Chapter 4

Data

4.1 Introduction

In the research that is conducted in this study, the ocean properties and the relationships between different ocean properties are investigated. In order to conduct a meaningful investigation and draw accurate conclusions, sufficient data is needed for the different oceanic properties. It is often difficult to obtain data measurements for the entire ocean. Thus, various data sources have to be combined. Some of these data sources are discussed in this chapter, including *in situ* measurements, satellite data, Argo floats, the GlobColour data set, and the calculations made for the MLD data.

These raw data sets require processing before they can be used for the investigation. In addition to the processing that is performed on the various data sets, validation is also carried out in order to determine the accuracy of the data collected. The processing and validation of the data sets are also discussed in this chapter.

4.2 Data Collection

4.2.1 Introduction

In order to find the empirical relationship between the pCO_2 and other ocean properties that affect the CO_2 fluxes in the ocean, a certain set of data is required. Data for the various different oceanic properties as well as the pCO_2 values at different locations in the ocean are required. There are several means to collect this data. Some of these means are discussed in this section.

4.2.2 *In situ* measurements

In situ measurements are referred to as measurements that are taken at the original (or natural) place (or position) [18]. In this case it means that measurements are taken by ships or floats that travel in the ocean. These vessels physically take measurements of ocean properties by sampling the water in which they sail or float. In the SRP project, the South African National Antarctic Expedition (SANAE) cruises, amongst others, are considered as a source of *in situ* data. Other *in situ* cruises conducted by the Southern Ocean Carbon and Climate Observatory (SOCCO) include the Gough cruise in spring (which travels towards the Gough island) and the Marion cruise in autumn (which travels towards the Marion island).

In situ trips are carried out across global oceans throughout the year. Databases are created to collect information acquired on all the trips. These compiled datasets contain information that allows researchers and modellers to work with global ocean data and to gain a better understanding of the global oceans.

Ocean and CO_2

The region of the Southern Ocean that is located south of Africa, is one of the main contributors to the equator-ward heat flux into the South Atlantic [57]. This part of the ocean thus becomes important when investigating ocean properties, as it affects the rest of the ocean as well. The Agulhas current carries water from the Indian Ocean to the South Atlantic, causing heat fluxes to occur [57]. The high energy meso-scale eddy activity in the region of the Southern Ocean that is south of Africa, serves as an exchange basin for heat in the ocean [57]. Even though the Southern Ocean is suspected to be a significant carbon sink, there are sparse measurements in this region, and the processes in this region are still only vaguely understood.

The Southern Ocean is unique in the effect it has on the coupling between the ocean to the atmosphere and cryosphere [57]. There are a number of processes in the Southern Ocean, that have an effect on circulation in the global ocean and, possibly, on climate change [57]. Some of the most relevant processes are discussed here:

- Arctic Circumpolar Current (ACC): This is the only current that connects three of the major ocean basins and serves as an important interface between these basins, transferring heat and fresh water [57]. The ACC stretches more than 12400 miles around Antarctica. Its great depth and width makes it the largest current in the world [1]. An illustration of where the ACC is

located in the ocean, is given in Figure 4.1.

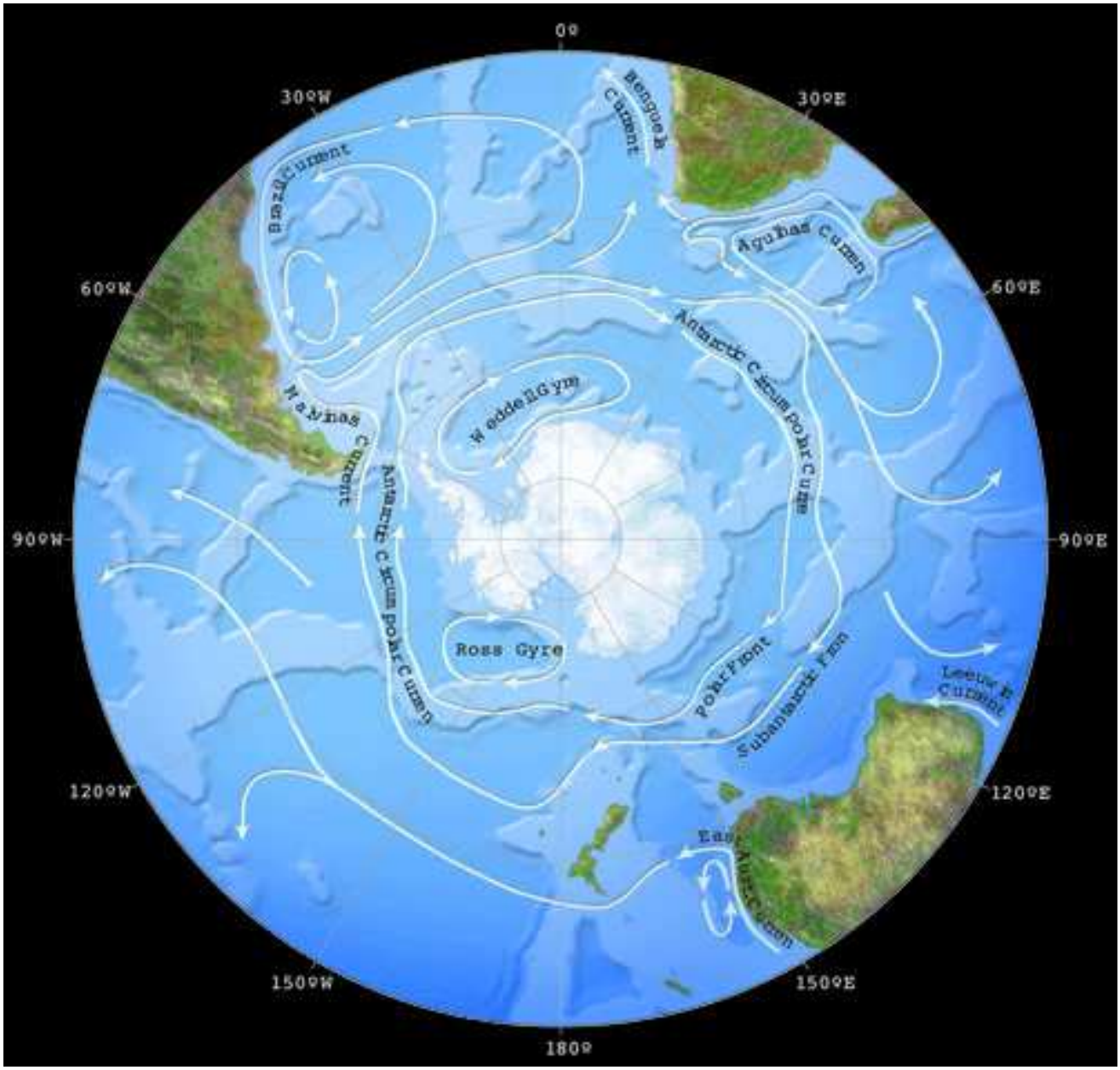

Figure 4.1: The ACC and other currents in the Southern Ocean [6].

- The Southern Ocean provides for strong coupling between the ocean and the atmosphere, in the sub-Antarctic belt, when water from this area is directed Northwards, mixing cool low salinity water with base water in the main thermocline [57].
- South of the Antarctic circumpolar current, the upwelling of deep water provides a means for the heat to be transferred up to 2000m into the atmosphere and cryosphere.

- Cold, dense Antarctic bottom water is produced in the Southern Ocean.
- Anomalies across different climate zones are caused by the large scale variability of the atmospheric circulation in the area above the Southern Ocean.

SANAE cruise

Data on ocean properties, that contribute to an improved understanding of ocean dynamics, are collected (amongst others) on SANAE cruises. The first leg of the SANAE cruise is from Cape Town to Antarctica. Thereafter, the cruise heads from Antarctica to South Georgia and back to Antarctica. Finally, the last leg of the cruise begins at Antarctica and ends at Cape Town. This trip lasts for about three months during austral summer (December to February). Figure 4.2 shows the hydrographic points where the measurements are made during the cruise. The six legs of the trip are identified as follows:

1. Leg 1: Southward GoodHope hydrographic transect.
2. Leg 2: Antarctic ice shelf Eulerian experiments (stationary at Antarctica).
3. Leg 3: Northward Buoy run to South Georgia.
4. Leg 4: Southward Buoy run from South Georgia.
5. Leg 5: Antarctic ice shelf Eulerian experiments (stationary at Antarctica).
6. Leg 6: Northward GoodHope hydrographic transect.

The purpose of the SANAE cruise is firstly to create a long term time series of *in situ* observations in the Southern Ocean, together with observations from the Marion and Gough cruises [57]. An additional purpose of the SANAE cruise is to extend the hydrographic measurements along the GoodHope line, in order to enable long term investigations of the ocean's physical state [57]. Furthermore, the data collected during the SANAE cruise can assist in the elucidation of the relationship between the variability of the surface pCO_2, and the biogeochemical and physical factors that play a role in climate change [57]. The data from the SANAE trip also allows for research to be done on phytoplankton community structures and it enables remote sensing verification of models [57].

The SANAE cruise crosses four different oceanic zones which include the seasonal marginal ice zone, the Permanent Open Ocean Zone, the shelf zone of the South Sandwich and the South Georgia Islands and the frontal zones of the Antarctic Polar front, and the subantarctic front and subtropical

front [57]. This cruise enables a comparison of the factors affecting the physical and biogeochemical attributes of the ocean and climate change factors. (These factors differ across various regions of the ocean [57]). Over time, the data collected on this cruise will result in improved estimations and predictions of climate change variability. This data will also enable further elucidation of the role played by the Southern Ocean in climate change [57].

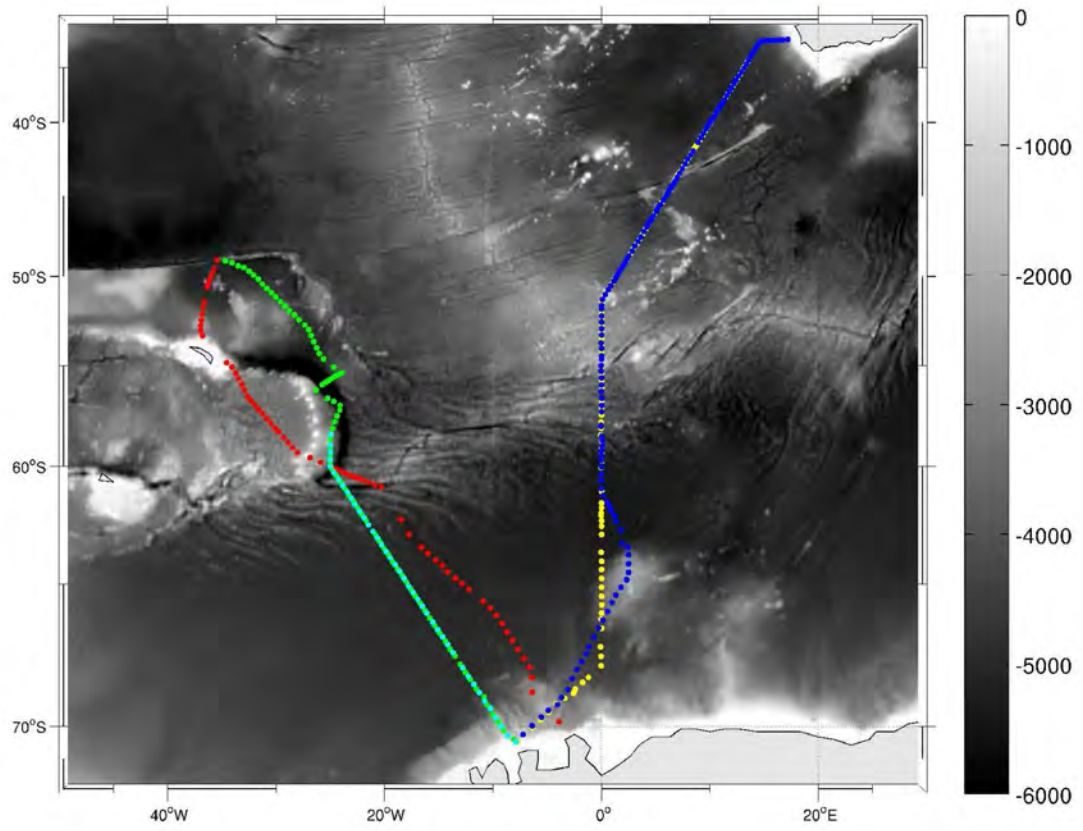

Figure 4.2: The map of the positions of the hydrographic stations during the SANAE cruise. XBT: blue and yellow dots; UCTD: red and green dots; CTD: cyan dots [57].

Physical measurements

The measurements of the ocean properties are taken in accordance with several procedures. Some of these methods and procedures are discussed here.

Changes in oceanic heat fluxes are typically measured using an expendable bathythermograph (XBT).

XBT measurements are especially used in regions of inter-basin exchange where high density observations are needed [57]. Furthermore, these measurements are taken in areas useful for the determination of inter-ocean exchange of heat [57].

The underway conductivity temperature and depth (UCTD) instrument is a new instrument that is used on SANAE cruises. It allows for the measurement of temperature and salinity profiles up to depths of 400 m, while the vessel is moving [57]. From these observations, the physical properties of the upper water column of the ocean can be measured and frontal features can be identified. UCTD profiles are obtained at two hour intervals during the cruise, except when icy conditions prohibit safe deployment and retrieval of the probe [57].

The conductivity temperature depth (CTD) instrument is used for water column sampling up to a depth of 500m [57]. The CTD instrument measures temperature, salinity, oxygen, density and fluorescence [57]. CTD data are collected 6 times a day on the Buoy run (Leg 3 and Leg 4 of the SANAE cruise), and the CTD and UCTD data are combined to create a high density set of observations for the UCTD leg of the Buoy run [57]. After the CTD and UCTD data are collected, the raw data is converted from binary units to engineering units and then the data is processed in order to improve the quality thereof [57]. The data can be affected by different errors that can be accidental or inherent.

Argo floats can also be used to physically measure ocean properties in the ocean. Argo floats are buoys placed in the ocean that take measurements of ocean properties at specified time intervals. Some Argo floats are released into the ocean during the cruise [57]. The deployments of these Argo floats into the ocean will hopefully add more clarity to the region of the Southern Ocean with sparse hydrographic data [57]. The profiles of the oceanic properties, obtained from the Argo floats, are available within 24 to 48 hours of the float surfacing [57].

Other continuous underway measurements include the measurement for TCO_2 and pCO_2. The TCO_2 is the total dissolved inorganic carbon; this is measured by means of the total alkalinity of the water from the uncontaminated underway lab supply [57]. The pCO_2 in the ocean and atmosphere is measured during cruises, using a general Oceanic equilibrator-based system with a Licor LI-700 infra-red gas analyser[57]. The ΔpCO_2, which is the difference between the atmospheric and ocean pCO_2, represents the thermodynamic driving potential for the CO_2 gas transfer across the surface of the ocean [57].

Biological properties of the ocean are also measured during the voyage. Discrete underway biological samples are collected every four hours from either the engine room or from the uncontaminated

surface seawater supply or from the Niskin bottles on the CTD rosette sampler closed at various depths through the water column [57]. In the frontal regions, the biological samples are taken every two hours.

The surface chlorophyll-a samples are taken at every station of the underway sampling. In addition, 250ml of seawater is vacuum filtered through 25 mm Whatman GFIF filters [57]. The filters are extracted with 8ml acetone for twelve to twenty four hours. The raw fluorescence is read using a Turner design trilogy laboratory fluorometer, which is calibrated before the voyage [57]. From this, the chlorophyll-a concentration is derived using the slope of a calibration curve for known chlorophyll-a concentrations [57].

Carbon flux

Various measurements of the CO_2 concentration in the surface of the ocean have been made in the last few decades. The concentration of CO_2 in the ocean can be expressed in various ways. These include: the mole fraction of CO_2 in the headspace (XCO_2), the partial pressure of CO_2 (pCO_2), and the fugacity of CO_2 (fCO_2) in the headspace [65]. The fugacity of CO_2 is a unit that takes the non-ideality of CO_2 gas into account [65]. Thus, the fugacity should be used when calculating the gas exchange. Any one of these different units of CO_2 concentration can be converted to another unit by using a standard set of manipulations [65].

The CO_2 data that is collected by various researchers have been reported differently. CO_2 data is collected in different units, including pCO_2, fCO_2 and XCO_2 [65]. In order to compare data sets with one another and to use the collected data as an uniform data set, all the data have to be converted to a global ocean surface fCO_2 dataset. This combined data set has been designed by the International Surface Ocean-Lower Atmosphere study (SOLAS) project [13] and the International Ocean Carbon Coordination Project (IOCCP) [5]. The data set is constructed by converting all the CO_2 concentrations to fCO_2, where possible, in order to ensure consistency in the data [65]. The XCO_2 values are assumed to be dry mole fractions, unless stated otherwise. It is assumed that with respect to calibration done, the investigators standardized the XCO_2 values [65]. The partial pressure of CO_2 is calculated from

$$(\mathrm{pCO_2})^{wet}_{\mathrm{T}_{equ}} = (\mathrm{XCO_2})^{dry}_{\mathrm{T}_{equ}}(\mathrm{P}_{equ} - \mathrm{pH_2O}), \tag{4.1}$$

where $(XCO_2)^{dry}_{\mathrm{T}_{equ}}$ refers to the CO_2 mole fraction [65]. pH_2O is the water vapour pressure at the equilibrator temperature [65], which is calculated from

$$\mathrm{pH_2O} = \exp(24.4543 - 67.4509(100/\mathrm{T}) - 4.8489\ln(\mathrm{T}/100) - 0.000544\mathrm{T}). \tag{4.2}$$

The difference between the intake temperature (T) and the equilibrator temperature (T_{equ}) is corrected by the empirical relationship

$$(\mathrm{pCO_2})^{wet}_{\mathrm{T}} = (\mathrm{CO_2})^{wet}_{\mathrm{T}_{equ}} \exp[0.0423(\mathrm{T} - \mathrm{T}_{equ})], \tag{4.3}$$

where T refers to the Sea Surface Temperature in the same units as the equilibrator temperature. To convert the $\mathrm{pCO_2}$ values to $\mathrm{fCO_2}$ values, the following is used :

$$(f\mathrm{CO_2})^{wet}_{\mathrm{T}} = (\mathrm{pCO_2})^{wet}_{\mathrm{T}} \exp\left[\frac{(B(\mathrm{CO_2},\mathrm{T}) + 2[1 - (X\mathrm{CO_2})^{wet}_{\mathrm{T}_{equ}}]^2\delta(\mathrm{CO_2},\mathrm{T}))\mathrm{P}_{equ}}{(R^*\mathrm{T})}\right], \tag{4.4}$$

where P_{equ} is the pressure of the equilibrator in atmosphere [65]. T is the sea surface temperature in Kelvin. R is the gas constant: $82.0578\,\mathrm{cm^3\,atm\,mol^{-1}\,K^{-1}}$. B($\mathrm{CO_2}$,T) and δ($\mathrm{CO_2}$,T) are the coefficients for $\mathrm{CO_2}$, that provides systematic corrections for the ideal gas law [65]. B($\mathrm{CO_2}$,T) (in $\mathrm{cm^3\,mol^{-1}}$) is given by

$$B(\mathrm{CO_2},\mathrm{T}) = 1636.75 + 12.0408\mathrm{T} - 3.27957 \times 10^{-2}\mathrm{T}^2 + 3.16528 \times 10^{-5}\mathrm{T}^3. \tag{4.5}$$

Finally, δ($\mathrm{CO_2}$,T) (in $\mathrm{cm^3\,mol^{-1}}$) is given by [65]

$$\delta(\mathrm{CO_2},\mathrm{T}) = 57.7 - 0.188\mathrm{T}. \tag{4.6}$$

These equations show that it is preferable that the $\mathrm{fCO_2}$ values are calculated from dry mole fractions as well as the reported equilibrator and intake temperatures, the equilibrator pressure and the surface salinity [65]. However, it often happens that not all of these variables are recorded. In such a case, data is collected from external sources [65]: The missing pressure values are supplemented by sea level pressure data from the combined reanalysis project from the National Center for Environmental Prediction (NCEP) [10] in the USA and the National Center for Atmospheric Research(NCAR) [9] in the USA. Missing salinity values are supplemented by monthly mean salinity data from the World Ocean Atlas 2005 [65]. If atmospheric pressure or NCEP/NCAR pressure is used, 3hPa is added to provide for the overpressure that normally exits on-board a ship [65].

SOCAT Database

The Surface Ocean CO_2 Atlas (SOCAT) is a database that contains data from observations of ocean properties [12]. SOCAT is an international database, supported by the international ocean carbon coordination project (IOCCP) [12]. The objective of the SOCAT database is to establish a standard global surface carbon dioxide dataset that combines all *in situ* data that is publicly available in a uniform data format [12]. This database includes data from more than 14 countries. The data were collected during more than 2150 cruises from 1968 to 2007. It contains more than 7.5 million measurements of the various carbon parameters that is compiled in a uniform format [12]. The SOCAT database contains data of the oceanic properties in the regions that are indicated by the red lines in Figure 4.3.

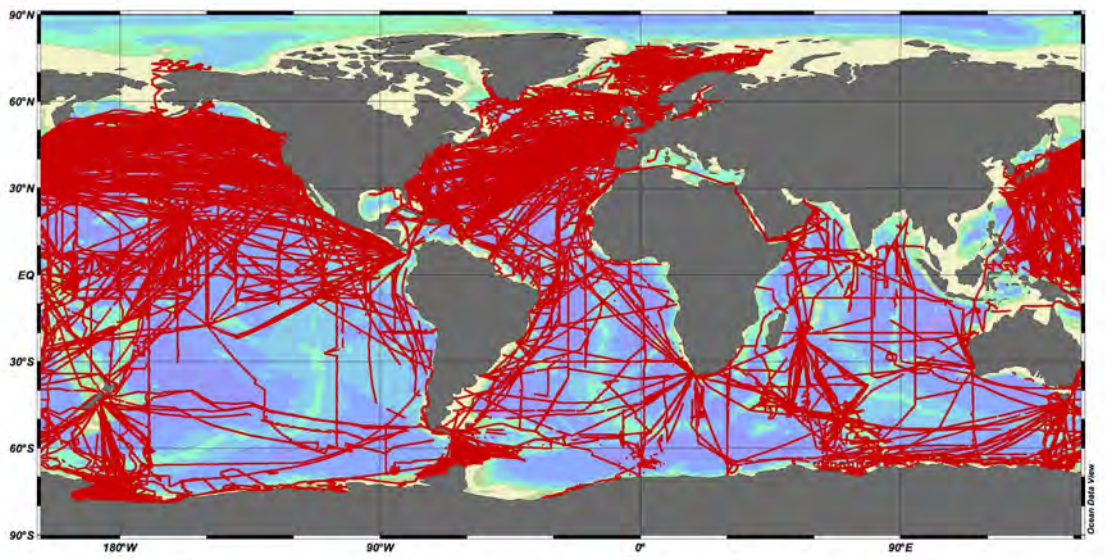

Figure 4.3: SOCAT database areas covered in the ocean [12]

The measurements that are considered for the SOCAT database undergo quality control to ensure that the data are of a high quality [62]. The datasets receive a "Wanninkhof Flag", and also each fCO_2 recorded value will receive a "WOCE flag" that is given to the datasets line by line in each data file [62]. The Wanninkhof flags indicate the quality of each cruise by evaluating both the data and meta-data from a specific cruise. When evaluating this, the following criteria is considered [62]:

- The degree to which approved methods or criteria are followed during data collection.
- The completeness of the metadata documentation.

- The acceptability of the second level quality control.
- The degree to which the data set compares to other data.

In order to obtain a good Wanninkhof flag the following needs to be true for the data set: The equilibrator pressure accuracy of 0.5hPa is sufficient for seawater fCO_2 [62]. In the case of high accuracy atmospheric data, the atmospheric pressure (outside the ship) needs to be measured with an accuracy of 0.1hPa [62]. In the case where only outside air pressure is recorded, and used to calculate the seawater fCO_2, the accuracy of 0.5hPa which is required is not met, because of the overpressure that is normally maintained within ships. When calculating the fCO_2 in this case, the 3hPa needs to be added [62].

The data obtained from different sources are recorded in various metadata formats [62]. These metadata sets are seen as complete when they provide information of the following [62]:

- The time, region and method of the reported data.
- The quality of the recorded data should be recorded.
- The list of sensors used on the cruises and the accuracy of these sensors should be mentioned.
- The accuracy of any other data in the data file should be recorded.

CDIAC database

Another source of *in situ* data is the Carbon Dioxide Information Analysis Center (CDIAC) database [4]. CDIAC is the primary Climate-change Data and Information Analysis Center of the United States Department of Energy (DoE) [4]. CDIAC is based at the Oak Ridge National Laboratory (ORNL) of the DoE and the World Data Center for Atmospheric Trace Gasses also forms part thereof [4]. CDIAC works with users globally and focus on the greenhouse effect and global climate change [4]. The CDIAC database contains, amongst others, information of the atmospheric CO_2 concentrations and other relatively active gasses, climate trends, the effect of the terrestrial biosphere and the oceans on one another, the fossil fuel emissions of CO_2 and the biogeochemical cycles of greenhouse gasses [4].

Approximately 4.75 million measurements of the pCO_2 in the ocean surface were collected globally from 1957 to 2009. These measurements are recorded in the Latmont Doherty Earth Observatory (LDEO) database [74]. Open water as well as coastal water data are recorded in this database [74].

Only the data obtained by the equilibrator-CO_2 analyser systems is included in this database, and quality control is applied, by considering the system performance, the degree to which the calibration is reliable, and the consistency of the data [74]. The database is supposed to be updated annually [74]. The LDEO database is available, free of charge, as a numeric package from CDIAC. The data in the LDEO database mainly include the following parameters [74]:

- Partial pressure of CO_2 (pCO_2).
- Sea surface temperature (SST).
- Sea surface salinity (SSS).
- Pressure of equilibration.
- Barometric pressure inside and outside the ship.

A significant number of the measurements that are made for the LDEO database are made by continuous underway systems. These systems are similar to those used aboard the National Science Foundation ice breakers Nathaniel B Palmer and Laurence B Could, which operate primarily in the Southern Ocean [74]. Data is also collected by research ships operated by the Atlantic Oceanographic and Meteorological Laboratory and the Pacific Marine Environmental Laboratory of the National Oceanic and Atmospheric Administration. Additionally, the LDEO database contains data that is obtained from a number of major national and international oceanographic programs, including the Geophysical Sections Experiment (GEOSECS), the Joint Global Ocean Flux Study (JGCFS), the World Ocean Circulation Experiment (WOCE), the Climate Variability(CLIVAR) project and others which are supported by the National Science Foundation, the National Oceanic and Atmospheric Administration and the Department of Energy [74]. A large amount of the data is also gathered from international collaborators including organizations from Germany, Japan, France, UK, Australia, Iceland, Canada, Norway, the Netherlands and others [74].

Takahashi *et al.*, that set up the LDEO data file, used their personal judgement to distinguish between acceptable and unacceptable files [74]. Measurements made at conditions where the water flow stopped or where flow is reduced, are not accepted. If the measurements coincide with rapid changes in temperature of the water in the equilibrator, the measurements are rejected [74]. Where measurements are made using only one calibration gas mixture, there are unspecified uncertainties that could be included, and these observations are also rejected [74]. Considering the complete data set, errors will exist in the database because of the differences in the equilibrator operations, the methods for calibration, and the possible interpolation of some parameters that may vary. It is

assumed that the data in the LDEO database has an uncertainty of approximately 2.5 μatm [74]. The coverage of the ocean by the CDIAC database is shown in Figure 4.4.

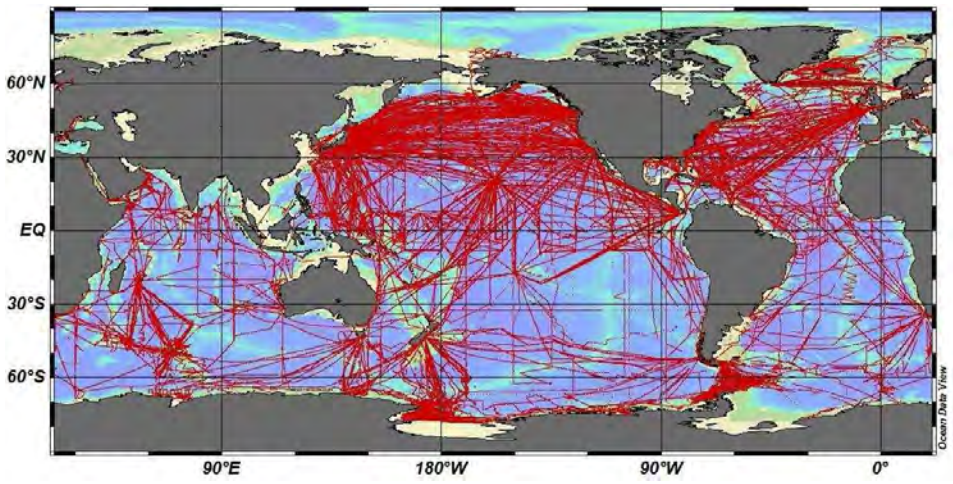

Figure 4.4: LDEO database pCO_2 information coverage of the global oceans [74]

The master file for the LDEO database includes the following elements:

- Cruise ID.
- Station number.
- Latitude (in decimal degrees) where N is assumed to be positive.
- Longitude (in decimal degrees) where E is assumed to be positive.
- Month/Day/Year.
- Julian date in decimal notation.
- Temperature at which the CO_2 was measured (°C).
- Sea surface temperature (°C).
- Sea surface salinity.
- pCO_2 in seawater (in units of micro-atmosphere) at the temperature in the SST column.
- pCO_2 in seawater (in units of pascals) at the temperature in the TEMP column.
- pCO_2 in seawater (in units of micro-atmosphere) at the temperature in the TEMP_pCO2 column, this is the ordinary value that is actually measured.

- Pressure in the equilibrium vessel in units of millibars.
- Barometric pressure in the outside air from ship's observation system in millibars.

During the international geophysical year (IGY), 1957-1960, the infra-red CO_2 gas analyser was introduced for the first time to investigate pCO_2 values aboard ships. Since then, the CO_2 gas analysers have undergone further developments and have been improved to a great extend. The improved analysers allow for measurements to be made more frequently, with greater accuracy, while keeping the basic procedures and principles of measuring the same [74]. In the LDEO database, a collection of high quality pCO_2 data is obtained from measurements of oceanic pCO_2 values that are measured by using the equilibrator-analyser method and processed using a standardized method [74]. Since the reduction method varies between databases, pCO_2 values are assumed to vary by about $1.5\,\mu$atm. These pCO_2 values are not recomputed when entered into the database, but the pCO_2 values are accepted as they are reported to CDIAC [74]. Each group that collected data are trusted to have selected the best data reduction method for their available resources. Where only one calibration gas mixture was used to make the measurements, these measurements are assumed to be unreliable, and are not included in the database [74].

The oceanic pCO_2 values that are included in the LDEO database are from direct measurements made by an equilibrator-CO_2-analyser system. Some measurements in the LDEO database are made in flowing water, whereas others are made from water samples taken at hydrographic stations along the trip. Various methods, equilibrators and gas analysers are used by different groups. Measurements of pCO_2 are accepted irrespective of the different analyser methods, provided that the gas analysers are frequently calibrated, and the CO_2-air gas mixtures and carrier gas are equilibrated with oceanic samples [74].

In a sample of seawater, pCO_2 values at the equilibrium temperature are computed from the reported CO_2 concentration values by

$$(\mathrm{pCO_2})_{equ} = \mathrm{XCO_2}(\mathrm{P}_{equ} - \mathrm{P}_{water}), \tag{4.7}$$

where XCO_2 is the mole fraction of CO_2 concentration in the carrier gas, P_{equ} is the total pressure of the gas in the equilibrator and P_{water} is the equilibrium of the vapour pressure at the temperature of equilibration, T_{equ} and salinity [74]. Some equilibrators are open to the room air. Thus, some P_{equ} values may be pressure values of the interior of the ship, or Barometric pressures outside the ship, depending on where the equilibrator is situated [74]. If no pressure values are reported from

the source of the data, a climatological value is used from the National Centres for Environmental Prediction/National Centre for Atmospheric Research (NCEP/NCAR) reanalysis II project file [74].

The effect of the temperature and the isothermal effect for seawater needs to be taken into account in order to calculate pCO_2 at the *in situ* water temperature. This is given by:

$$(pCO_2)_{sw}@T_{insitu} = [(pCO_2)_{sw}@T_{equ}] \exp\left[0.0433\,(T_{insitu} - T_{equ}) - 4.35{\times}10^5\left(T^2_{insitu} - T^2_{equ}\right)\right], \tag{4.8}$$

where, "*sw*" indicates the seawater *in situ* conditions, "*equ*" indicates the equilibrator conditions, "@T_{insitu}" indicates that this measurement is taken at the *in situ* temperature and "@T_{equ}" indicates that this measurement is taken at the equilibrium temperature. [74]. In these computations, the CO_2 gas is assumed to behave as an ideal gas and it is assumed that it mixes ideally with air and water vapour [74]. To calculate the pCO_2 for air, the following equation is used:

$$(pCO_2)_{air} = (XCO_2)_{air}(P_{baro} - P_{sw}), \tag{4.9}$$

where P_{baro} is the barometric pressure at the sea surface, and P_{sw} is the equilibrium vapour pressure at the temperature and salinity for the mixed layer water [74]. The ΔpCO_2 (or the difference between the air pCO_2 and the seawater pCO_2) can be computed using [74]

$$\Delta pCO_2 = (pCO_2)_{sw} - (pCO_2)_{air}\ . \tag{4.10}$$

Since both seawater and air pCO_2 values are computed assuming that CO_2 behaves like an ideal gas, the effects for the actual non-ideal behaviour of CO_2 is cancelled out. The ΔpCO_2 values indicate the size of the oceanic source or sink. If ΔpCO_2 is a positive value, the ocean is a source for atmospheric CO_2 and it releases CO_2 into the atmosphere. If ΔpCO_2 is a negative value, the ocean is a sink for atmospheric CO_2 and the ocean takes up CO_2 from the atmosphere [74].

The salinity values are measured only at hydrographic stations, and not as regularly as the other oceanic properties. Salinity values are therefore interpolated between the stations in order to have values for all the other corresponding points. Where salinity values are missing, climatological values were used [74].

The sea-air CO_2 flux is determined by [74]

$$(\text{sea-air}CO_2\text{flux}) = (\text{transfer coefficient}) \times (\Delta pCO_2)\ . \tag{4.11}$$

The transfer coefficient is dependent on the degree of turbulence near the surface of the water, which is often expressed as a function of the wind speed [74].

4.2.3 GlobColour dataset

The GlobColour project was initiated in 2005 by the International Ocean Colour Coordinating Group (IOCCG). The aim of this project is to create a long time series (10+ years) of consistently calibrated global ocean colour information with the best possible coverage of the ocean [27]. The IOCCG have proved that the improved methods of merging ocean colour data is sufficient for creating such a database [27]. The GlobColour database consists of a collection of satellite based ocean colour data that aims to improve global carbon-cycle research [27]. The collection of data is collected from the three sensors that are believed to be the most capable ocean colour sensors. These include the SeaWiFS sensor on the GeoEye's Orbview-2 mission, the Moderate Resolution Imaging Spectrometer (MODIS) sensor on NASA's Aqua mission and the Medium Resolution Imaging Spectrometer (MERIS) sensor on the ESA's ENVISAT mission [27].

The error estimates for the initial sensors and the merging processes are included in the dataset, since this will be of significance for a number of dataset users, including modellers who need to include it in ocean simulations [27]. The GlobColour data and merging processes are also validated to ensure that the data and the techniques are sufficiently accurate [27]. The merged dataset from GlobColour will be distributed by the Ocean Colour Thematic Assembly Centre (OC-TAC), who aims to bridge the gap between space agencies who provide ocean colour data and the Global Monitoring and Environment security (GMES) marine applications for ocean colour data [27].

The GlobColour processor is designed to be a stand-alone system, and can operate with the minimum system consisting of a PC with a Linux operating system [27]. The GlobColour processor consists of four primary modules [27]:

- A pre-processor module.
- A spatial binning module.
- A merging module.
- A temporal binning module.

Any data that can be acquired about the global ocean colour, by any means to serve as input to the GlobColour database, are referred to as level-2 data. The level-2 data used as inputs to the

GlobColour processing system, include:

- Envisat MERIS Reduced Resolution with a 1 km resolution for the period 2002-2006.
- MODIS Aqua Ocean Color with a 1 km resolution for the period 2002-2006.
- SeaWiFS Ocean Color GAC with a 4 km resolution for the period 1997-2006.
- SeaWiFS Ocean Color LAC with a 1 km resolution for the period 1997-2006.

The level-2 data serve as input to the processing system, and then the system has the ability to provide level-3 daily merged products within 24 hours [27]. In the processing, the processor can switch between different merging processes as needed, without having to stop or restart the process [27]. Two different groups of merging strategies exist in the processing of the ocean colour data. The first is the merging of bio-optical properties obtained from the different sensors. The second is the merging of the normalised water-leaving radiances, that are obtained from different sensors and methods, and then applying the bio-optical models correctly to determine the data for the ocean colour products [27].

The output products for the period from 1997-2006 with daily, weekly (8-days) and monthly data from the GlobColour processing are as follows [27]:

- GlobColour level-3 product with a 4.63 km equal area grid with integerised sinusoidal projection (ISIN).
- GlobColour level-3 LowRes meteorology product with an equal angle, Plate-Carré projection (PC), 4.6 km grid.
- GlobColour level-3 product with an equal angle, Plate-Carré projection (PC), 4.6 km grid.

The full GlobColour product set includes the following parameters for the period 1997-2007 on a spatial grid of 4.6 km, on a daily, weekly and monthly time scale [27]:

- Chlorophyll-a concentration (Chl).
- Diffuse attenuation coefficient @ 490 nm (K_d490).
- Total suspended matter.
- CDM absorption (aCDM443).

- Particle backscattering coefficient (bbp443).
- Aerosol optical thickness (T865).
- Exact normalised water-leaving radiance @ 412, 443, 490, 510, 531, 555, 620 nm.
- Water-leaving radiance @ 670, 681, 709 nm.
- Data quality flags.
- Cloud fraction.
- Excess of radiance at 555 nm (turbidity index) (EL555).
- Error estimates per pixel for each layer.

The GlobColour dataset is validated to determine the margins of error in the measurement and the processing of data. It is often difficult to validate satellite data with *in situ* data, since the time and spatial grids differ between the two sets of data and because the methods of gathering *in situ* data are not consistent [27]. Additionally, the degradation of the satellite sensors over time needs to be considered in the validation process [34].

The coastal areas are influenced by other factors than the rest of the ocean, and data algorithms are not expected to yield accurate results of ocean parameters in the coastal areas [34]. The overall ocean colour data in the merged GlobColour data set turns out to be more accurate than any one individual sensor's data [34].

The currently available ocean colour merged data sets include [34]:

- REASoN: Input data is obtained from SeaWiFS and MODIS. It uses the GSM01 merging model and yields data for the parameters Chl, CDM, BBP and the uncertainties for the daily products. It has a 9 km spatial resolution and a Daily, 4-Day, 8-Day and Monthly temporal resolution.
- NASA OBPG: Input data is obtained from SeaWiFS and MODIS. The weighted average merging method is used and the Chl is the only product from this merging. It has a 9 km spatial resolution and a Daily, 8-Day, Monthly, Seasonally and Yearly temporal resolution.
- GlobColour: The MERIS, SeaWiFS and the MODIS sensor data is used for input data. The GSM01 model and the weighted average methods are used for merging. There are 19 products in this dataset, with uncertainties in some of them. It has a 4.5 km, 0.25° and 1° spatial resolution and a Daily, 8-Day and Monthly temporal resolution.

From this it can be seen that when validating by comparing to other merged databases, the GlobColour database proves to have a better accuracy, more products and a better spatial resolution [34]. It is seen that whenever more than one set of data from a sensor is used, a better coverage of the ocean is obtained than any one sensor will on its own [34]. The individual sensors cover between 8% (MERIS) and 16% (SeaWiFS) of the ocean daily. When the sensors are combined in sets of two, between 20-25% of the ocean is covered daily. In the GlobColour dataset, where three sensors are combined, approximately 30% of the ocean is covered daily [34].

The merged dataset is validated globally, and it turns out that the merged product set does not degrade the data values from the individual data sets [34]. The GlobColour full product set is verified to be usable in global climate and ocean models [34]. The GlobColour dataset, however, is difficult to validate in the coastal areas and *in situ* data of a higher quality and quantity is required for accurate validation [34].

Some chlorophyll values obtained by the GlobColour processing are shown in Figure 4.5 [27].

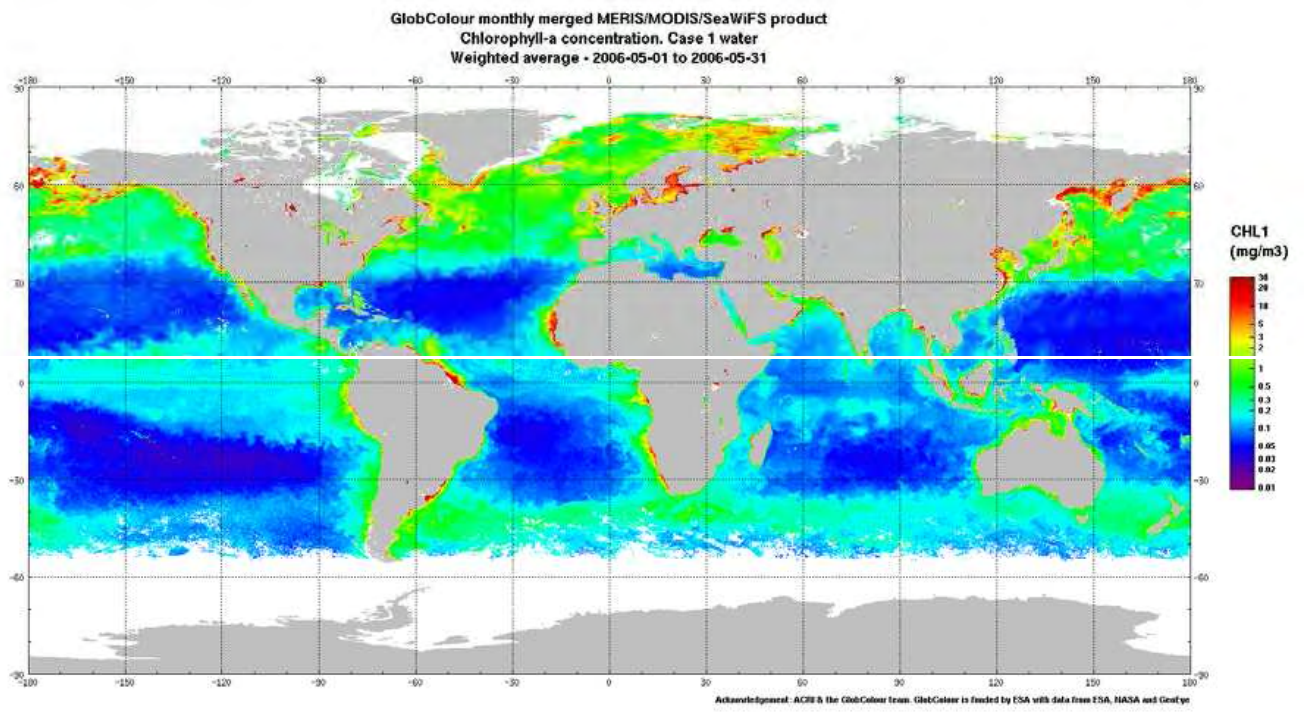

Figure 4.5: The weighted average chlorophyll-a concentration distribution over the global ocean for May 2006 as determined by the GlobColour processing [27]

4.2.4 MLD data

MLD data collection

The ocean is divided into layers according to different temperatures, salinity and lighting in the ocean. The surface ocean mixed layer is linked to the atmosphere and is also the layer in the ocean where primary production takes place. Thus, this layer plays an important role in the carbon cycle and has a significant influence on whether the ocean is a carbon sink or source [24]. The Southern Ocean is predicted to be, on average, the largest annual CO_2 sink [24]. However, observations of ocean properties in the Southern Hemisphere are sparse, and models are needed to estimate missing measurements [24].

The mixed layer is defined differently by researchers, depending on the parameters that are taken into account, including temperature, salinity or density. The mixed layer depth (MLD) is an indication of the temperatures and the biological activity in the layer. The measurements of the mixed layer depth is important in order to allow the validation of the prediction models and satellite measurements [24].

The mixed layer is the ocean layer that is directly in contact with the atmosphere and thus plays an important role in the carbon exchange that takes place between the ocean and the atmosphere [24]. Both the solubility and the biological activity in the mixed layer has an effect on the amount of CO_2 that is absorbed or released by the ocean through the mixed layer [24]. CO_2 is much more soluble in cold water than in warm water. In the water formation regions of the high latitude waters, the water become denser and it sinks, causing the CO_2 that is absorbed in the water to be carried into deeper oceanic waters [24]. Biological activities include the production and termination of organic molecules in water and the effect this has on the carbon exchange [24]. Increased production of the organic molecules takes place at the frontal regions as well as regions where meanders and eddies are observed [24].

The surface mixed layer depth varies on a temporal and a geographical scale [24]. The layer is influenced by heating and cooling as the seasons change. Wave actions also influence the layer, as the tides and winds change. Additionally, eddies and movements at the fronts affect the mixed layer. Both the effect of the atmospheric motions and the oceanic motions on the mixed layer plays a role in the CO_2 exchange that takes place between the ocean and the atmosphere [24]. On a seasonal time-scale, in general the MLD behaves as follows: In summer, the solar isolations is a maximum due to the weak winds, and the MLD is the thinnest during summer months; In the autumn months, the MLD increases when the first storms appear and a small amount of the heat in the mixed layer

is lost; In winter, the heat is lost to the atmosphere due to increasing storms and the mixed layer is even deeper. In late winter, the mixed layer is at its deepest; In spring, the solar irradiation is more, and less vigorous winds allow for the heat to stay trapped in the mixed layer, and the depth of the layer then slowly decreases again [24]. In the summer, the mixed layer is the shallowest and varies between 20 m to 150 m. The deepest mixed layers are formed in winter; as deep as 750 m is found in the Greenland-Iceland-Norway sea and 550 m in the Labrador sea [24].

Both vertical and horizontal motions in the ocean affect the activities in the mixed layer [24]. At the frontal regions, vertical motions in the ocean take place, and often in these regions, there is an increased productivity of organisms and enhanced carbon absorption by the ocean [24]. The Antarctic Circumpolar Current (ACC) spans the globe and assists in the exchange between the ocean basins [24]. The ACC is driven by the westerly winds resulting from the pressure differences over the Antarctic and subtropics [24]. This influences the geographical variations of the mixed layer [24]. Eddies in the ocean affect the flow of the ACC and the heat transport in the ocean and mixed layer [24]. Eddies create poleward heat fluxes that balance out the heat lost in the Antarctic region [24]. The heat transport due to the eddies plays a significant role in the total heat transport variability across the mixed layers [24].

It is common for temperature to be used to calculate the mixed layer depth, but due to the weak temperature gradient south of the Antarctic Polar Front (APF) and salinity that varies due to winter water present in this region, it is often recommended that not only temperature should be taken into account [24]. In this region for example, it is suggested that the density difference will be a better criterion for determining the MLD, since, as the temperature and the salinity changes, the densities of the water change correspondingly [24]. The Southern Ocean in particular has sparse measurements of the oceanic variables, compared to the more abundant measurements in the Northern Hemisphere [24]. There is however an increasing number of measurements of the temperature, salinity and density profiles due to increasing Argo floats and cruises in the Southern Hemisphere [24]. The high resolution data of temperature, salinity and density that is needed to calculate the MLD can be obtained from various sources including cruises and Argo floats [24].

The salinity in the mixed layer is not affected as much by the freshwater inputs into the ocean, but rather by the vertical entrainment and advection. Temperature variations, on the other hand, are greatly dominated by surface fluxes [24]. The variations in the mixed layer are also affected by mesoscale and submesoscale activities; these activities can be observed in the chlorophyll distributions in the ocean [24].

MLD from the CFSR project

The Climate Forecast System Reanalysis (CFSR) is a reanalysis project of the atmosphere, ocean, sea ice and land for the period 1999 to 2009. In this reanalysis project the data sets are updated and revised and the forecast models are re-analysed to improve the quality of the climate forecast system. It was recently completed at the National Centres for Environmental Prediction (NCEP) [80]. There are several improvements in the oceanic part of the reanalysis, including [80]:

- The MOM4 Ocean model with interactive sea-ice.
- The six hourly coupled model forecast as the initial guess value.
- The inclusion of the mean climatological run off.
- The high spatial resolution ($0.5^\circ \times 0.5^\circ$) and high temporal (hourly) model outputs. This results in a high resolution ocean and atmosphere model.
- Observed variations in the CO_2 aerosols and other trace gasses.
- The inclusion of satellite radiances.

The net ocean surface heat flux proves to have smaller biases than in previous reanalysis models [80]. Furthermore, the ocean surface wind stress of the CFSR has smaller biases and higher correlation with the ERA 40 produced by the European Centre for Medium Range Weather Forecasts than previous reanalysis projects [80]. Also, the CFSR proves to have a smaller margin of error compared to other reanalysis projects [80].

The CFSR makes use of a partially coupled ocean and atmosphere data assimilation system [80], [71]. In general, the aim of carrying out this reanalysis is to determine the atmospheric state of the earth over several years by using a constant data assimilation system [71]. The main aim of the CFSR is to provide initial conditions for the ocean, atmosphere and land climate forecast systems [80]. The oceanic component of CFSR is intended to replace current operational ocean analysis system produced by the Global Ocean Data Assimilation System (GODAS). The CFSR is an improvement on the GODAS system in various different areas [71]. Amongst others, the CFSR SST values compare better to the Optimum Interpolation Sea Surface Temperature (OISST) than the GODAS SST [80].

The temperatures used in the CFSR project are measured using expendable bathythermographs (XBTs), fixed mooring arrays, the research moored array for African-Asian-Australian monsoon analysis and prediction in the tropical Indian ocean (RAMA), the prediction and research moored

array in the Tropical Atlantic (PIRATA) and from Argo floats [80]. The salinity data were sparse in the Southern Hemisphere before the Argo floats were introduced. For this reason, the synthetic salinity profiles are included in the CFSR [80]. The synthetic values are based on temperature profiles that are observed, and calculated by the Temperature-Salinity correlation [80].

On average, the CFSR SST values are approximately 0.05–0.1°C warmer than the daily OISST in the tropical Indian Ocean, the western tropical pacific and the tropical North Atlantic [80]. These warmer values are mainly due to the strong South Westerly in these regions [80]. Overall, the CFSR SST values are on average 0.2–0.4°C colder than the daily OISST values, apart from the western boundary currents. The GODAS system on the other hand has regions where the SST differs up to 1°C from the daily OISST values [80].

The sea surface salinity (SSS) plays an important role in oceanic processes and especially in the formation of the Barrier Layer (due to density differences) of the surface layer of the ocean, which directly affects the mixed layer [71]. SSS is calculated in the CFSR project keeping in mind the seasonal and inter annual variability, as well as the rivers that enter the ocean that have an effect on the regional SSS values [71] [80].

The CFSR is validated by various parameters from *in situ* measurements [80]. Monthly climatological temperature and salinity values from the World Ocean Atlas 2005 (WOA05) are used to validate values for the SSS and MLD recorded in the CFSR [80]. Seasonal temperature variations, as they are recorded by the National Oceanographic Data Center (NODC), are used to validate the seasonal temperature in the upper layer of the ocean in the CFSR [71]. Temperature profiles in the NOCD database is up to a depth of 700 m and a geographical resolution of $1° \times 1°$ for the dates from 1955 to 2009 [71]. The temperature data in the NOCD is interpolated to fit the resolution of the CFSR data and the average of the temperature values in the top 300 m of the ocean is taken as the upper ocean heat values [71]. The salinity values in the CFSR are also validated using values from the NODC database [71].

MLD values for global oceans are calculated using individual temperature and salinity values in the CFSR. The climatology of MLD is based on the annual climatology values for temperature and salinity, using a threshold value of a density difference of $0.25\,\mathrm{kgm}^{-3}$ between the surface and the depth of the mixing layer. Thus, at the point where the density in the ocean differs with $0.25\,\mathrm{kgm}^{-3}$ from the density at the surface of the ocean, the mixing layer is said to end, and the MLD is at this depth [80]. When the MLD values in the CFSR are compared to the MLD values in the WOA05, it is found that the values in the CFSR and the GODAS are approximately 10–20 m deeper over a large area in the tropics [80]. This could be due to uncertainties in the heat and salinity variations

at the surface or incorrect modelling due to sparse measurements [80]. It should be noted that the WOA05 MLD values are calculated by sparse measurements in the ocean, while the GODAS and CFSR MLD values are calculated using monthly values for the period from 1982 to 2004 [80].

The MLD temperature is in general primarily affected by vertical heat flux from the atmosphere [24]. Turbulence in the ocean affects the mixing in the mixed layer as well as the formation of layers in the ocean, affecting the mixed layer [24]. In the deeper ocean, below the mixed layer, the density field is primarily influenced by salinity. Vertical variations in the mixed layer results in a consistent overall salinity in the mixed layer, because fresh water entering the mixed layer from rain, is balanced out by salt water taken up by the mixed layer from deep waters [24].

4.2.5 Satellite data

Satellites are one of the sources of measurements of oceanic properties. To measure sea surface temperature (SST) from space, a high performance radiometer is required as well as effective atmospheric correction [51]. Measuring ocean properties from space has the advantage of coverage, consistency and continuity [51]. Climate is strongly temperature dependent. Thus, accurate observations of SST are required, since a change of the ocean temperature of about 0.3 °C can have a significant impact on the climate [51]. The satellites are required to measure SST irrespective of cloud cover, since the Southern Ocean is covered with clouds from June to August [55].

One of these satellites is the Advanced Very High Resolution Radiometer (AVHRR). The AVHRR was designed in the 1960's as the world's first general access Earth Imager [51]. It is currently used by the Naval Oceanographic office to retrieve sea surface temperature values [42]. It contains a telescope, to define the field of view (FoV), as well as filters to define the spectral response of detectors [51]. The AVHRR makes use of infra-red sensors to enable readings during daytime as well as night-time [42]. The AVHRR is designed for the following tasks: Firstly, channels one and two of the AVHRR are used to recognize clouds, to determine boundaries between land and water, to establish regions of ice in the ocean, to monitor melting snow and to keep track of terrestrial vegetation [42]. Secondly, channels three, four and five are used to measure temperatures of the clouds and sea surface, as well as to map clouds at night time [42].

The AVHRR does not fly alone on the TIROS-N satellites, but flies in conjunction with the TIROS-N operational vertical sounder (TOVS). The TOVS is a combination of a High-resolution Infra-red Radiation Sounder (HIRS), as well as a microwave sounding unit and a Stratospheric Sounding Unit (SSU). The combination of these three units results in a collection of microwave and infra-red

channels that can be used to set up a profile for atmospheric temperature and to a small extent atmospheric moisture [42]. Infra-red measurements can be made at a resolution of 1 km, but cannot be made in cloudy areas and are problematic when aerosols are present [51]. Microwave observations on the other hand cannot be MADE near coast lines, are problematic when there is radio interference and rain and has a low spatial resolution of 20 km–100 km, but is not greatly affected by clouds [51].

Early in the operation of the AVHRR, it was determined that one of the easiest variables to determine using the infra-red sensors is sea surface temperature (SST) [42]. Originally, in the VHRR and the AVHRR-1 a single thermal infra-red channel was used to measure SST [42]. By adding the 11 micrometer channel in the AVHRR used currently, the atmospheric attenuation of the SST signal due to the presence of water vapour in the atmosphere is taken into account [42]. These are now channels four and five in the AVHRR [42]. These measurements of SST need to be calibrated. Currently it is being calibrated using SST values obtained from drifting buoys in the ocean [42]. It can also alternatively be calibrated using SST measurements made by a radiometer fixed to ships [42].

The infra-red measurements of sea surface temperature can also be used successfully to determine currents and the way in which these currents vary [42]. This is because the currents can be observed by investigating gradients of sea surface temperature [42].

The polar orbit in which the AVHRR orbits, as well as the 300 km path of the AVHRR, results in a significant overlap of observations in the polar regions [42]. This is crucial to deal with the clouds that cover these areas which often obstruct the infra-red sensor and reduce the amount of accurate readings [42].

The advanced along track scanning radiometer (AATSR) is an improved observational method from the AVHRR in the sense that it is an independent observation system [51]. Unlike the AVHRR, the AATSR does not need SST values from buoys to calibrate the readings for the ocean surface temperature [51]. The AATSR uses global validation, by using continuous global checks by comparing observations from readings to the following: data from drifting buoys; other satellites' observations; measurements from radiometers fixed on ships; and precise observations at selected sites [51]. From this validation, it is shown that the AATSR obtains an accuracy of $\pm 0.1^{\circ}$C in many different regions [51]. Some of the problems of the AATSR come in when determining the presence of aerosols and clouds [51].

4.2.6 Argo Floats

Argo floats are more evenly distributed across the ocean and therefore yield a dataset with a more even distribution of points [39]. Argo floats do not measure pCO_2 values of the ocean, but instead measures sea surface temperature (SST) and sea surface salinity (SSS) quite accurately over a period of approximately 10 days [39]. The information about the SST and the SSS can be used to calculate the MLD values of the ocean at the Argo float positions [39]. Thus, the Argo floats collect data on some of the variables that controls the CO_2 fluxes in the ocean [39].

Since Argo floats decrease the area of the ocean where no observations are available, it contributes to much better estimates of the ocean-atmosphere CO_2 fluxes [39]. Argo floats often give a better coverage of the ocean than satellites, since satellites are often unable to make readings due to clouds covering the ocean or low solar irradiation at high latitudes [39]. The CURRENT Argo float distribution is shown in Figure 4.6.

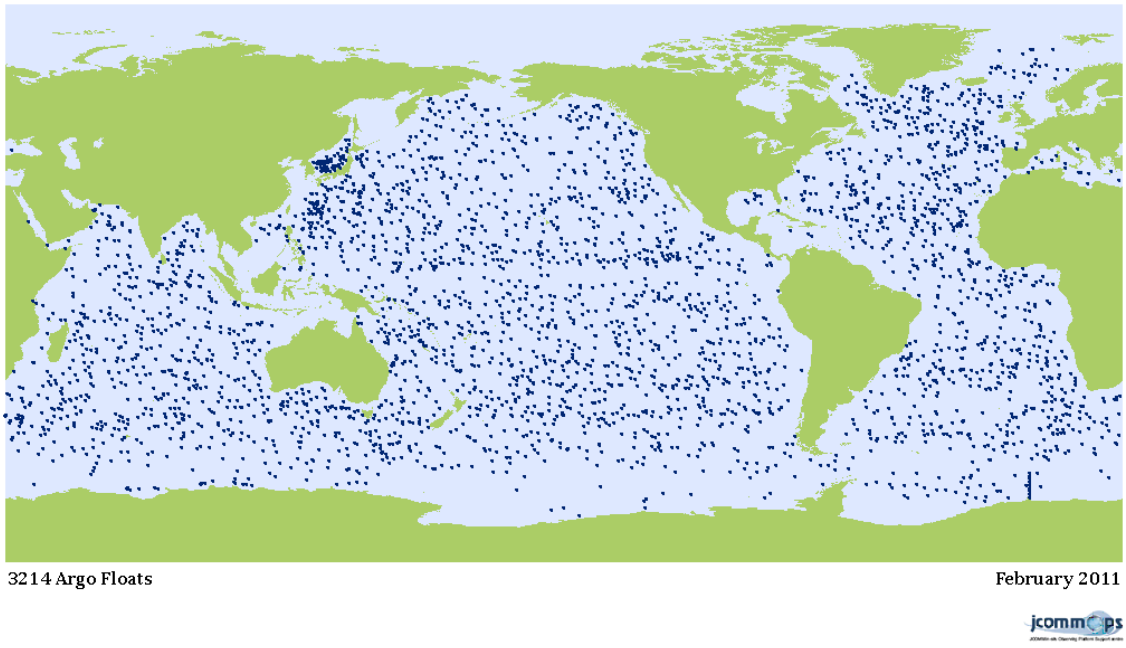

Figure 4.6: ARGO floats distribution in the ocean [2].

The mixed layer depth (MLD) at Argo float positions can be derived from the SST and SSS values, but there are currently no accurate way to estimate the MLD values of the open ocean [39]. If an accurate measurement of MLD was available, it is shown that the RMS error when calculating the annual oceanic pCO_2 values can be decreased by 10%, when including MLD as well as SST and SSS in the estimations [39].

By additionally also taking into account chlorophyll values when calculating annual pCO_2 values of the ocean, it is shown that the RMS error of estimating pCO_2 values can be decreased by another 12% [39]. Chlorophyll data is an indication of the biological processes affecting the ocean dynamics, and the information on chlorophyll also allows for more accurate extrapolation across the ocean surface [68].

4.2.7 Conclusion

There are many different sources of data for the ocean, including *in situ* measurements from several different cruises conducted in the ocean, various satellite measurements and floats spread out across the ocean. These sources have been validated and verified in order to check the accuracy of the data and to ensure that the most accurate measurements and data sets are used in modelling. Different data collection methods and their corresponding data sets were discussed in this chapter. The *in situ* measurements are known to be the most accurate measurements, as they are measured by sampling actual ocean water and measuring ocean properties from this. The other datasets used are in most cases validated against *in situ* measurements.

The CDIAC and SOCAT data sets consist of *in situ* measurements collected during various different cruises conducted across the entire ocean. The GlobColour data set measures oceanic properties by analyzing the ocean colour. We use the chlorophyll values from this data set. The MLD data is obtained by deriving the MLD values from temperature and salinity values otherwise collected. Another data collection method is Argo floats, which are distributed across the entire ocean, and collect data of various parameters throughout the year.

By using a collection of these datasets, modelling of the ocean and the investigation of its properties can be done more accurately.

4.3 Data processing

4.3.1 *In situ* data

The *in situ* data used in the first stage of the modelling was gathered during the SANAE49 cruise in 2009-2010. The *in situ* data is assumed to be the most accurate data and the initial modelling is done using this dataset. The *in situ* data set contains, amongst others, measurements of pCO_2, MLD, SST (or intake temperature, which is the temperature at which water samples are taken),

latitude and longitude values and chlorophyll-a concentrations. These are the parameters needed to carry out the investigations in this dissertation.

Interpolation of MLD values

The MLD measurements for the SANAE49 cruise was only taken every ∼20 nautical miles during the cruise and every ∼10 nautical miles in the frontal zones. This is a lower resolution than the measurements made for other parameters during the cruise, which were made at smaller spatial intervals. Furthermore, the MLD data were measured with both XBT and UCTD instruments at different latitudes. The XBT and UCTD are different instruments used to measure MLD, but will yield the same MLD values if they were used at the same time in the same point in the ocean. The data from the XBT and UCTD are combined to form a dataset for MLD values for the SANAE49 trip. They are simply combined by putting the two sets together in the same matrix and sorting the values so that they appear in the matrix from a high latitude value to a low latitude value.

Because the MLD values for the SANAE49 cruise are not at the same latitude and longitude positions as the other parameters, some technique is needed to obtain estimates for MLD values at the positions where the other parameters are measured. It was decided that a weighted average of *in situ* MLD values of the points closest to the required point (from both above and below the point in space) will be used to interpolate a MLD value for a point along the cruise line. The point closest to the required point in both the northern and southern latitude direction have to be determined. The point closest to the required point is defined as the point with the closest Euclidean distance to the required point. Once the north and south closest point is determined, the total distance between the two points via the target point is calculated. This total distance is then used to assign a weight to each of the two points in order to calculate a single MLD value for the point required. The point closer to the target point is assigned a heavier weight and the point furthest from the target point is assigned a lower weight such that the two weights add up to one. Let B be the target point, the point at which we want to calculate the MLD value. Then let A be the point north of B and let C be the point south of B. Let the distance AB be equal to D_1 and let the distance BC be equal to D_2. Then the MLD at the point B is calculated by Equation (4.12).

$$\text{MLD}_B = \frac{D_2}{D_1 + D_2} \cdot \text{MLD}_A + \frac{D_1}{D_1 + D_2} \cdot \text{MLD}_C. \tag{4.12}$$

Table 4.1: Summary of statistics of variables in the final SANAE49L6 data set.

	N	Missing	Mean	SD	Min	Max	Q1	Median	Q3
pCO2	6103	2	360.19	37.72	251.19	435.98	351.2	368.62	380.18
Salinity	6105	0	34.16	0.55	33.36	35.69	33.82	33.98	34.18
Chl	6105	0	1.16	1.23	0.12	5.14	0.46	0.62	1.44
SST	6105	0	6.3	5.68	-0.28	21.3	2.66	3.61	8.37
MLD	6105	0	61.59	24.92	13.15	127.93	42.09	55.85	82.45

Data cleaning process

The data from the SANAE49L6 cruise is used to do the initial modelling. This is the SANAE cruise from December 2009 to February 2010. The L6 refers to Leg 6, which is the trip Northwards from Antarctica back to Cape Town. The original dataset obtained for this part of the cruise started on 12 February 2010 (GPS time 00:04:48) and ended on 22 February 2010 (GPS time 23:55:54). During this part of the cruise, the boat travelled between the coordinates (-70.6245,-0.0001) and (-34.073,17.4585). Only the rows in the original data set that had "Type=EQU" or "Type=EQU-DRAIN" are used for further data processing and analysis. The variables that are of interest for determining the correlations between parameters in the ocean in this case include the pCO_2, salinity, intake temperature, chlorophyll-a concentration and MLD.

In the original data file, there were three latitude locations, around 60°S, 50°S and 40°S that had spikes in more than one variable. Extensive checks have been done to find the points in the data where the spikes occur. It is confirmed that the points at which these errors occur are points where errors in measurements occurred, and these lines are removed from the original data set. The data points below the most South MLD value, as well as the data points North of 37°S are deleted from the dataset. The interpolation was done on the MLD data set in order to obtain MLD values at all the points where the other parameter values are available for the trip. Once the data sets are merged and interpolation is carried out, the data set has 6103 rows. The start date in this particular data set is 13 February 2010 (GPS time 18:07:50) and end date 21 February 2010 (GPS time 18:30:53). The coordinates for this cruise in this data set vary between (-69.5998,-5.9036) and (-37.0004,12.918). A summary of the statistics of the variables are given in Table 4.1 and the data is shown in Figures 4.7, 4.8, 4.9 and 4.10. A multivariate plot of the data is given in Figure 4.11 which gives an indication of how the different variables relate to one another.

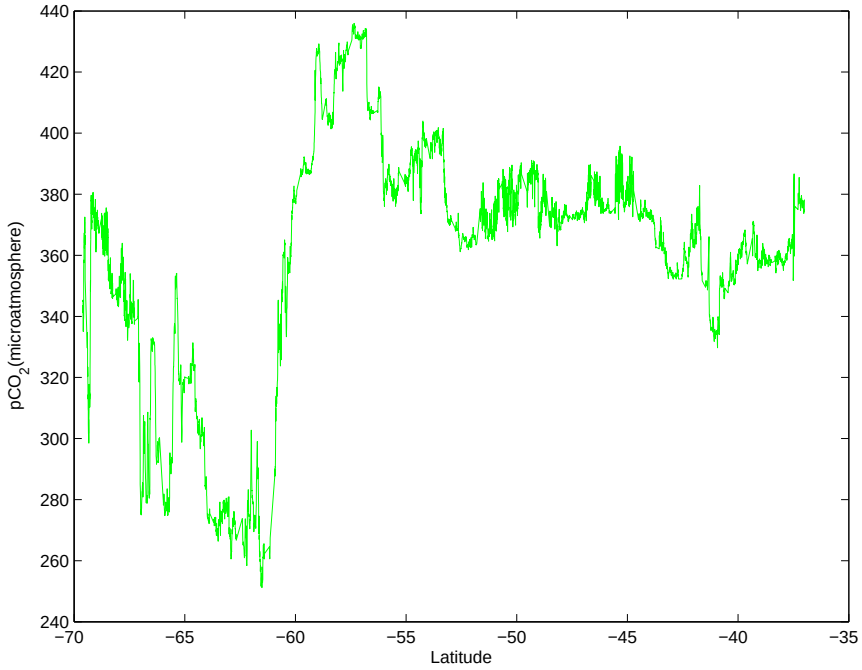

Figure 4.7: The pCO_2 data in the SANAE49L6 data set.

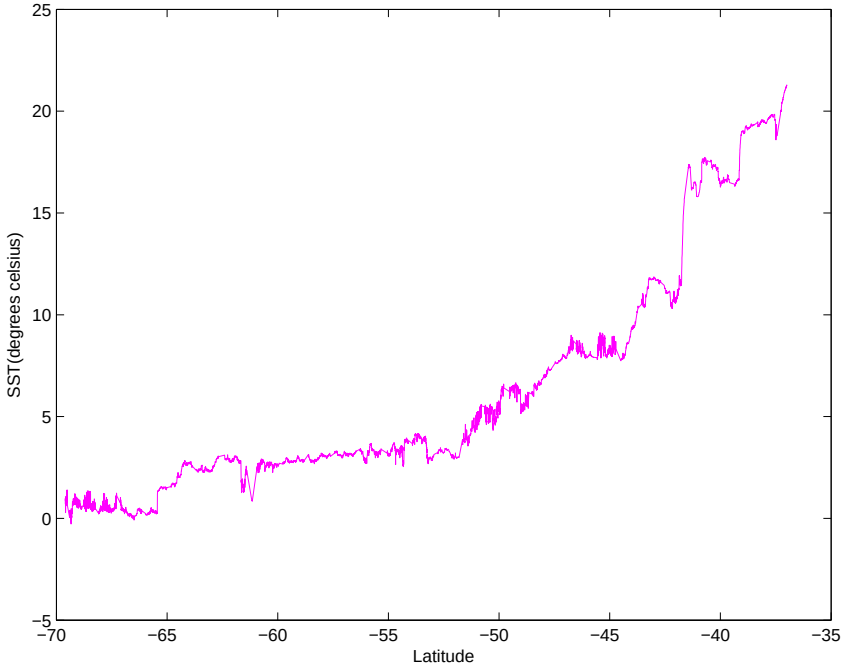

Figure 4.8: The temperature data in the SANAE49L6 data set.

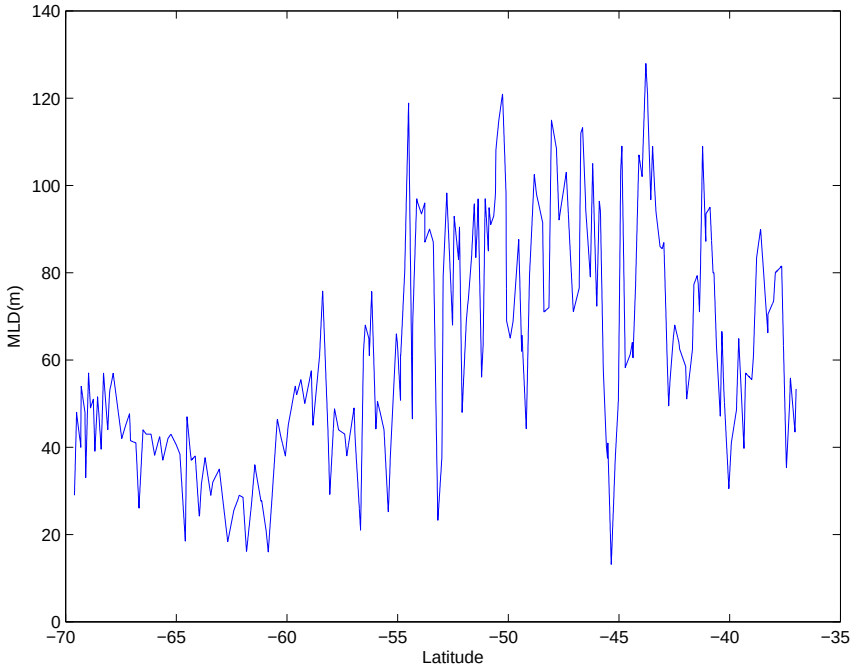

Figure 4.9: The MLD data in the SANAE49L6 data set.

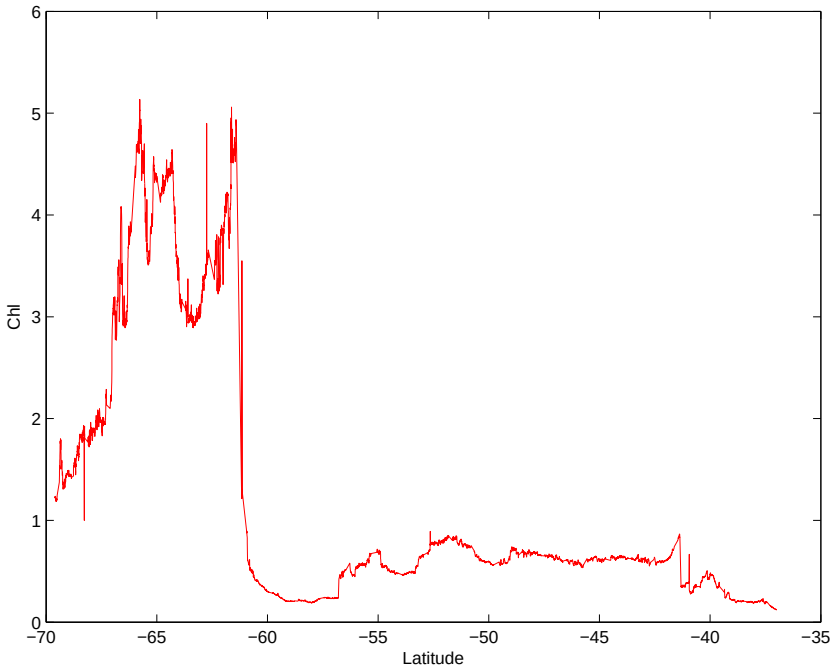

Figure 4.10: The chlorophyll-a concentration in the SANAE49L6 data set.

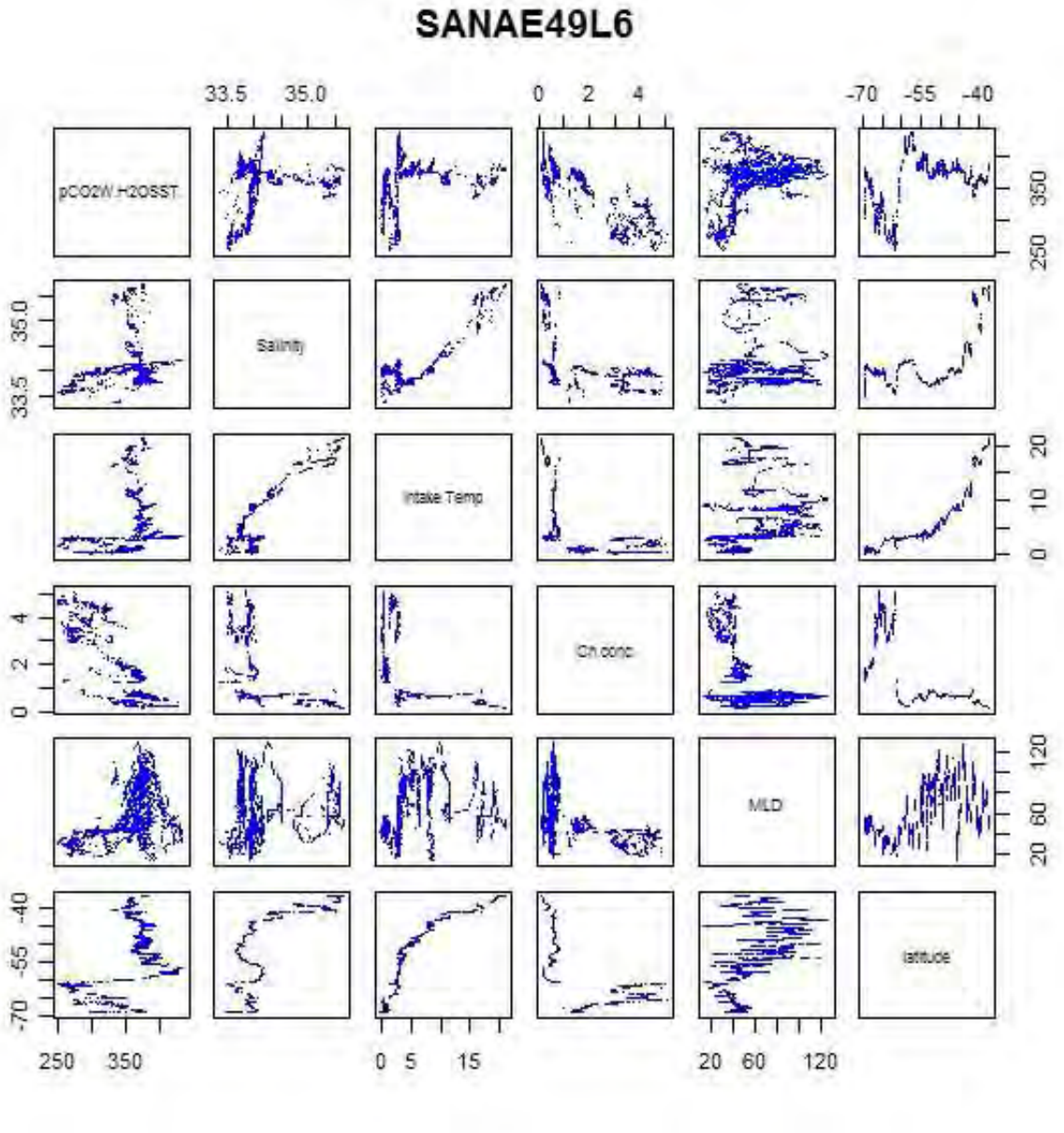

Figure 4.11: Multivariate plot of the SANAE49L6 data set.