.data)
```

```
M6.predict.fCO2 <- predict(mlr.M6.fco2, newdata = test50.data)
```

```
M7.predict.fCO2 <- predict(mlr.M7.fco2, newdata = test40.data)
```

```
M8.predict.fCO2 <- predict(mlr.M8.fco2, newdata = test30.data)
```

```
M9.predict.fCO2 <- predict(mlr.M9.fco2, newdata = test20.data)
```

```
M10.predict.fCO2.trainstats <-
```

```
(predict(mlr.M10.fco2, newdata = test.standard.trainstats)*sd.train[8]) + means.train[8]
```

```
M10.predict.fCO2.teststats <-
```

```
(predict(mlr.M10.fco2, newdata = test.standard.teststats)*sd.test[8]) + means.test[8]
```

```
# Model MSE's fCO2
```

```
M1.MSE.fCO2 <- sum((M1.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
```

```
M2.MSE.fCO2 <- sum((M2.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
```

```
M3.MSE.fCO2 <- sum((M3.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
```

```
M4.MSE.fCO2 <- sum((M4.predict.fCO2 - test80.data[, "fCO2.Water"])^2)/nrow(test80.data)
```

```
M5.MSE.fCO2 <- sum((M5.predict.fCO2 - test60.data[, "fCO2.Water"])^2)/nrow(test60.data)
```

```
M6.MSE.fCO2 <- sum((M6.predict.fCO2 - test50.data[, "fCO2.Water"])^2)/nrow(test50.data)
```

```
M7.MSE.fCO2 <- sum((M7.predict.fCO2 - test40.data[, "fCO2.Water"])^2)/nrow(test40.data)
```

```
M8.MSE.fCO2 <- sum((M8.predict.fCO2 - test30.data[, "fCO2.Water"])^2)/nrow(test30.data)
```

```
M9.MSE.fCO2 <- sum((M9.predict.fCO2 - test20.data[, "fCO2.Water"])^2)/nrow(test20.data)
```

```
M10.MSE.fCO2.trainstats <-
```

```
sum((M10.predict.fCO2.trainstats - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
```

```
M10.MSE.fCO2.teststats <-
```

```
sum((M10.predict.fCO2.teststats - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
```

```
fCO2.models.MSE <- c(M1.MSE.fCO2, M2.MSE.fCO2, M3.MSE.fCO2, M4.MSE.fCO2, M5.MSE.fCO2,
```

```
M6.MSE.fCO2, M7.MSE.fCO2, M8.MSE.fCO2, M9.MSE.fCO2, M10.MSE.fCO2.trainstats,
```

```
M10.MSE.fCO2.teststats)
```

```
#Model MAE's fCO2
```

APPENDIX
R CODE

131

```

M1.MAE.fCO2 <- sum(abs(M1.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
M2.MAE.fCO2 <- sum(abs(M2.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
M3.MAE.fCO2 <- sum(abs(M3.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
M4.MAE.fCO2 <- sum(abs(M4.predict.fCO2 - test80.data[, "fCO2.Water"]))/nrow(test80.data)
M5.MAE.fCO2 <- sum(abs(M5.predict.fCO2 - test60.data[, "fCO2.Water"]))/nrow(test60.data)
M6.MAE.fCO2 <- sum(abs(M6.predict.fCO2 - test50.data[, "fCO2.Water"]))/nrow(test50.data)
M7.MAE.fCO2 <- sum(abs(M7.predict.fCO2 - test40.data[, "fCO2.Water"]))/nrow(test40.data)
M8.MAE.fCO2 <- sum(abs(M8.predict.fCO2 - test30.data[, "fCO2.Water"]))/nrow(test30.data)
M9.MAE.fCO2 <- sum(abs(M9.predict.fCO2 - test20.data[, "fCO2.Water"]))/nrow(test20.data)
M10.MAE.fCO2.trainstats <-
sum(abs(M10.predict.fCO2.trainstats - test70.data[, "fCO2.Water"]))/nrow(test70.data)
M10.MAE.fCO2.teststats <-
sum(abs(M10.predict.fCO2.teststats - test70.data[, "fCO2.Water"]))/nrow(test70.data)
fCO2.models.MAE <- c(M1.MAE.fCO2, M2.MAE.fCO2, M3.MAE.fCO2, M4.MAE.fCO2, M5.MAE.fCO2,
M6.MAE.fCO2, M7.MAE.fCO2, M8.MAE.fCO2, M9.MAE.fCO2, M10.MAE.fCO2.trainstats,
M10.MAE.fCO2.teststats)

# Error Tables

fCO2.error.table <- cbind(fCO2.models.MSE, fCO2.models.MAE)
colnames(fCO2.error.table) <- c("Mean Square Error", "Mean Absolute Error")
rownames(fCO2.error.table) <- c("M1", "M2", "M3", "M4", "M5", "M6", "M7", "M8", "M9",
"M10 (Training Stats)", "M10 (Test Stats)")

cat("fCO2 Error Rate Table")
print(fCO2.error.table)

train.test.split.MSE.fco2 <- c(M4.MSE.fCO2, M1.MSE.fCO2, M5.MSE.fCO2, M6.MSE.fCO2,
M7.MSE.fCO2, M8.MSE.fCO2, M9.MSE.fCO2)
train.test.split.MAE.fco2 <- c(M4.MAE.fCO2, M1.MAE.fCO2, M5.MAE.fCO2, M6.MAE.fCO2,
M7.MAE.fCO2, M8.MAE.fCO2, M9.MAE.fCO2)

# Prediction Plots (fCO2)

plot(x = test70.data[, "latitude"], y = test70.data[, "fCO2.Water"],
main = "Predictive Plot of fCO2 from MLR Model M1", xlab = "Latitude", ylab = "fCO2",
pch = ".", col = "blue")
lines(x = test70.data[, "latitude"], y = M1.predict.fCO2, lty = 1, col = "red")

windows()

plot(x = test70.data[, "latitude"], y = test70.data[, "fCO2.Water"],

```

```

main = "Predictive Plot of fCO2 from MLR Model M2", xlab = "Latitude", ylab = "fCO2",
pch = ".", col = "blue")
lines(x = test70.data[, "latitude"], y = M2.predict.fCO2, lty = 1, col = "red")

windows()

plot(x = test70.data[, "latitude"], y = test70.data[, "fCO2.Water"],
main = "Predictive Plot of fCO2 from MLR Model M3", xlab = "Latitude", ylab = "fCO2",
pch = ".", col = "blue")
lines(x = test70.data[, "latitude"], y = M3.predict.fCO2, lty = 1, col = "red")

# Training - Test Split Error Histograms

windows()

barplot(height = train.test.split.MSE.fco2, space = 0.2,
names.arg = c("M4", "M1", "M5", "M6", "M7", "M8", "M9"), col = c("green", "orange"),
ylim = c(0, 450), main = "Subset Division Mean Square Errors",
xlab = "Training - Test % Split", ylab = "Mean Square Error")

windows()

barplot(height = train.test.split.MAE.fco2, space = 0.2,
names.arg = c("M4", "M1", "M5", "M6", "M7", "M8", "M9"), col = c("blue", "violet"),
ylim = c(0, 20), main = "Subset Division Mean Absolute Errors",
xlab = "Training - Test % Split", ylab = "Mean Absolute Error")

windows()

barplot(height = sqrt(train.test.split.MSE.fco2), space = 0.2,
names.arg = c("M4", "M1", "M5", "M6", "M7", "M8", "M9"), col = c("black", "red"),
ylim = c(0, 20), main = "Subset Division Root Mean Square Errors",
xlab = "Training - Test % Split", ylab = "Root Mean Square Error")
}

```

MLR error simulation

```

function (comb.fco2.mld.data, reps = 100, mlr.reps.error.file, mlr.)
{

co2.data <- comb.fco2.mld.data
co2.data <- co2.data[-c(4353, 4354), c("latitude", "longitude", "Salinity",

```

APPENDIX
R CODE

133

```
"Ch.conc", "Intake.Temp", "pCO2W.H2OSST.", "MLD", "fCO2.Water"]]  
log.chl <- log10(co2.data[, "Ch.conc"])  
co2.data <- cbind(co2.data, log.chl)  
  
M1.MSE.fCO2 <- NULL  
M1.MAE.fCO2 <- NULL  
  
for(i in 1:reps){  
  # Dividing data into Training and Test sets  
  
  split <- runif(nrow(co2.data),0,1)  
  training70 <- which(split > 0.3)  
  
  test70 <- which(split < 0.3 | split == 0.3)  
  
  train70.data <- co2.data[training70,]  
  
  test70.data <- co2.data[test70,]  
  
  # Model Building(fCO2)  
  
  mlr.M1.fco2 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train70.data)  
  # Model Parameters (fCO2)  
  
  mlr.parameter.M1 <- mlr.M1.fco2$coef  
  
  print(mlr.parameter.M1)  
  
  # Model predictions fCO2  
  
  M1.predict.fCO2 <- predict(mlr.M1.fco2, newdata = test70.data)  
  
  # Model Errors  
  M1.MSE.fCO2[i] <- sum((M1.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)  
  M1.MAE.fCO2[i] <- sum(abs(M1.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)  
  
  }  
  
MLR.error.table <- cbind(M1.MSE.fCO2, sqrt(M1.MSE.fCO2), M1.MAE.fCO2)  
colnames(MLR.error.table) <- c("MLR Mean Square Errors", "MLR Root Mean Square Errors",  
"MLR Mean Absolute Errors")
```

```

write.csv(MLR.error.table, file = mlr.reps.error.file)

hist(M1.MSE.fCO2, main = c("MLR Mean Square Errors", "100 Repetitions"),
xlab = "Mean Square Error", ylab = "Frequency", xlim = c(0, 3500), ylim = c(0, 100),
breaks = seq(0, 3500, by = 100), col = c("green", "orange"))

windows()

hist(sqrt(M1.MSE.fCO2), main = c("MLR Root Mean Square Errors", "100 Repetitions"),
xlab = "Root Mean Square Error", ylab = "Frequency", xlim = c(0, 60), ylim = c(0, 100),
breaks = seq(0, 60, by = 2), col = c("red", "black"))

windows()

hist(M1.MAE.fCO2, main = c("MLR Mean Absolute Errors", "100 Repetitions"),
xlab = "Mean Absolute Error", ylab = "Frequency", xlim = c(0, 20), ylim = c(0, 100),
breaks = seq(0, 20, by = 1), col = c("blue", "purple"))
}

```

Non-parametric kernel regression

Models M1 - M10

```

function (comb.fco2.mld.data, seed = 5000)
{
  set.seed(seed)
  library(np)

  # Reading Data in
  co2.data <- comb.fco2.mld.data
  co2.data <- co2.data[-c(4353, 4354), c("latitude", "longitude", "Salinity",
"Ch.conc", "Intake.Temp", "pCO2W.H2OSST.", "MLD", "fCO2.Water")]
  log.chl <- log10(co2.data[, "Ch.conc"])
  co2.data <- cbind(co2.data, log.chl)

  # Dividing data into Training and Test sets
  split <- runif(nrow(co2.data), 0, 1)
  training80 <- which(split > 0.2)
  training70 <- which(split > 0.3)
  training60 <- which(split > 0.4)

```


APPENDIX
R CODE

135

```
training50 <- which(split > 0.5)
training40 <- which(split > 0.6)
training30 <- which(split > 0.7)
training20 <- which(split > 0.8)

test80 <- which(split < 0.2 | split == 0.2)
test70 <- which(split < 0.3 | split == 0.3)
test60 <- which(split < 0.4 | split == 0.4)
test50 <- which(split < 0.5 | split == 0.5)
test40 <- which(split < 0.6 | split == 0.6)
test30 <- which(split < 0.7 | split == 0.7)
test20 <- which(split < 0.8 | split == 0.8)

train80.data <- co2.data[training80,]
train70.data <- co2.data[training70,]
train60.data <- co2.data[training60,]
train50.data <- co2.data[training50,]
train40.data <- co2.data[training40,]
train30.data <- co2.data[training30,]
train20.data <- co2.data[training20,]

test80.data <- co2.data[test80,]
test70.data <- co2.data[test70,]
test60.data <- co2.data[test60,]
test50.data <- co2.data[test50,]
test40.data <- co2.data[test40,]
test30.data <- co2.data[test30,]
test20.data <- co2.data[test20,]

# Standardizing the training data
means.train <- apply(train70.data, 2, mean)
sd.train <- apply(train70.data, 2, sd)
means.test <- apply(test70.data, 2, mean)
sd.test <- apply(test70.data, 2, sd)

train.standard <- matrix(0, ncol = ncol(train70.data), nrow = nrow(train70.data))

for(i in 1:ncol(train70.data)){
train.standard[,i] <- (train70.data[,i] - means.train[i])/sd.train[i]}
colnames(train.standard) <- c("latitude", "longitude", "Salinity", "Ch.conc",
"Intake.Temp", "pCO2W.H2OSSST.", "MLD", "fCO2.Water", "log.chl")
train.standard <- as.data.frame(train.standard)
```

```
# Model Building(fCO2)
np.M1.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train70.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M1.fco2 <- npreg(bws = np.M1.fco2.bands)

print("Model M1 Done")

np.M2.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + lati-
tude,
data = train70.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M2.fco2 <- npreg(bws = np.M2.fco2.bands)

print("Model M2 Done")

np.M3.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Salinity,
data = train70.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M3.fco2 <- npreg(bws = np.M3.fco2.bands)

print("Model M3 Done")

np.M4.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train80.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M4.fco2 <- npreg(bws = np.M4.fco2.bands)

print("Model M4 Done")

np.M5.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train60.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M5.fco2 <- npreg(bws = np.M5.fco2.bands)

print("Model M5 Done")

np.M6.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train50.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M6.fco2 <- npreg(bws = np.M6.fco2.bands)
```

APPENDIX
R CODE

137

```

print("Model M6 Done")

np.M7.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train40.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M7.fco2 <- npreg(bws = np.M7.fco2.bands)

print("Model M7 Done")

np.M8.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train30.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M8.fco2 <- npreg(bws = np.M8.fco2.bands)

print("Model M8 Done")

np.M9.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train20.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M9.fco2 <- npreg(bws = np.M9.fco2.bands)

print("Model M9 Done")

np.M10.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train.standard, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
np.M10.fco2 <- npreg(bws = np.M10.fco2.bands)

print("Model M10 Done")

# Standardizing the test data
test.standard.trainstats <- matrix(0, ncol = ncol(test70.data), nrow = nrow(test70.data))
test.standard.teststats <- matrix(0, ncol = ncol(test70.data), nrow = nrow(test70.data))

for(j in 1:ncol(test70.data)){
test.standard.trainstats[,j] <- (test70.data[,j] - means.train[j])/sd.train[j]}
for(k in 1:ncol(test70.data)){
test.standard.teststats[,j] <- (test70.data[,j] - means.test[j])/sd.test[j]}
colnames(test.standard.trainstats) <- c("latitude", "longitude", "Salinity",
"Ch.conc", "Intake.Temp", "pCO2W.H2OSST.", "MLD", "fCO2.Water", "log.chl")
colnames(test.standard.teststats) <- c("latitude", "longitude", "Salinity",

```

```

"Ch.conc", "Intake.Temp", "pCO2W.H2OSST.", "MLD", "fCO2.Water", "log.chl")

test.standard.trainstats <- as.data.frame(test.standard.trainstats)
test.standard.teststats <- as.data.frame(test.standard.teststats)

# Model predictions fCO2
M1.predict.fCO2 <- predict(np.M1.fco2, newdata = test70.data)
M2.predict.fCO2 <- predict(np.M2.fco2, newdata = test70.data)
M3.predict.fCO2 <- predict(np.M3.fco2, newdata = test70.data)
M4.predict.fCO2 <- predict(np.M4.fco2, newdata = test80.data)
M5.predict.fCO2 <- predict(np.M5.fco2, newdata = test60.data)
M6.predict.fCO2 <- predict(np.M6.fco2, newdata = test50.data)
M7.predict.fCO2 <- predict(np.M7.fco2, newdata = test40.data)
M8.predict.fCO2 <- predict(np.M8.fco2, newdata = test30.data)
M9.predict.fCO2 <- predict(np.M9.fco2, newdata = test20.data)
M10.predict.fCO2.trainstats <-
(predict(np.M10.fco2, newdata = test.standard.trainstats)*sd.train[6]) + means.train[6]
M10.predict.fCO2.teststats <-
(predict(np.M10.fco2, newdata = test.standard.teststats)*sd.test[6]) + means.test[6]

# Model MSE's fCO2
M1.MSE.fCO2 <- sum((M1.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
M2.MSE.fCO2 <- sum((M2.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
M3.MSE.fCO2 <- sum((M3.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
M4.MSE.fCO2 <- sum((M4.predict.fCO2 - test80.data[, "fCO2.Water"])^2)/nrow(test80.data)
M5.MSE.fCO2 <- sum((M5.predict.fCO2 - test60.data[, "fCO2.Water"])^2)/nrow(test60.data)
M6.MSE.fCO2 <- sum((M6.predict.fCO2 - test50.data[, "fCO2.Water"])^2)/nrow(test50.data)
M7.MSE.fCO2 <- sum((M7.predict.fCO2 - test40.data[, "fCO2.Water"])^2)/nrow(test40.data)
M8.MSE.fCO2 <- sum((M8.predict.fCO2 - test30.data[, "fCO2.Water"])^2)/nrow(test30.data)
M9.MSE.fCO2 <- sum((M9.predict.fCO2 - test20.data[, "fCO2.Water"])^2)/nrow(test20.data)
M10.MSE.fCO2.trainstats <-
sum((M10.predict.fCO2.trainstats - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
M10.MSE.fCO2.teststats <-
sum((M10.predict.fCO2.teststats - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
fCO2.models.MSE <- c(M1.MSE.fCO2, M2.MSE.fCO2, M3.MSE.fCO2, M4.MSE.fCO2, M5.MSE.fCO2,
M6.MSE.fCO2, M7.MSE.fCO2, M8.MSE.fCO2, M9.MSE.fCO2, M10.MSE.fCO2.trainstats,
M10.MSE.fCO2.teststats)

# Model MAE's fCO2
M1.MAE.fCO2 <- sum(abs(M1.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
M2.MAE.fCO2 <- sum(abs(M2.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
M3.MAE.fCO2 <- sum(abs(M3.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)

```

APPENDIX
R CODE

139

```

M4.MAE.fCO2 <- sum(abs(M4.predict.fCO2 - test80.data[, "fCO2.Water"]))/nrow(test80.data)
M5.MAE.fCO2 <- sum(abs(M5.predict.fCO2 - test60.data[, "fCO2.Water"]))/nrow(test60.data)
M6.MAE.fCO2 <- sum(abs(M6.predict.fCO2 - test50.data[, "fCO2.Water"]))/nrow(test50.data)
M7.MAE.fCO2 <- sum(abs(M7.predict.fCO2 - test40.data[, "fCO2.Water"]))/nrow(test40.data)
M8.MAE.fCO2 <- sum(abs(M8.predict.fCO2 - test30.data[, "fCO2.Water"]))/nrow(test30.data)
M9.MAE.fCO2 <- sum(abs(M9.predict.fCO2 - test20.data[, "fCO2.Water"]))/nrow(test20.data)
M10.MAE.fCO2.trainstats <-
sum(abs(M10.predict.fCO2.trainstats - test70.data[, "fCO2.Water"]))/nrow(test70.data)
M10.MAE.fCO2.teststats <-
sum(abs(M10.predict.fCO2.teststats - test70.data[, "fCO2.Water"]))/nrow(test70.data)
fCO2.models.MAE <- c(M1.MAE.fCO2, M2.MAE.fCO2, M3.MAE.fCO2, M4.MAE.fCO2, M5.MAE.fCO2,
M6.MAE.fCO2, M7.MAE.fCO2, M8.MAE.fCO2, M9.MAE.fCO2, M10.MAE.fCO2.trainstats,
M10.MAE.fCO2.teststats)

# Error Table
fCO2.error.table <- cbind(fCO2.models.MSE, fCO2.models.MAE)
colnames(fCO2.error.table) <- c("Mean Square Error", "Mean Absolute Error")
rownames(fCO2.error.table) <- c("M1", "M2", "M3", "M4", "M5", "M6", "M7", "M8", "M9",
"M10 (Training Stats)", "M10 (Test Stats)")

train.test.split.MSE.fco2 <- c(M4.MSE.fCO2, M1.MSE.fCO2, M5.MSE.fCO2, M6.MSE.fCO2,
M7.MSE.fCO2, M8.MSE.fCO2, M9.MSE.fCO2)
train.test.split.MAE.fco2 <- c(M4.MAE.fCO2, M1.MAE.fCO2, M5.MAE.fCO2, M6.MAE.fCO2,
M7.MAE.fCO2, M8.MAE.fCO2, M9.MAE.fCO2)

# Barplot of Errors from Subset Divisions
barplot(height = train.test.split.MSE.fco2, space = 0.2, names.arg = c("80-20", "70-30",
"60-40", "50-50", "40-60", "30-70", "20-80"), col = c("green", "orange"), ylim = c(0, 450),
main = c("Nonparametric Regression", "fCO2 Mean Square Errors"),
xlab = "Training - Test % Split", ylab = "MSE")

windows()

barplot(height = train.test.split.MAE.fco2, space = 0.2, names.arg = c("80-20", "70-30",
"60-40", "50-50", "40-60", "30-70", "20-80"), col = c("blue", "violet"), ylim = c(0, 20),
main = c("Nonparametric Regression", "fCO2 Mean Absolute Errors"),
xlab = "Training - Test % Split", ylab = "MAE")

write.csv(fCO2.error.table, file = "NEW NP models MSE (fCO2).csv")

list("M1 Bandwidths" = np.M1.fco2.bands$bw, "M2 Bandwidths" = np.M2.fco2.bands$bw,
"M3 Bandwidths" = np.M3.fco2.bands$bw, "M4 Bandwidths" = np.M4.fco2.bands$bw,

```

```
"M5 Bandwidths" = np.M5.fco2.bands$bw, "M6 Bandwidths" = np.M6.fco2.bands$bw,
"M7 Bandwidths" = np.M7.fco2.bands$bw, "M8 Bandwidths" = np.M8.fco2.bands$bw,
"M9 Bandwidths" = np.M9.fco2.bands$bw, "M10 Bandwidths" = np.M10.fco2.bands$bw)
}
```

NPKR error simulation

```
function (comb.fco2.mld.data, reps = 50, npkr.rep.bands.file, npkr.rep.errors.file)
{
  library(np)

  co2.data <- comb.fco2.mld.data
  co2.data <- co2.data[-c(4353, 4354), c("latitude", "longitude", "Salinity", "Ch.conc",
    "Intake.Temp", "pCO2W.H2OSST.", "MLD", "fCO2.Water")]
  log.chl <- log10(co2.data[, "Ch.conc"])
  co2.data <- cbind(co2.data, log.chl)

  model.errors <- matrix(0, ncol = 3, nrow = reps)
  model.bands <- matrix(0, ncol = 3, nrow = reps)

  for(i in 1:reps){
    # Dividing data into Training and Test sets
    split <- runif(nrow(co2.data),0,1)

    training70 <- which(split > 0.3)

    test70 <- which(split < 0.3 | split == 0.3)

    train70.data <- co2.data[training70,]

    test70.data <- co2.data[test70,]

    # Model Building (fCO2)
    np.M1.fco2.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
      data = train70.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
      na.action = na.omit)
    np.M1.fco2 <- npreg(bws = np.M1.fco2.bands)

    # Model predictions fCO2
    M1.predict.fCO2 <- predict(np.M1.fco2, newdata = test70.data)

    # Model MSE's fCO2
```

```

M1.MSE.fCO2 <- sum((M1.predict.fCO2 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)

# Model MAE's fCO2
M1.MAE.fCO2 <- sum(abs(M1.predict.fCO2 - test70.data[, "fCO2.Water"]))/nrow(test70.data)

model.errors[i,] <- c(M1.MSE.fCO2, M1.MAE.fCO2, sqrt(M1.MSE.fCO2))
model.bands[i,] <- np.M1.fco2.bands$bw
print(paste("Model", i, "Done"))
}

write.csv(model.bands, file = npkr.rep.bands.file)
write.csv(model.errors, file = npkr.rep.errors.file)
}

```

Mixed regression model

MLR and NPKR model M11 and Mixed models M1, M3 and M11

```

function (fco2.alt.data, pure.model.file, mixed.model.error.file, npkr.model.bands.file)
{
  library(np)
  log.chl <- log10(fco2.alt.data[, "Ch.conc"])
  fco2.alt.data <- cbind(fco2.alt.data, log.chl)
  set.seed(5000)

  # Dividing data into Training and Test sets
  split <- runif(nrow(fco2.alt.data),0,1)

  training70 <- which(split > 0.3)

  test70 <- which(split < 0.3 | split == 0.3)

  train70.data <- fco2.alt.data[training70,]

  test70.data <- fco2.alt.data[test70,]

  # Alt Models
  mlr.alt <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,

```

APPENDIX
R CODE

142

```
data = train70.data)
npkr.alt.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,
data = train70.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
npkr.alt <- npreg(bws = npkr.alt.bands)
print("Model 1 Done")

# Mixed Models
mlr.m1 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD, data = train70.data)
npkr.m1.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD,
data = train70.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
npkr.m1 <- npreg(bws = npkr.m1.bands)

print("Model 2 Done")

mlr.sal <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Salinity,
data = train70.data)
npkr.sal.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Salinity,
data = train70.data, regtype = "lc", bwtype = "fixed", ckertype = "epanechnikov",
na.action = na.omit)
npkr.sal <- npreg(bws = npkr.sal.bands)

print("Model 3 Done")

# Pure Model Predictions and Error Rates
mlr.predict.70 <- predict(mlr.alt, newdata = test70.data)
npkr.predict.70 <- predict(npkr.alt, newdata = test70.data)

alt.mlr.mse <- sum((mlr.predict.70 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
alt.npkr.mse <- sum((npkr.predict.70 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)

alt.mlr.mae <- sum(abs(mlr.predict.70 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
alt.npkr.mae <- sum(abs(npkr.predict.70 - test70.data[, "fCO2.Water"]))/nrow(test70.data)

alt.mlr.rmse <- sqrt(alt.mlr.mse)
alt.npkr.rmse <- sqrt(alt.npkr.mse)

alt.table <- cbind(c(alt.mlr.mse, alt.npkr.mse), c(alt.mlr.mae, alt.npkr.mae),
c(alt.mlr.rmse, alt.npkr.rmse))
colnames(alt.table) <- c("Mean Square Error", "Mean Absolute Error",
"Root Mean Square Error")
```


APPENDIX
R CODE

143

```
rownames(alt.table) <- c("MLR", "NPKR")
write.csv(alt.table, file = pure.model.file)

# Mixed Model Predictions
mlr.M1.predict.70 <- predict(mlr.m1, newdata = test70.data)
npkr.M1.predict.70 <- predict(npkr.m1, newdata = test70.data)
npkr.M1.predict.70[npkr.M1.predict.70 == 0] <- mlr.M1.predict.70[npkr.M1.predict.70 == 0]
mixed.M1.predict.70 <- npkr.M1.predict.70

mlr.sal.predict.70 <- predict(mlr.sal, newdata = test70.data)
npkr.sal.predict.70 <- predict(npkr.sal, newdata = test70.data)
npkr.sal.predict.70[npkr.sal.predict.70 == 0] <- mlr.sal.predict.70[npkr.sal.predict.70 == 0]
mixed.sal.predict.70 <- npkr.sal.predict.70

npkr.predict.70[npkr.predict.70 == 0] <- mlr.predict.70[npkr.predict.70 == 0]
mixed.alt.predict.70 <- npkr.predict.70

# Mixed Model Error Rates
mixed.M1.70.mse <-
sum((mixed.M1.predict.70 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
mixed.sal.70.mse <-
sum((mixed.sal.predict.70 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
mixed.alt.70.mse <-
sum((mixed.alt.predict.70 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)

mixed.M1.70.mae <-
sum(abs(mixed.M1.predict.70 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
mixed.sal.70.mae <-
sum(abs(mixed.sal.predict.70 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
mixed.alt.70.mae <-
sum(abs(mixed.alt.predict.70 - test70.data[, "fCO2.Water"]))/nrow(test70.data)

mixed.M1.70.rmse <- sqrt(mixed.M1.70.mse)
mixed.sal.70.rmse <- sqrt(mixed.sal.70.mse)
mixed.alt.70.rmse <- sqrt(mixed.alt.70.mse)

mixed.table <- cbind(c(mixed.M1.70.mse, mixed.sal.70.mse, mixed.alt.70.mse),
c(mixed.M1.70.mae, mixed.sal.70.mae, mixed.alt.70.mae),
c(mixed.M1.70.rmse, mixed.sal.70.rmse, mixed.alt.70.rmse))
colnames(mixed.table) <- c("Mean Square Error", "Mean Absolute Error",
"Root Mean Square Error")
rownames(mixed.table) <- c("M1", "Salinity", "Altimetry")
```

```
write.csv(mixed.table, file = mixed.model.error.file)

npkr.bands.table <- rbind(c(npkr.m1.bands$bw, 0), npkr.sal.bands$bw,
npkr.alt.bands$bw)
colnames(npkr.bands.table) <- c("Sea Surface Temperature",
"Log Chlorophyll-a Concentration", "MLD", "Salinity or Altimetry")
rownames(npkr.bands.table) <- c("Model M1", "Salinity Model", "Altimetry Model")
write.csv(npkr.bands.table, file = npkr.model.bands.file)

}
```

Mixed regression model subset division

```
function (fco2.alt.data, mixed.subsets.errors.file, mixed.subsets.bands.file)
{
library(np)
log.chl <- log10(fco2.alt.data[, "Ch.conc"])
fco2.alt.data <- cbind(fco2.alt.data, log.chl)
set.seed(5000)

# Dividing data into Training and Test sets
split <- runif(nrow(fco2.alt.data), 0, 1)
training80 <- which(split > 0.2)
training70 <- which(split > 0.3)
training60 <- which(split > 0.4)
training50 <- which(split > 0.5)
training40 <- which(split > 0.6)
training30 <- which(split > 0.7)
training20 <- which(split > 0.8)

test80 <- which(split < 0.2 | split == 0.2)
test70 <- which(split < 0.3 | split == 0.3)
test60 <- which(split < 0.4 | split == 0.4)
test50 <- which(split < 0.5 | split == 0.5)
test40 <- which(split < 0.6 | split == 0.6)
test30 <- which(split < 0.7 | split == 0.7)
test20 <- which(split < 0.8 | split == 0.8)

train80.data <- fco2.alt.data[training80,]
train70.data <- fco2.alt.data[training70,]
train60.data <- fco2.alt.data[training60,]
```

APPENDIX
R CODE

145

```
train50.data <- fco2.alt.data[training50,]
train40.data <- fco2.alt.data[training40,]
train30.data <- fco2.alt.data[training30,]
train20.data <- fco2.alt.data[training20,]

test80.data <- fco2.alt.data[test80,]
test70.data <- fco2.alt.data[test70,]
test60.data <- fco2.alt.data[test60,]
test50.data <- fco2.alt.data[test50,]
test40.data <- fco2.alt.data[test40,]
test30.data <- fco2.alt.data[test30,]
test20.data <- fco2.alt.data[test20,]

# Subset Divisions Altimetry Mixed
mlr.alt.80 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,
data = train80.data)
npkr.alt.bands.80 <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +
Altimetry, data = train80.data, regtype = "lc", bwtype = "fixed",
ckertype = "epanechnikov", na.action = na.omit,
bws = c(0.171998349, 0.042561028, 7.721106395, 0.040249224), bandwidth.compute = FALSE)
npkr.alt.80 <- npreg(bws = npkr.alt.bands.80)

print("Model 1 Done")

mlr.alt.70 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,
data = train70.data)
npkr.alt.bands.70 <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +
Altimetry, data = train70.data, regtype = "lc", bwtype = "fixed",
ckertype = "epanechnikov", na.action = na.omit,
bws = c(0.443904215, 0.035362319, 7.131079051, 0.02499924), bandwidth.compute = FALSE)
npkr.alt.70 <- npreg(bws = npkr.alt.bands.70)

print("Model 2 Done")

mlr.alt.60 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,
data = train60.data)
npkr.alt.bands.60 <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +
Altimetry, data = train60.data, regtype = "lc", bwtype = "fixed",
ckertype = "epanechnikov", na.action = na.omit,
bws = c(0.443904215, 0.035362319, 7.112209135, 0.02499924), bandwidth.compute = FALSE)
npkr.alt.60 <- npreg(bws = npkr.alt.bands.60)
```

APPENDIX
R CODE

146

```
print("Model 3 Done")
```

```
mlr.alt.50 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,  
data = train50.data)  
npkr.alt.bands.50 <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +  
Altimetry, data = train50.data, regtype = "lc", bwtype = "fixed",  
ckertype = "epanechnikov", na.action = na.omit,  
bws = c(1.884781698, 0.03476244, 5.431964405, 0.013416408), bandwidth.compute = FALSE)  
npkr.alt.50 <- npreg(bws = npkr.alt.bands.50)
```

```
print("Model 4 Done")
```

```
mlr.alt.40 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,  
data = train40.data)  
npkr.alt.bands.40 <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +  
Altimetry, data = train40.data, regtype = "lc", bwtype = "fixed",  
ckertype = "epanechnikov", na.action = na.omit,  
bws = c(0.172490284, 0.101954864, 5.720449306, 0.011627553), bandwidth.compute = FALSE)  
npkr.alt.40 <- npreg(bws = npkr.alt.bands.40)
```

```
print("Model 5 Done")
```

```
mlr.alt.30 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,  
data = train30.data)  
npkr.alt.bands.30 <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +  
Altimetry, data = train30.data, regtype = "lc", bwtype = "fixed",  
ckertype = "epanechnikov", na.action = na.omit,  
bws = c(0.172490284, 0.043134419, 6.120456725, 0.02236068), bandwidth.compute = FALSE)  
npkr.alt.30 <- npreg(bws = npkr.alt.bands.30)
```

```
print("Model 6 Done")
```

```
mlr.alt.20 <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,  
data = train20.data)  
npkr.alt.bands.20 <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +  
Altimetry, data = train20.data, regtype = "lc", bwtype = "fixed",  
ckertype = "epanechnikov", na.action = na.omit,  
bws = c(0.209391551, 0.016529894, 6.837186568, 0.050535136), bandwidth.compute = FALSE)  
npkr.alt.20 <- npreg(bws = npkr.alt.bands.20)
```

```
print("Model 7 Done")
```

APPENDIX
R CODE

147

```
# Subset Division Predictions
mlr.alt.predict.80 <- predict(mlr.alt.80, newdata = test80.data)
npkr.alt.predict.80 <- predict(npkr.alt.80, newdata = test80.data)
npkr.alt.predict.80[npkr.alt.predict.80 == 0]
<- mlr.alt.predict.80[npkr.alt.predict.80 == 0]
mixed.alt.predict.80 <- npkr.alt.predict.80

mlr.alt.predict.70 <- predict(mlr.alt.70, newdata = test70.data)
npkr.alt.predict.70 <- predict(npkr.alt.70, newdata = test70.data)
npkr.alt.predict.70[npkr.alt.predict.70 == 0]
<- mlr.alt.predict.70[npkr.alt.predict.70 == 0]
mixed.alt.predict.70 <- npkr.alt.predict.70

mlr.alt.predict.60 <- predict(mlr.alt.60, newdata = test60.data)
npkr.alt.predict.60 <- predict(npkr.alt.60, newdata = test60.data)
npkr.alt.predict.60[npkr.alt.predict.60 == 0]
<- mlr.alt.predict.60[npkr.alt.predict.60 == 0]
mixed.alt.predict.60 <- npkr.alt.predict.60

mlr.alt.predict.50 <- predict(mlr.alt.50, newdata = test50.data)
npkr.alt.predict.50 <- predict(npkr.alt.50, newdata = test50.data)
npkr.alt.predict.50[npkr.alt.predict.50 == 0]
<- mlr.alt.predict.50[npkr.alt.predict.50 == 0]
mixed.alt.predict.50 <- npkr.alt.predict.50

mlr.alt.predict.40 <- predict(mlr.alt.40, newdata = test40.data)
npkr.alt.predict.40 <- predict(npkr.alt.40, newdata = test40.data)
npkr.alt.predict.40[npkr.alt.predict.40 == 0]
<- mlr.alt.predict.40[npkr.alt.predict.40 == 0]
mixed.alt.predict.40 <- npkr.alt.predict.40

mlr.alt.predict.30 <- predict(mlr.alt.30, newdata = test30.data)
npkr.alt.predict.30 <- predict(npkr.alt.30, newdata = test30.data)
npkr.alt.predict.30[npkr.alt.predict.30 == 0]
<- mlr.alt.predict.30[npkr.alt.predict.30 == 0]
mixed.alt.predict.30 <- npkr.alt.predict.30

mlr.alt.predict.20 <- predict(mlr.alt.20, newdata = test20.data)
npkr.alt.predict.20 <- predict(npkr.alt.20, newdata = test20.data)
npkr.alt.predict.20[npkr.alt.predict.20 == 0]
<- mlr.alt.predict.20[npkr.alt.predict.20 == 0]
mixed.alt.predict.20 <- npkr.alt.predict.20
```

```
# Subset Division Predictions Plots (Mixed Model and NPKR Model)
plot(x = test80.data[, "latitude"], y = test80.data[, "fCO2.Water"], type = "p",
     pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "Mixed Model (80-20 Subsets)"),
     xlab = "Latitude", ylab = "fCO2")
lines(x = test80.data[, "latitude"], y = mixed.alt.predict.80, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
      pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test80.data[, "latitude"], y = test80.data[, "fCO2.Water"], type = "p",
     pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "NPKR Model (80-20 Subsets)"),
     xlab = "Latitude", ylab = "fCO2")
lines(x = test80.data[, "latitude"], y = npkr.alt.predict.80, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
      pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test70.data[, "latitude"], y = test70.data[, "fCO2.Water"], type = "p",
     pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "Mixed Model (70-30 Subsets)"),
     xlab = "Latitude", ylab = "fCO2")
lines(x = test70.data[, "latitude"], y = mixed.alt.predict.70, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
      pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test70.data[, "latitude"], y = test70.data[, "fCO2.Water"], type = "p",
     pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "NPKR Model (70-30 Subsets)"),
     xlab = "Latitude", ylab = "fCO2")
lines(x = test70.data[, "latitude"], y = npkr.alt.predict.70, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
      pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test60.data[, "latitude"], y = test60.data[, "fCO2.Water"], type = "p",
     pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "Mixed Model (60-40 Subsets)"),
     xlab = "Latitude", ylab = "fCO2")
lines(x = test60.data[, "latitude"], y = mixed.alt.predict.60, col = "red")
```

APPENDIX
R CODE

149

```
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))

windows()

plot(x = test60.data[, "latitude"], y = test60.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "NPKR Model (60-40 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test60.data[, "latitude"], y = npkr.alt.predict.60, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))

windows()

plot(x = test50.data[, "latitude"], y = test50.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "Mixed Model (50-50 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test50.data[, "latitude"], y = mixed.alt.predict.50, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))

windows()

plot(x = test50.data[, "latitude"], y = test50.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "NPKR Model (50-50 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test50.data[, "latitude"], y = npkr.alt.predict.50, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))

windows()

plot(x = test40.data[, "latitude"], y = test40.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "Mixed Model (40-60 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test40.data[, "latitude"], y = mixed.alt.predict.40, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))

windows()

plot(x = test40.data[, "latitude"], y = test40.data[, "fCO2.Water"], type = "p",
```

APPENDIX
R CODE

150

```
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "NPKR Model (40-60 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test40.data[, "latitude"], y = npkr.alt.predict.40, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test30.data[, "latitude"], y = test30.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "Mixed Model (30-70 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test30.data[, "latitude"], y = mixed.alt.predict.30, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test30.data[, "latitude"], y = test30.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "NPKR Model (30-70 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test30.data[, "latitude"], y = npkr.alt.predict.30, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test20.data[, "latitude"], y = test20.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "Mixed Model (20-80 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test20.data[, "latitude"], y = mixed.alt.predict.20, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))
```

```
windows()
```

```
plot(x = test20.data[, "latitude"], y = test20.data[, "fCO2.Water"], type = "p",
pch = ".", col = "blue", main = c("Prediction Plot of fCO2", "NPKR Model (20-80 Subsets)"),
xlab = "Latitude", ylab = "fCO2")
lines(x = test20.data[, "latitude"], y = npkr.alt.predict.20, col = "red")
legend(locator(1), legend = c("Observed fCO2", "Predicted fCO2"),
pch = c(".", "-"), col = c("blue", "red"))
```


APPENDIX
R CODE

151

```
# Subset Division Predictions (Mixed Models)
mixed.alt.80.mse <-
sum((mixed.alt.predict.80 - test80.data[, "fCO2.Water"])^2)/nrow(test80.data)
mixed.alt.70.mse <-
sum((mixed.alt.predict.70 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
mixed.alt.60.mse <-
sum((mixed.alt.predict.60 - test60.data[, "fCO2.Water"])^2)/nrow(test60.data)
mixed.alt.50.mse <-
sum((mixed.alt.predict.50 - test50.data[, "fCO2.Water"])^2)/nrow(test50.data)
mixed.alt.40.mse <-
sum((mixed.alt.predict.40 - test40.data[, "fCO2.Water"])^2)/nrow(test40.data)
mixed.alt.30.mse <-
sum((mixed.alt.predict.30 - test30.data[, "fCO2.Water"])^2)/nrow(test30.data)
mixed.alt.20.mse <-
sum((mixed.alt.predict.20 - test20.data[, "fCO2.Water"])^2)/nrow(test20.data)

mixed.alt.80.mae <-
sum(abs(mixed.alt.predict.80 - test80.data[, "fCO2.Water"]))/nrow(test80.data)
mixed.alt.70.mae <-
sum(abs(mixed.alt.predict.70 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
mixed.alt.60.mae <-
sum(abs(mixed.alt.predict.60 - test60.data[, "fCO2.Water"]))/nrow(test60.data)
mixed.alt.50.mae <-
sum(abs(mixed.alt.predict.50 - test50.data[, "fCO2.Water"]))/nrow(test50.data)
mixed.alt.40.mae <-
sum(abs(mixed.alt.predict.40 - test40.data[, "fCO2.Water"]))/nrow(test40.data)
mixed.alt.30.mae <-
sum(abs(mixed.alt.predict.30 - test30.data[, "fCO2.Water"]))/nrow(test30.data)
mixed.alt.20.mae <-
sum(abs(mixed.alt.predict.20 - test20.data[, "fCO2.Water"]))/nrow(test20.data)

mixed.alt.80.rmse <- sqrt(mixed.alt.80.mse)
mixed.alt.70.rmse <- sqrt(mixed.alt.70.mse)
mixed.alt.60.rmse <- sqrt(mixed.alt.60.mse)
mixed.alt.50.rmse <- sqrt(mixed.alt.50.mse)
mixed.alt.40.rmse <- sqrt(mixed.alt.40.mse)
mixed.alt.30.rmse <- sqrt(mixed.alt.30.mse)
mixed.alt.20.rmse <- sqrt(mixed.alt.20.mse)

mixed.table <- cbind(c(mixed.alt.80.mse, mixed.alt.70.mse, mixed.alt.60.mse,
mixed.alt.50.mse, mixed.alt.40.mse, mixed.alt.30.mse, mixed.alt.20.mse),
c(mixed.alt.80.mae, mixed.alt.70.mae, mixed.alt.60.mae, mixed.alt.50.mae,
```

```

mixed.alt.40.mae, mixed.alt.30.mae, mixed.alt.20.mae), c(mixed.alt.80.rmse,
mixed.alt.70.rmse, mixed.alt.60.rmse, mixed.alt.50.rmse, mixed.alt.40.rmse,
mixed.alt.30.rmse, mixed.alt.20.rmse))
colnames(mixed.table) <- c("Mean Square Error", "Mean Absolute Error",
"Root Mean Square Error")
rownames(mixed.table) <- c("80 - 20", "70 - 30", "60 - 40", "50 - 50",
"40 - 60", "30 - 70", "20 - 80")
write.csv(mixed.table, file = mixed.subsets.errors.file)

npkr.bands.table <- rbind(npkr.alt.bands.80$bw, npkr.alt.bands.70$bw,
npkr.alt.bands.60$bw, npkr.alt.bands.50$bw, npkr.alt.bands.40$bw,
npkr.alt.bands.30$bw, npkr.alt.bands.20$bw)
colnames(npkr.bands.table) <- c("Sea Surface Temperature",
"Log Chlorophyll-a Concentration", "MLD", "Altimetry")
rownames(npkr.bands.table) <- c("80 - 20", "70 - 30", "60 - 40", "50 - 50",
"40 - 60", "30 - 70", "20 - 80")
write.csv(npkr.bands.table, file = mixed.subsets.bands.file)
}

```

Mixed regression model error simulation

```

function (fco2.alt.data, reps = 50, mixed.rep.error.file, npkr.rep.bands.file)
{
  library(np)
  log.chl <- log10(fco2.alt.data[, "Ch.conc"])
  fco2.alt.data <- cbind(fco2.alt.data, log.chl)

  mixed.table.errors <- matrix(0, nrow = reps, ncol = 3)
  colnames(mixed.table.errors) <- c("Mean Square Error",
"Mean Absolute Error", "Root Mean Square Error")

  npkr.bands.table <- matrix(0, nrow = reps, ncol = 4)
  colnames(npkr.bands.table) <- c("Sea Surface Temperature",
"Log Chlorophyll-a Concentration", "MLD", "Altimetry")

  for(i in 1:reps){
    # Dividing data into Training and Test sets
    split <- runif(nrow(fco2.alt.data), 0, 1)
    training70 <- which(split > 0.3)
    test70 <- which(split < 0.3 | split == 0.3)
    train70.data <- fco2.alt.data[training70,]
  }
}

```

APPENDIX
R CODE

153

```
test70.data <- fco2.alt.data[test70,]

# Alt Models
mlr.alt <- lm(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD + Altimetry,
data = train70.data)
npkr.alt.bands <- npregbw(formula = fCO2.Water ~ Intake.Temp + log.chl + MLD +
Altimetry, data = train70.data, regtype = "lc", bwtype = "fixed",
ckertype = "epanechnikov", na.action = na.omit)
npkr.alt <- npreg(bws = npkr.alt.bands)

# Alt Predictions
mlr.predict.70 <- predict(mlr.alt, newdata = test70.data)
npkr.predict.70 <- predict(npkr.alt, newdata = test70.data)

# Alt Mixed Predictions
zeros <- which(npkr.predict.70 == 0)
npkr.predict.70[zeros] <- mlr.predict.70[zeros]
mixed.alt.predict.70 <- npkr.predict.70

# Mixed Errors
mixed.alt.70.mse <-
sum((mixed.alt.predict.70 - test70.data[, "fCO2.Water"])^2)/nrow(test70.data)
mixed.alt.70.mae <-
sum(abs(mixed.alt.predict.70 - test70.data[, "fCO2.Water"]))/nrow(test70.data)
mixed.alt.70.rmse <- sqrt(mixed.alt.70.mse)

mixed.table.errors[i,] <- c(mixed.alt.70.mse, mixed.alt.70.mae,
mixed.alt.70.rmse)
npkr.bands.table[i,] <- npkr.alt.bands$bw
print(paste("Model", i, "Done"))
}

write.csv(mixed.table.errors, file = mixed.rep.error.file)
write.csv(npkr.bands.table, file = npkr.rep.bands.file)
}
```

List of References

- Al-anezi, K., Somerfield, C., Mee, D. and Hilal, N. (2008). Parameters affecting the solubility of carbon dioxide in seawater at the conditions encountered in MSF desalination plants. *Desalination*, vol. 222, pp. 548–571.
- Archer, D., Winguth, A., Lea, D. and Mahowald, N. (2000). What caused the glacial/interglacial atmospheric pCO₂ cycles? *Reviews of Geophysics*, vol. 38, no. 2, pp. 159–189.
- Atlas of Surface Marine Data (1994). Ocean Climate Laboratory, NOAA, Washington, DC. CD-ROM NODC-56.
- Bakker, D., Debaar, H. and Bathmann, U. (1997). Changes of carbon dioxide in surface waters during spring in the Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 44, no. 1–2, pp. 91–127.
- Bates, N.R., Pequignet, A.C. and Sabine, C.L. (2006). Ocean carbon cycling in the Indian Ocean: 1. Spatiotemporal variability of inorganic carbon and air-sea CO₂ gas exchange. *Global Biogeochemical Cycles*, vol. 20, no. GB3020. Doi:10.1029/2005GB002491.
- Böning, C.W., Dispert, A., Visbeck, M., Rintoul, S.R. and Schwarzkopf, F.U. (2008). The response of the Antarctic Circumpolar Current to recent climate change. *Nature Geoscience*, vol. 1, pp. 864–869.
- Boutin, J., Etcheto, J., Dandonneau, Y., Bakker, D.C.E., Feely, R.A., Inoue, H.Y., Ishii, M., Ling, R.D., Nightingale, P.D., Metzl, N. and Wanninkhof, R.H. (1999). Satellite sea surface temperature: a powerful tool for interpreting in situ pCO₂ measurements in the equatorial Pacific Ocean. *Tellus*, vol. 51B, pp. 490–508.

- Bye, J.A.T. (1996). Coupling ocean - atmosphere models. *Earth-Science Reviews*, vol. 40, pp. 149–162.
- Dandonneau, Y. (1995). Sea-surface partial pressure of carbon dioxide in the Eastern Equatorial Pacific (August 1991 to October 1992): A multivariate analysis of physical and biological factors. *Deep-Sea Research II*, vol. 42, no. 2–3, pp. 349–364.
- Deng, F. and Chen, J.M. (2011). Recent global CO₂ flux inferred from atmospheric CO₂ observations and its regional analyses. *Biogeosciences Discussions*, vol. 8, no. 2, pp. 3497–3536.
- Dickson, A.G. and Goyet, C. (eds.) (1994). *Handbook of Methods for Analysis of the Various Parameters of the Carbon Dioxide System in Seawater: Version 2*. ORNL/CDIAC-74.
- Dong, S., Sprintall, J. and Gille, S.T. (2006). Location of the Antarctic Polar Front from AMSR-E satellite sea surface temperature measurements. *Journal of Physical Oceanography*, vol. 36, pp. 2075–2089.
- Fox, J. (2005). Introduction to nonparametric regression. Economic and Social Research Council, Available at: <http://socserv.mcmaster.ca/jfox/Courses/Oxford-2005/slides-handout.pdf>, Accessed: 14 November 2011.
- Friedrich, T. and Oeschler, A. (2009). Neural network-based estimates of North Atlantic surface pCO₂ from satellite data: A methodological study. *Journal of Geophysical Research*, vol. 114. C03020, doi:10.1029/2007JC004646.
- Fu, L., Chelton, D.B., Le Traon, P. and Morrow, R. (2010). Eddy dynamics from satellite altimetry. *Oceanography*, vol. 23, no. 4, pp. 14–25.
- Gaines, S. and Airame, S. (2012). Upwelling. <http://oceanexplorer.noaa.gov/explorations/02quest/background/upwelling/upwelling.html>, Accessed: July 2012.

- GLOBALVIEW-CO₂ (2000). Cooperative atmospheric data integration project—carbon dioxide. CD-ROM. Boulder, Colorado, Also available on Internet via anonymous FTP to <ftp.cmdl.noaa.gov>, Path: `cag/co2/GLOBALVIEW`.
- Golub, G.H. and van Loan, C.F. (1996). *Matrix Computations*, chap. 1, pp. 1–47. 3rd edn. The John Hopkins University Press, 2715 North Charles Street, Baltimore, Maryland.
- Goyet, C. (1998). Observations of the CO₂ system properties in the tropical Atlantic Ocean. *Marine Chemistry*, vol. 60, no. 1–2, pp. 49–61.
- Jamet, C., Moulin, C. and Lefèvre, N. (2007). Estimation of the oceanic pCO₂ in the North Atlantic from VOS lines *in situ* measurements: Parameters needed to generate seasonally mean maps. *Annales Geophysicae*, vol. 25, pp. 2247–2257.
- Kohonen, T. (2001). *Self-Organizing Maps*. 3rd edn. Springer, Berlin, Heidelberg, New York. ISBN 3-540-67921-9.
- Le Quéré, C., Raupach, M.R., Canadell, J.G., Marland, G., Bopp, L., Ciais, P., Conway, T.J., Doney, S.C., Feely, R.A., Foster, P., Friedlingstein, P., Gurney, K., Houghton, R.A., House, J.I., Huntingford, C., Levy, P.E., Lomas, M.R., Majkut, J., Metzl, N., Ometto, J.P., Peters, G.P., Colin Prentice, I., Randerson, J.T., Running, S.W., Sarmiento, J.L., Schuster, U., Sitch, S., Takahashi, T., Viovy, N., van der Werf, G.R. and Ian Woodward, F. (2009). Trends in the sources and sinks of carbon dioxide. *Nature Geoscience*, vol. 2, pp. 831–836.
- Le Quéré, C., Rödenbeck, C., Buitenhuis, E.T., Conway, T.J., Langenfelds, R., Gomez, A., Labuschagne, C., Ramonet, M., Nakazawa, T., Metzl, N., Gillett, N. and Heimann, M. (2007). Saturation of the Southern Ocean CO₂ sink due to recent climate change. *Science (New York, N. Y.)*, vol. 316, no. 5832, pp. 1735–1738.
- Lee, K., Wanninkhof, R.H., Takahashi, T., Doney, S.C. and Feely, R.A. (1998). Low interannual variability in recent oceanic uptake of atmospheric carbon dioxide. *Nature*, vol. 396, pp. 151–159.

- Lefèvre, N. and Taylor, A. (2002). Estimating pCO₂ from sea surface temperatures in the Atlantic gyres. *Deep-Sea Research I*, vol. 49, pp. 539–554.
- Lefèvre, N., Watson, A.J. and Watson, A.R. (2005). A comparison of multiple regression and neural network techniques for mapping *in situ* pCO₂ data. *Tellus*, vol. 57B, pp. 375–384.
- Lüger, H., Wallace, D.W.R. and Körtzinger, A. (2004). The pCO₂ variability in the midlatitude North Atlantic Ocean during a full annual cycle. *Global Biogeochemical Cycles*, vol. 18, no. GB3023.
- Li, Q. and Racine, J.S. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, vol. 14, pp. 485–512.
- Li, Q. and Racine, J.S. (2007). *Nonparametric Econometrics: Theory and Practice*, chap. 2, pp. 57–115. Princeton University Press.
- Lindstrom, E.J. (2012). Wind driven surface currents: Upwelling and downwelling background. <http://oceanmotion.org/html/background/upwelling-and-downwelling.html>, Accessed: July 2012.
- Longhurst, A. (1995). Seasonal cycles of pelagic production and consumption. *Progress in Oceanography*, vol. 36, no. 2, pp. 77–167.
- Lueker, T.J. (2000). Ocean pCO₂ calculated from dissolved inorganic carbon, alkalinity, and equations for K₁ and K₂: validation based on laboratory measurements of CO₂ in gas and seawater at equilibrium. *Marine Chemistry*, vol. 70, no. 1–3, pp. 105–119.
- Markham, A.E. and Kobe, K.A. (1941). The solubility of carbon dioxide and nitrous oxide in aqueous salt solutions. *Journal of the American Chemical Society*, vol. 63, no. 2, pp. 449–454.
- Maximenko, N. and Niiler, P. (2011). 1992-2002 mean dynamic ocean topography. <http://apdrc.soest.hawaii.edu/projects/DOT/>, Accessed: October 2012.

- Maximenko, N., Niiler, P., Centurioni, L., Rio, M., Melnichenko, O., Chambers, D., Zlotnicki, V. and Galperin, B. (2009). Mean dynamic topography of the ocean derived from satellite and drifting buoy data using three different techniques. *Journal of Atmosphere and Ocean Technology*, vol. 26, no. 9, pp. 1910–1919.
- McNeil, B.I., Metzl, N., Key, R.M., Matear, R.J. and Corbiere, A. (2007). An empirical estimate of the Southern Ocean air-sea CO₂ flux. *Global Biogeochemical Cycles*, vol. 21, no. GB3011.
- McNeil, B.I., Tilbrook, B. and Matear, R.J. (2001). Accumulation and uptake of anthropogenic CO₂ in the Southern Ocean, south of Australia between 1968 and 1996. *Journal of Geophysical Research*, vol. 106, no. C12. Doi:10.1029/2000JC000331.
- Monteiro, P.M.S. (2010). Understanding and predicting the seasonal cycle of carbon in the Southern Ocean: A high resolution global carbon-climate model study and model development platform in CSIRO. Internal Project Proposal at the CSIR.
- Monterey Bay Aquarium Research Institute (2005). The carbon footprint of the coastal ocean. Tech. Rep., Monterey Bay Aquarium Research Institute.
- Moore, J.K., Abbott, M.R. and Richman, J.G. (1999). Location and dynamics of the Antarctic Polar Front from satellite sea surface temperature data. *Journal of Geophysical Research*, vol. 104, no. C2, pp. 3059–3073.
- Nelson, N.B., Bates, N.R., Siegel, D.A. and Michaels, A.F. (2001). Spatial variability of the CO₂ sink in the Sargasso Sea. *Deep-Sea Research II*, vol. 48, pp. 1801–1821.
- Olsen, A., Bellerby, R.G.J., Johannessen, T., Omar, A.M. and Skjelvan, I. (2003). Interannual variability in the wintertime air-sea flux of carbon dioxide in the northern North Atlantic, 1981-2001. *Deep-Sea Research I*, vol. 50, pp. 1323–1338.
- Olsen, A., Triñanes, J.A. and Wanninkhof, R.H. (2004). Sea-air flux of CO₂ in the Caribbean Sea estimated using *in situ* and remote sensing data. *Remote Sensing of Environment*, vol. 89, pp. 309–324.

- Ono, T., Saino, T., Kurita, N. and Sasaki, K. (2004). Basin-scale extrapolation of shipboard pCO₂ data by using satellite SST and Chl *a*. *International Journal of Remote Sensing*, vol. 25, no. 19, pp. 3803–3815.
- R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Racine, J.S. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, vol. 119, no. 1, pp. 99–130.
- Rangama, Y. (2005). Variability of the net air-sea CO₂ flux inferred from shipboard and satellite measurements in the Southern Ocean south of Tasmania and New Zealand. *Journal of Geophysical Research*, vol. 110, no. C9, pp. 1–17. Doi:10.1029/2004JC002619.
- Sarma, V.V.S.S., Saino, T., Sasaoka, K., Nojiri, Y., Ono, T., Ishii, M., Inoue, H.Y. and Matsumoto, K. (2006). Basin-scale pCO₂ distribution using satellite sea surface temperature, Chl *a*, and climatological salinity in the North Pacific in spring and summer. *Global Biogeochemical Cycles*, vol. 20, no. GB3005. Doi:10.1029/2005GB002594.
- Sarmiento, J.L. and Gruber, N. (2002). Sinks for anthropogenic carbon. *Physics Today*, vol. 55, no. 8, p. 30.
- Sheather, S.J. (2009). *A Modern Approach to Regression with R*, chap. 5, pp. 130–135. Springer Science+Business Media, LLC, 223 Spring Street, New York, NY 10013, USA.
- Sigman, D.M. and Boyle, E.A. (2000). Glacial/interglacial variations in atmospheric carbon dioxide. *Nature*, vol. 407, pp. 859–869.
- Sigman, D.M., Hain, M.P. and Haug, G.H. (2010). The polar ocean and glacial cycles in atmospheric CO₂ concentration. *Nature*, vol. 466, pp. 47–55.

- Slansky, C.M., Feely, R.A. and Wanninkhof, R.H. (1997). The stepwise linear regression method for calculating anthropogenic CO₂ invasion into the North Pacific Ocean. In: Tsunogai, S. and Foundation, J.M.S. (eds.), *Biogeochemical Processes in the North Pacific*, pp. 70–79.
- Stephens, M.P., Samuels, G., Olson, D.B., Fine, R.A. and Takahashi, T. (1995). Sea-air flux of CO₂ in the North Pacific using shipboard and satellite data. *Journal of Geophysical Research*, vol. 100, no. C7, pp. 13571–13583. Doi:10.1029/95JC00901.
- Takahashi, T. and Chipman, D. (1982). Carbon dioxide partial pressure in surface waters of the Southern Ocean. *Antarctic Journal*, vol. 17.
- Takahashi, T., Feely, R.A., Weiss, R.F., Wanninkhof, R.H., Chipman, D.W., Sutherland, S.C. and Takahashi, T.T. (1997). Global air-sea flux of CO₂: an estimate based on measurements of sea-air pCO₂ difference. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 16, pp. 8292–8299.
- Takahashi, T., Sutherland, S., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R.H., Feely, R.A. and Sabine, C. (2002). Global sea-air CO₂ flux based on climatological surface ocean pCO₂ and seasonal biological and temperature effects. *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 49, no. 9–10, pp. 1601–1622.
- Takahashi, T. and Sutherland, S.C. (2007). Global ocean surface water partial pressure of CO₂ database: Measurements performed during 1968-2007 (version 2007). ORNL/CDIAC-152, NDP-088a, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee.
- Takahashi, T., Sutherland, S.C. and Kozyr, A. (2009a). Global ocean surface water partial pressure of CO₂ database: Measurements performed during 1957–2009 (version 2009). ORNL/CDIAC-152, NDP-088(V2009), Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee, doi: 10.3334/CDIAC/otg.ndp088(V2009).

- Takahashi, T., Sutherland, S.C., Wanninkhof, R., Sweeney, C., Feely, R.A., Chipman, D.W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A., Bakker, D.C.E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T.S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R., Wong, C.S., Delille, B., Bates, N.R. and deBaar, H.J.W. (2009*b*). Climatological mean and decadal change in surface ocean pCO₂, and net sea-air CO₂ flux over the global oceans. *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 56, no. 8–10, pp. 554–577.
- Telszewski, M., Chazottes, A., Schuster, U., Watson, A.J., Moulin, C., Bakker, D.C.E., González-Dávila, M., Johannessen, T., Körtzinger, A., Lüger, H., Olsen, A., Omar, A., Padin, X.A., Ríos, A.F., Steinhoff, T., Santana-Casiano, M., Wallace, D.W.R. and Wanninkhof, R.H. (2009). Estimating the monthly pCO₂ distribution in the North Atlantic using a self-organizing neural network. *Biogeosciences*, vol. 6, pp. 1405–1421.
- Tsunogai, S., Watanabe, S. and Sato, T. (1999). Is there a “continental shelf pump” for the absorption of atmospheric CO₂? *Tellus*, vol. 51B, no. 3, pp. 701–712.
- Wallace, D.W.R. (1995). *Ocean Observing System Development Panel Report*, chap. 5, p. 54. Texas A&M University, College Station, Texas.
- Weiss, R.F. (1974). Carbon dioxide in water and seawater, the solubility of a non-ideal gas. *Marine Chemistry*, vol. 2, pp. 203–215.
- Yasunishi, A. and Yoshida, F. (1979). Solubility of carbon dioxide in aqueous electrolyte solutions. *Journal of Chemical and Engineering Data*, vol. 24, no. 1, pp. 11–14.