Estimating the spatial and temporal variability of primary production from a combination of *in situ* and remote sensing data: A southern Benguela case study

by

Robert Williamson

Thesis presented for the degree of
DOCTOR OF PHILOSOPHY

Department of Oceanography
Marine Research Institute
University of Cape Town
August 2013

Supervisors

Prof. J.G. Field
Prof. F.A. Shillington
Prof. A. Jarre
Dr. A. Potgieter

Plagiarism Declaration

I know the meaning of plagiarism and declare that all of the work in the thesis, save for that which is properly acknowledged, is my own.

Signed

Acknowledgements

I would like to thank the following individuals, who all had a significant role to play in the completion of this study:

My supervisor, Prof. John Field (Marine Research Insitutue - UCT) for constant guidance and my co-supervisors Prof. Frank Shillington (Department of Oceanography - UCT), Prof. Astrid Jarre (Marine Research Institute and Department of Biological Sciences - UCT) and Dr. Anet Potgieter for their unwavering support.

I would also like to thank Dr. Christo Whittle, Dr. Tarron Lamont and Dr. Stewart Bernard for their useful and much appreciated inputs.

A special thank-you to family and friends who constantly encouraged me along the way.

Financial support was provided by: the SEAChange Project of the South African Network for Coastal and Oceanic Research, funded by the former branch Marine and Coastal Management and the National Research Foundation; and by the South African Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation, through the Research Chair in Marine Ecology and Fisheries. Part funding was further received from the UCT Vice-Chancellor's Strategic Initiative through the Marine Research in the Benguela and Agulhas Systems for supporting Interdisciplinary Climate-Change Science (Ma-Re BASICS) programme, as well as through the Marine Biology Research Centre of the Department of Biological Sciences.

Table of Contents

Plagiarism Declaration i
Acknowledgements ii
Table of Contents iii
Abstract vii
Chapter 1 - General Introduction 1
1.1 Historical background of studies on phytoplankton 1
1.1.1 Understanding the physiology and distribution 1
1.1.2 Estimates of primary production 3
1.1.3 Subdividing the ocean in characteristic production zones 4
1.1.4 Characteristics of the coastal zone 5
1.2 The Benguela upwelling system 7
1.2.1 Physical forcing and biological response in the upwelling system 9
1.2.2 Climate change impacts on eastern boundary upwelling systems 18
1.3 Relating the vertical distribution of Chl *a* to environmental variables 19
1.4 Statistical methodology to recognize patterns 21
1.5 Thesis structure 23
Chapter 2 - Data Acquisition, Pre-processing and Clustering 26
2.1 Introduction 26
2.2 Data and methods 27
2.2.1 *In Situ* data 27
2.2.2 Satellite data 34
2.3 Results 39
2.3.1 *In Situ* data 39
2.3.2 Satellite data 49
2.4 Discussion 64
2.4.1 *In situ* data 64
2.4.2 Satellite data 69
2.5 Conclusion 73

Chapter 3 - Developing a Static Bayesian Network 74
3.1 Introduction 74
3.2 Data and methods 76
3.2.1 Remote sensing data 76
3.2.2 Discretizing continuous data 77
3.2.3 Identifying characteristic profiles 79
3.2.4 Experiment 1: Predicting surface Chl *a* from satellite data 81
3.2.5 Experiment 2: Relating vertical profiles to surface variables 83
3.3 Results 84
3.3.1 Experiment 1: Predicting surface Chl *a* from satellite data 84
3.3.2 Experiment 2: Relating vertical profiles to surface variables 86
3.4 Discussion 91
Chapter 4 - Classification Using Time Series Data 93
4.1 Introduction 93
4.2 Pre-processing wind data 95
4.3 Classification with sequences 100
4.3.1 Developing a classifier 100
4.3.2 Incorporating sequence data 102
4.3.3 Rule scoring algorithms 103
4.3.4 Classifying surface variables to profile classes 106
4.3.5 Results 108
4.4 Predicting missing values from sequences 110
4.4.1 Interpolation 110
4.4.2 Hidden Markov Models 115
4.4.3 Conditional Random Fields 117
4.5 Conditional Random Fields for missing values 118
4.5.1 Feature functions 118
4.5.2 Potential functions 122
4.5.3 Calculating conditional probabilities 123
4.5.4 Training phase 124
4.5.5 Testing phase 131
4.5.6 Results 132

4.6 Revisiting the classification model 134
4.7 Conclusion 138
Chapter 5 - Estimation of Daily Primary Production 141
5.1 Introduction 141
5.2 Methods 143
5.2.1 Photosynthesis parameters 143
5.2.2 Attenuation of surface PAR 144
5.2.3 Depth of the euphotic zone 147
5.2.4 Daily variability of PAR 148
5.2.5 Primary production model 149
5.3 Results and discussion 151
5.3.1 Attenuation of PAR 151
5.3.2 Euphotic depth 153
5.3.3 Validation of daily primary production 155
5.4 Conclusion 157
Chapter 6 - Time-Series of Depth-Integrated Primary Production 159
6.1 Introduction 159
6.2 Methods 161
6.3 Results 165
6.3.1 Frequency distributions of spring profiles 167
6.3.2 Frequency distributions of autumn profiles 172
6.3.3 Depth distribution of primary production 176
6.3.4 Depth-integrated production 182
6.3.5 Modelled versus *in situ* primary production 184
6.3.6 Annual primary production 185
6.4 Discussion 187
6.4.1 Spring processes 187
6.4.2 Autumn processes 188
6.4.3 Modelled and observed profile distributions 190
6.4.4 Modelled primary production 195
6.4.5 Long-term primary production 203
6.4.6 Annual primary production 207

6.5 Conclusion 208
Chapter 7 - Conclusion 211
References 218

Robert Williamson

Estimating the spatial and temporal variability of primary production from a combination of in situ and remote sensing data: A southern Benguela case study.

July 2013

Abstract

The aim of this thesis is to produce fine resolution estimates of primary production in three-dimensional space at the temporal scale that these events develop. It is hypothesized that complex relationships among time sequences of physical and biological processes that influence primary production can be automatically discovered from archives of data.

This study uses an archive of *in situ* ship-board data containing subsurface temperature and phytoplankton distribution profiles. Each profile is associated in time and space with satellite remotely-sensed wind, sea surface temperature and surface chlorophyll *a* data. The bottom depth, season and location of each profile are also recorded. The archive of depth profiles is simplified by mapping each profile onto one of twelve representative profile clusters obtained using the *k*-means clustering algorithm so that each cluster contains a set of similar profiles and their corresponding data. Relationships between remotely sensed surface features and chlorophyll *a* profiles are first obtained from a static Bayesian network using same day data. This is then taken further by analysing time-series of satellite data to predict likely temperature and chlorophyll *a* profiles for each pixel of a 4 km resolution satellite image.

The time-series of data associated with each profile and their influence on the profile shape is determined using statistical models based on Bayesian theory and methods based on rule discovery. These sequences are then used to predict the profile shape. The predicted chlorophyll *a* profile is used in a primary production model to estimate hourly production in the water column of the St Helena Bay sub-region. The model is run over sixteen months to illustrate the short-term variability of primary production on a fine scale.

The total annual production estimated by the model shows very good agreement with other estimates made in the Benguela system. An average annual production rate of 3.2 $gC.m^{-2}.day^{-1}$ is obtained for the sub-region. The model differs from previous methods for estimating annual production in that the primary production estimates are three-dimensional and derived from physical processes, which allows for analysis of complex tropho-dynamic relationships and impacts of hypothetical climate scenarios.

Chapter 1 - General Introduction

A major goal of biological oceanography is the determination of temporal and regional changes in phytoplankton production in the world's oceans. Research in this field can be attributed to the combined efforts that discovered photosynthesis and those by Ørsted (1844) and Hooker (1874) who independently discovered in the mid 19th century that unicellular planktonic plants support oceanic food webs. The energy transfer processes through these webs have vital implications for all ecosystem components as well as for fisheries yields. Understanding phytoplankton growth within its environment and what controls its abundance, distribution and productivity at various temporal scales continues to be the focus of much research. Besides the obvious implications for marine food webs, in the past century phytoplankton have become a major focal point regarding the recycling of nutrients, the flux of carbon between the ocean and atmosphere and the transfer of organic matter from the ocean surface to the ocean floor (for example, the Joint Global Ocean Flux Study (JGOFS) and the International Marine Biogeochemistry and Ecosystem Research (IMBER) programmes).

1.1 Historical background of studies on phytoplankton

1.1.1 Understanding the physiology and distribution

Work by Brandt (1899) and Nathansohn (1906) proposed nutrients as a major regulator of phytoplankton productivity and highlighted the importance of upwelling and mixing for the supply of subsurface nutrients. In the present study, phytoplankton productivity refers to the rates of carbon fixation by a single organism

or per unit biomass, whereas primary production refers to total organic carbon produced by a particular biomass per unit area or volume and per unit time. The variability of upwelling and mixing from open-ocean to coastal waters was hypothesized by Gran (1912) to determine the spatial variability of production. Following this, studies by Atkins (1928) and Marshall and Orr (1928) highlighted seasonal changes in water column structure and nutrient supply and the need for the thermocline to retain phytoplankton in the sunlit and initially nutrient-rich upper layer. Sverdrup (1953), quantifying the earlier work of Gran (1931), incorporated the interactions between light, nutrients, mixing and stability to explain the onset of the North Atlantic spring bloom. Nutrients, light and temperature control the individual algal cell growth whereas environmental conditions, which control water column stability, advection, grazing and sinking, control population growth through the control of concentrated layers.

In 1952, Steemann Nielsen (1952) introduced the radiotracer ^{14}C method to replace the labour-intensive oxygen titration method for quantifying photosynthesis. Using this method, Koblentz-Mishke *et al*. (1970) were able to product the first global map of oceanic primary production from over 7000 primary production incubations. However, the ^{14}C method was not without problems, particularly those that arose from long incubations. Alternative short-term methods based on ^{14}C uptake were explored and resulted in the first photosynthesis versus irradiance (P vs. E) curves (Platt *et al*. 1975). In the late 1980s fluorescence based measurements were introduced (Falkowski *et al*. 1986; Kolber *et al*. 1990), which allowed continuous measurements of photosynthetic activity although it still required the radiocarbon approach for calibration.

Esaias (1980) recognised that ocean colour potentially could be used to derive global maps of oceanic chlorophyhll concentration. The advent of satellite oceanography was a major leap forward for biological oceanography as it enabled the knowledge of the physiological controls of phytoplankton growth to be applied across the world's oceans. This led to a new understanding of the horizontal variability and abundance of the organisms on a global scale at almost daily frequencies. Images obtained from 1978 by the Coastal Zone Colour Scanner (CZCS) sensor on board the NIMBUS satellite, provided biological oceanographers with an unprecedented look at regional and seasonal variations in ocean colour, representing near-surface phytoplankton chlorophyll *a* (Chl *a*) concentration.

1.1.2 Estimates of primary production

Models of net primary production for large spatial scales require images of ocean colour, the incident solar irrandiance and the photosynthetic rates (Behrenfeld and Falkowski, 1997). They may also require SST and the mixed layer depth (MLD). Models can be classified according to their level of integration but all are fundamentally similar (Behrenfeld and Falkowski, 1997): The models use ocean-colour from satellites to estimate biomass and incident solar photosynthetically available radiation (PAR) to derive a vertically integrated estimate of primary production for each pixel or pixel set of the satellite image. Numeous models have been developed to estimate global production but many have been focused on the open ocean where Chl *a* concentrations are typically < 1.0 mgChl.m^{-3} and as a result perform poorly at continental margins where concentrations exceed ca. 5 mgChl.m^{-3}. A number of comparative studies have been performed to evaluate the most current models for predicting primary production, for example the Primary Productivity

Algorithm Round Robin (PPARR; Campbell *et al*. 2002; Carr *et al*. 2006; Friedrichs *et al*. 2009; Saba *et al*. 2010; Saba *et al*. 2011). Similar findings were obtained by Carr *et al*. (2006) and Saba *et al*. (2011), where the models were tested on a variety of oceanic regions characterised by their biophysical properties. In particular, almost all the models over-estimated primary production in shallow coastal water and struggled to capture the variability.

Many of these models are parameterized by in situ point measurements for parameterization which are few in number or may depend on average parameters specific to region, biome or regime.

1.1.3 Subdividing the ocean in characteristic production zones

Studies by Platt and Sathyendranath (1988) and Platt *et al*. (1991), suggested that regions with characteristic seasonal physical variations that affect the depth of the mixing layer could be used to describe local biological dynamics. Longhurst *et al*. (1995) created a biogeography of the global ocean based on how the ecology of plankton responds to regional oceanography. The authors initially used the forcing described by Sverdrup (1953) *viz*. forcing of the mixing layer depth by local wind and irradiance at the surface, but found this insufficient to describe the dynamics in all regions. They recognised four primary domains by their characteristic forcing, some of which required additional factors; the seasonal cycle of sea ice in the *Polar Domain*; the westerly wind stress from quasi-permanent low-pressure cells in the *Westerlies Domain*; the frictional wind stress by the seasonal trade winds in the *Trade Winds Domain*; and the modified oceanic circulation from interaction with the coastal topography and coastal wind regime in the *Coastal Domain*.

Regions do not only differ according to their physical dynamics but as a consequence, the vertical distribution of phytoplankton. Sathyendranath *et al.* (1995) found that the relationship between surface Chl *a* and Chl *a* as a function of depth varied significantly to warrant the use of biogeochemical provinces that describe the reality of seasonal phytoplankton growth as initially proposed by Platt and Sathyendranath (1988). Surface Chl *a* values, to some extent predict the subsurface profile but it is more useful for analytical models to have depth related concentrations. Sathyendranath *et al.* (1995) classified the North Atlantic Basin according to the differences in the physical environment that would most likely have an influence on regional algal dynamics. Within each of these provinces the authors established seasonally differentiated photosynthesis-light curve parameters and parameters that determine the vertical structure in the Chl *a* profile. Thus, they advocate the need for both satellite data to resolve the temporal and spatial issues and *in situ* data to resolve the dynamics below the surface.

1.1.4 Characteristics of the coastal zone

The mixing layer in temperate coastal zones (which are described as extending from the outer edge of the continental shelf to the high water mark), as with the open ocean, are subject to seasonal cycles of warming and cooling and variations in its depth. Here, processes affecting biological productivity are more complex than the open ocean due to relatively shallow depths, tidal currents, fronts and the proximity to the coast (Mann and Lazier, 1996). In shallow water the mixing layer may frequently extend to the bottom and entrain accumulated nutrients into the water column making them available for photosynthetic production (Mann and Lazier, 1996). Tidal currents may cause turbulence in shallow water as they flow over

bottom topography; sometimes this turbulence reaches the surface. Internal waves generated by the tidal flow typically propagate along the pycnocline when vertical stratification is strong. The oscillating displacement of the pycnocline may cause vertical transport of nutrients and movement of phytoplankton into a more brightly lit environment (Lande and Yentsch, 1988). Fronts mark a sharp gradient in water properties such as temperature, salinity and density. They may form at the shelf-break where cooler mixed continental shelf waters meet the warmer stratified open ocean water or where the offshore Ekman transport of upwelled water meets resistance from a stratified offshore water mass and sinks below it (Mann and Lazier, 1996). Frontal regions are almost always regions of enhanced primary production as a result of the convergence of two different water masses (Shannon, 1985). Proximity to the coastline may be an important property of coastal zones as the coastline acts as a barrier to advection during along-shore winds and can result in either upwelling or downwelling close inshore depending on the wind direction. Some of the most biologically productive regions in our oceans are found where wind-driven upwelling is the dominant method of replenishing nutrient concentrations in the upper layers (Ryther, 1969; Cushing, 1971; Mann, 2000).

The concentrated food environment of coastal upwelling areas provides good spawning grounds for pelagic fish and for fish productivity (Crawford *et al*., 1980; Longhurst *et al*., 1995; Boyd *et al.*, 1998). Production can vary between upwelling systems as well as spatially within each region in response to variations in wind strength, bathymetry, latitude and the hydrographic properties of the water column (Mann and Lazier, 1996). The variation in production, particularly on inter-annual time scales is of interest as the availability of phytoplankton may limit the occurrence and productivity of organisms at higher trophic levels (Borchers and Hutchings,

1986). The region of interest for this study is a particularly productive upwelling area, the Benguela upwelling system.

1.2 **The Benguela upwelling system**

The Benguela Current is one of the four major eastern boundary current regions of the global ocean (Wooster and Reid, 1963) and is similar to those off California, Peru and North West Africa (Shannon, 1985). The Benguela upwelling system consists mainly of the cool Benguela Current and is bounded by the warm Angola Current to the north and the warm Agulhas Current to the south. The slow, wide northward drifting current is driven by the equatorward winds around the semi-permanent South Atlantic atmospheric anticyclone (Shannon, 1985). The equatorward flow is also carried in a series of jet currents and eddies associated with the coastal and shelf topography. The current flows along the semi-arid to arid narrow coastal plain on the west coast of southern Africa that is bounded to the east by the main continental escarpment. The most northern extent of the Benguela Current system reaches as far as 12-13°S (Moroshkin *et al.*, 1970) to the Angola-Benguela frontal region, and the most southern extent as far as the Agulhas retroflection area at approximately 40°S (Shannon *et al.*, 1981). The perennial intense wind stress at the Lüderitz upwelling cell (ca. 27°S) effectively divides the Benguela system into a northern and southern part both physically and biologically (Shannon, 1985; Pitcher *et al*., 1992). The system extends eastward to the eastern limit of the Agulhas Bank at 26°E (Shannon and Nelson, 1996; Lutjeharms *et al*., 2000). Between Lüderitz and the southern limit of the Benguela system the coastline orientation is roughly north-south, whereas from Cape Point (34°S) to the eastern limit of the Agulhas Bank the orientation is east-west. The coastal bathymetry is varied along its length and

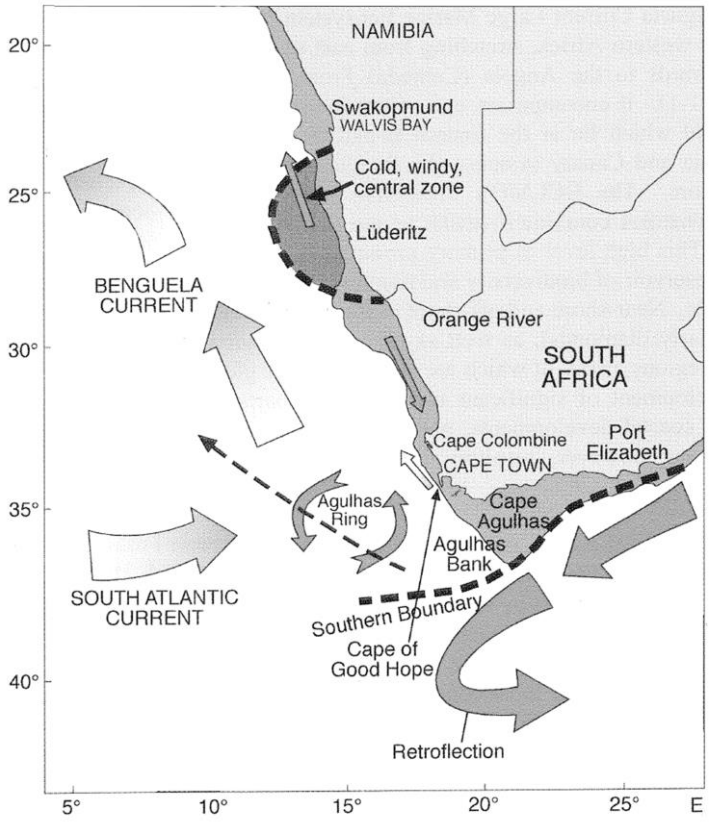


Figure 1.1 Illustration of the major upwelling features of the southern Benguela (adapted from Shannon, 1985)

breadth with several bays and capes. The continental shelf to the north of Cape Columbine (24°S) extends 120 km offshore whereas off the Cape Peninsula it is 40 km wide. The variability of the coastline creates alternating patterns of active and passive upwelling circulation. A prominent front is frequently observed near the shelf break where upwelled water sinks on the shoreward side. Beyond the shelf break front an eddy field marks the edge of the clear waters of the subtropical gyre of the South Atlantic. Cold core eddies appear as isolated areas of Chl *a* enhancement along this frontal zone (Shelton and Hutchings, 1990). The Agulhas Bank is a relatively wide and shallow shelf region (Fig. 1.1). It is a region of interaction between the cool Benguela Current to the west and the warm Agulhas Current to the

east (Probyn *et al*., 1994). The western Agulhas Bank is characterised by a stratified water column in summer and a well mixed water column in winter (Schumann, 1998). The eastern Agulhas Bank tends to have shallower more intense thermoclines as a result of the frequent intrusions of Agulhas Current water onto the bank (Walker, 1986).

1.2.1 Physical forcing and biological response in the upwelling system

Within the Benguela upwelling system the interaction of the South Atlantic Anticyclone with the adjacent continental pressure fields (Fig. 1.2) produce annual mean southerly winds along the west coast that force the upwelling of cold nutrient-rich South Atlantic Central Water from depths of 200-300 m (Hart and Currie, 1960; Shannon, 1966). Differences in seasonal wind along the length of the southern Benguela coastal region result in varied upwelling between the northern and southern regions (Shannon and Nelson, 1996; Strub *et al.,* 1998; Hardman-Mountford *et al*., 2003). The local wind is closely related to upwelling, particularly in the correspondence between peak cyclonic wind-stress curl (Shannon and Nelson, 1996) and major upwelling cells (Cape Frio (18°S), Lüderitz (27°S), Hondeklip Bay (31°S) and Cape Columbine (33 °S; Hardman-Mountford *et al.*, 2003). Elevated levels of Chl *a* biomass are noted downstream of these cells. Between these upwelling cells there tends to be a local minimum in mean wind stress and offshore Ekman transport during the winter leading to strong stabilization of the water column. The influence of cyclonic wind-stress curl is limited in its poleward extent to approximately 27°S in June-July but extends to Cape Agulhas from late spring to autumn (Bakun and Nelson, 1991). In the southern Benguela seasonal variation results as the South Atlantic Anticyclone drifts south and intensifies in summer in

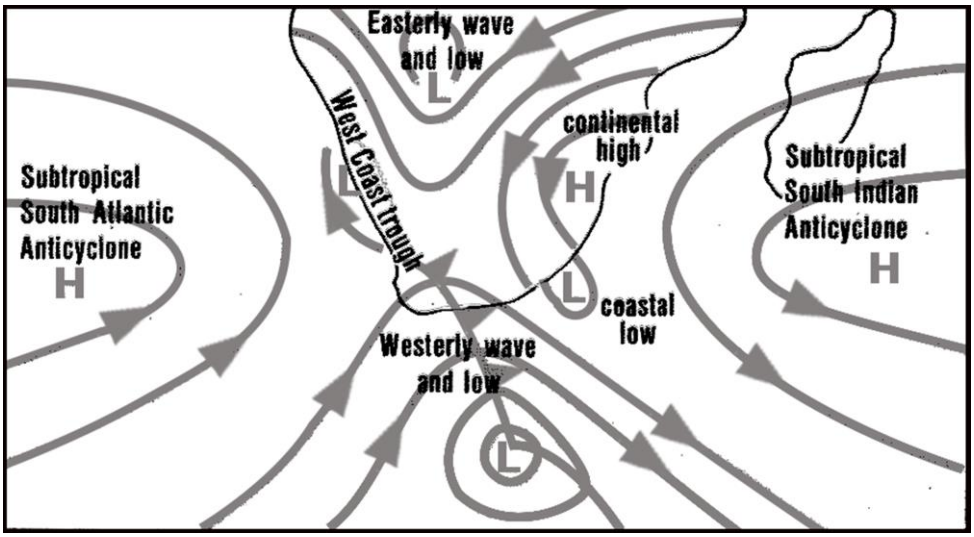


Figure 1.2 Illustration of the mesoscale atmospheric phenomena that drive local weather patterns (Preston-Whyte and Tyson, 1988).

conjunction with the development of a thermal low pressure over the central continent, affecting local winds. During winter the northward shifts results in a stronger role of the non-upwelling westerly winds (Andrews and Hutchings, 1980). North of approximately 31°S there is perennial upwelling, although a slight seasonality in intensity does exist.

Over the Agulhas Bank region easterly winds, frequent in spring and summer tend to move inshore water southwards causing upwelling. Winter is dominated by strong westerly wind but there is considerable switching of wind within a few days (Schumann, 1998). Upwelling can extend along the south coast east of Cape Agulhas (35°S, 20°E) (Shannon and Nelson, 1996; Lutjeharms *et al.*, 2000).

Short-term variability

The biological response to the environmental forcing differs significantly between the northern and southern Benguela regions. The predominant reason suggested by

Chapman and Shannon (1985) is the short-pulsed forcing in the south relative to the more continuous forcing in the north. The mean seasonal features of the troposphere seldom reflect the short-term perturbations from the mean in the form of short baroclinic waves and low pressure systems that control daily weather. These atmospheric low pressure systems and waves are the main agents of horizontal momentum and are superimposed on the semi-stationary barotropic westerly wave. On any one day there may be four to eight of these high frequency perturbations in the westerlies that exhibit a range of periods with peaks in the wave spectrum of two to eight days (Preston-Whyte and Tyson, 1988). Around South Africa these travelling wave disturbances create near-surface cyclonic vorticity and instability ahead of them, which manifest in the form of cold fronts (Preston-Whyte and Tyson, 1988). Cold fronts tend to follow well defined storm tracks in the jet stream with a seasonal migration, being furthest north in winter when the upper level westerly belt is at its widest. They are associated with distinctive patterns of surface divergence ahead and convergence behind them and a change in wind direction from southerly to north-west to south-west (Preston-Whyte and Tyson, 1988).

Associated with the approach of these perturbations are cells of low pressure or coastal lows and ridging anticyclonic cells which sometimes occur behind the fronts when a high pressure cell breaks off from the South Atlantic Anticyclone and drifts eastward (Fig. 1.3). The ridging anticyclone intensifies the offshore pressure gradient that leads to strong low-level south-easterly winds (Preston-Whyte and Tyson, 1988). Coastal lows are usually initiated on the west coast due to gradient flow from the high inland plateau towards the coast and propagate southward and then eastward along the coast as internal Kelvin waves trapped vertically by a low

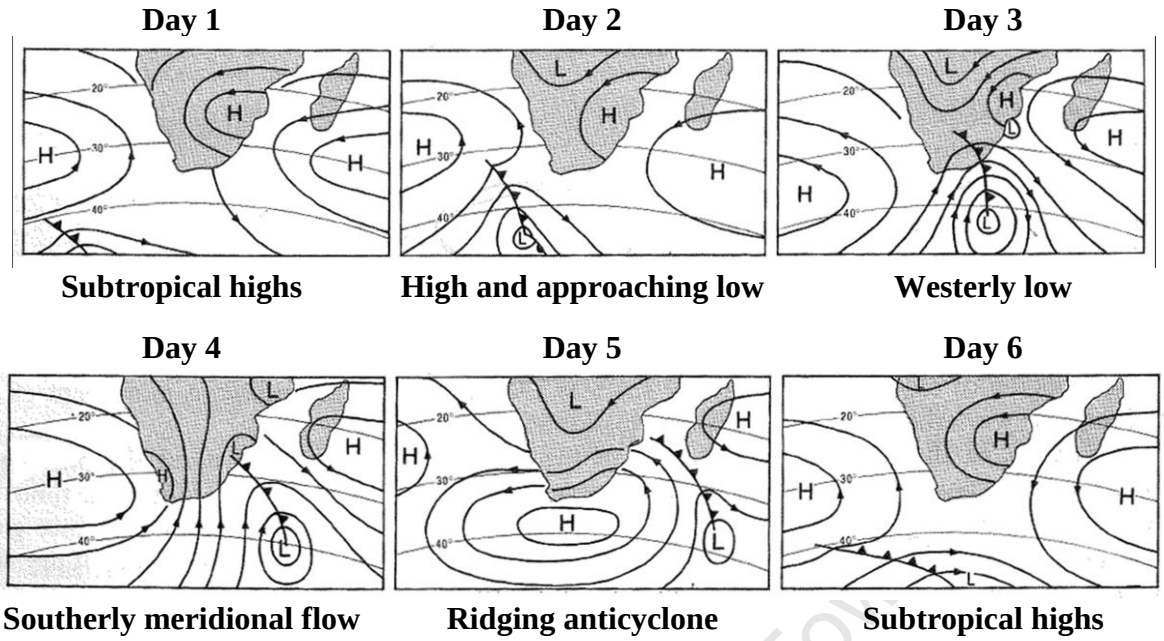


Figure 1.3 An illustration of a typical sequence of surface weather disturbances over southern Africa (Preston-Whyte and Tyson, 1988).

level inversion and horizontally by the escarpment. Offshore airflow occurs ahead of these coastal lows and onshore flow behind them (Preston-Whyte and Tyson, 1988).

The passing of atmospheric frontal systems in the south weaken the South Atlantic Anticyclone for short periods, causing slackening or abatement of the southerly winds. The pulsing of upwelling favourable winds allows conditions to continuously alternate between nutrient replenishment (turbulent events) and phytoplankton growth (water column stabilization). Further pulsing may occur in association with coastal low pressure systems, which can cause variations in sea-level that continue along the south coast in concert with the coastal low (Mann and Lazier, 1996; Nelson, 1992; Jury and Brundrit, 1992). At the crests of the waves the deep, nutrient-rich water is raised nearer than normal to the surface, which leads to more productive upwelling. At the wave troughs the opposite occurs as the nutrient-rich water is pushed deeper than normal.

In addition to the physical forcing, local shelf bathymetry plays a significant role in determining the locations of surface features (Longhurst, 1998). Nelson and Hutchings (1983) suggest that Ekman upwelling occurs preferentially where the continental shelf is narrowest such as off salient capes. Upwelling can also occur where the Agulhas Current flows against the shelf edge (Bakun, 1993; Lutjeharms *et al*., 2000). Wind fields on a smaller scale to mid-latitude perturbations are also influenced by local orography. Jury (1985) discusses how the longshore winds over the Cape Peninsula and Cape Columbine are affected by the orography. The curvature of the south-easterly wind in the lee of orographic features is variable depending on the depth of the airflow. Hutchings and Taunton-Clark (1990) compared the wind from Cape Point and Cape Columbine and concluded the poor spatial coherence between the two proximate sites can be attributed partly to orography.

Subsurface structures

Changes in surface wind along the west coast have been shown to rapidly influence changes in thermal structure and hence Chl *a* profiles (Hutchings *et al*., 1984; Brown, 1986; Brown and Hutchings, 1987). This creates a highly variable system with regards to profile shape particularly inshore. The effect of the wind may be modified by the structure of the water column; a weak temperature gradient being more susceptible to wind erosion that a highly stratified one (Brown and Hutchings, 1987). Low wind speeds (mean of 7.5 $m.s^{-1}$) result in a stable water column with a shallow upper mixing layer and rapid phytoplankton growth, whereas strong wind speeds (mean speed for one to three days of 13.9 $m.s^{-1}$) result in a much deeper surface mixing layer (mean depth 32 m). During a study conducted in the summer of

1985-1986, Largier and Swart (1987) noted that winds of the order 20-25 $m.s^{-1}$ were required to break down the thermocline that developed over the central Agulhas Bank. Further west the authors suggested that stability of the water column would remain unaffected by wind speed below ca. 26 $m.s^{-1}$. Wind direction can also influence water column stability. Short periods of onshore winds may create a more stable structure by advecting warmer surface water shoreward. Longer periods are more likely to cause significant mixing. Hutchings *et al*. (1984) found that south-westerly winds off the Cape Peninsula during January 1983 were not as efficient in forcing strong upwelling as the south-easterly winds that followed. Over the Agulhas Bank offshore easterly wind stress, which predominates over westerly wind in summer, drives coastal upwelling (Jury, 1988; Lutjeharms and Stockton, 1991). A well developed upwelling front is established with the onset of south-easterly winds in September (Probyn *et al*., 1994). In winter, westerly winds dominate, eroding the thermocline and resulting in a uniformly well-mixed upper layer (Boyd and Shillington, 1994). In reality, however, the weather systems bring alternately easterly then westerly wind within a few days (Schumann, 1989).

The thermal structure is an important determinant of the Chl *a* distribution. Stable conditions are noted by stratified waters and often a mixing layer of ca. 10 m when wind speed is low, which can result in prolific phytoplankton growth if conditions remain stable (Andrews and Hutchings, 1980; Brown and Hutchings, 1987). Long-term stability can result in a distinct subsurface Chl *a* maxima at 20-30 m, presumably due to sinking phytoplankton (Mitchell-Innes and Walker, 1991). A deep mixing layer below the thermocline can restrict the growth of phytoplankton (Brown and Hutchings, 1987).

Mitchell-Innes *et al.* (2001) described the characteristic winter profiles of four regions within the Benguela system using a shifted-Gaussian curve. The profiles generally changed from inshore well-defined surface or near-surface peaks to progressively weaker and deeper peaks offshore. The height of the peak declined offshore as well as getting deeper whereas the spread of the peak increased. The authors found that temperature profiles had little thermal structure with mixing layer depths that ranged from 13-41 m over the shelf and 42-60 m over the slope. Mean concentrations of Chl *a* were > 5 $mg.m^{-3}$ in the upper 30 m with peaks of 5-11 $mg.m^{-3}$. Chl *a* peaks were in the upper 17 m over the shelf and between 18 and 26 m over the slope. On occasion the concentrations may be more uniform across the shelf (Hutchings *et al*., 1984).

A number of papers have investigated the thermal and biological structure of the Agulhas Bank (for example, Largier and Swart, 1987; Largier *et al*., 1992; Probyn *et al*., 1994; Mitchell-Innes *et al*., 1999). Over the Agulhas Bank seasonal winds modulate the depth of the thermocline, which results from the advection of cold bottom water onto the shelf, particularly towards the west (Largier and Swart, 1987). Strong winds deepen the thermocline and may suppress productivity (Carter *et al*., 1987). A two-layer structure exists over the larger shelf area where cold upwelled water, forced by the Agulhas Current on the eastern bank, forms a basal layer and intrusive plumes of warmer Agulhas Current water continually replenish a mixed upper layer. During winter-spring there is a predominantly well mixed water column that is low in nutrients, which becomes more stratified over summer-autumn due to insolation and an increase in the relatively nutrient-rich upwelled bottom water. The intensive thermocline that results at the interface of the two layers has a tendency to deepen and weaken towards the west. Over the eastern Agulhas Bank shelf-edge

there is a doming of the isotherms as bottom water upwells. A cool ridge of surface water may occur inshore and parallel to the shelf edge above the 100 m isobath (Largier and Swart, 1987). The ridge has been proposed as an interaction between the advection of coastal upwelling and the Agulhas Current (Boyd and Shillington, 1994). Westward of this the bottom water tends to thin out and slow down over the broader western Agulhas Bank. As such the western region is more susceptible to thermocline break-down (Probyn *et al*., 1994). The shallow thermocline inshore is particularly susceptible to break-down from surface wind stress and large amplitude internal tides particularly on the western side in summer where upwelling winds are stronger and more frequent (Largier and Swart, 1987; Boyd and Shillington, 1994).

Probyn *et al*. (1994) reviewed primary production on the Agulhas Bank and describe the Central and East Agulhas Bank to be roughly east of 21°E (a line running from Cape Agulhas (20°E) to the southern tip of the bank) as opposed to a West Agulhas Bank. Typically, strong summer upwelling inshore on the West Agulhas Bank supports phytoplankton populations that reach surface Chl *a* concentrations of 2-5 $mg.m^{-3}$ and subsurface peaks of ca. 10 $mg.m^{-3}$ at 10-30 m. Eastward along the coast, peak values are typically located at 10-20 m. Further offshore over the bank a spring bloom occurs as a result of a more stable water column but declines by mid-summer as nutrients are depleted. On the West Agulhas Bank shelf-edge, Chl *a* peaks occur near the thermocline at 30-50 m with values ca. 1 $mg.m^{-3}$, whereas on the East Agulhas Bank shelf-edge Chl *a* peaks of 1-5 $mg.m^{-3}$ are found between the surface and 20 m due to the shallower thermocline. At the cool upwelling ridge inshore of the eastern shelf edge maximum concentrations > 2 $mg.m^{-3}$ are found between 10 and 20 m.

Characteristic temperature and Chl a features

Changes in sea surface temperature (SST) have been linked to the abrupt changes in wind stress (Brown and Hutchings, 1987). Along the west coast, Brown and Hutchings (1987) generally considered surface temperature below 11.5°C to indicate actively upwelling water in summer. Andrews and Hutchings (1980) classified newly upwelled water as < 10°C; maturing upwelled water as 10-12°C; and aged upwelled water as 12-16°C. Similarly, Barlow (1992) used the intervals 8-10°C to describe newly upwelled water, 10-15°C for mature upwelled water and 12-16°C for aged upwelled water. As the newly upwelled water begins to warm and stabilize, diatoms are able to exploit the rich nutrient conditions. Mitchell-Innes and Pitcher (1992) proposed that there is a temperature window from 12-15°C when conditions are good for diatom population development and beyond these limits Chl *a* concentrations decline. The decline in concentration is part of the succession to flagellates which are able to survive on the low concentration of recycled nutrients (Mitchell-Innes *et al*., 1991).

Over the West Agulhas Bank deep mixing in winter maintains surface water temperature at 13-16°C (although cooler water has been observed inshore) whereas in summer upwelled inshore water is ca. 11-12°C (Mitchell-Innes *et al*., 1999). Water temperature > 18°C is considered to result from intrusions of Agulhas Current subtropical water (Shannon, 1985; Boyd *et al*., 1985; Largier, 1992) and is low in Chl *a* (< 0.1 $mg.m^{-3}$, Mitchell-Innes *et al*., 1999).

Chl *a* concentration in newly upwelled water is generally < 1 $mg.m^{-3}$, peaks at between 10 and 20 $mg.m^{-3}$ and then decreases to between 1 and 3 $mg.m^{-3}$ when recycling of nutrients probably maintains low levels (Brown and Hutchings, 1987).

Mitchell-Innes and Pitcher (1992) observed concentrations > 3 $mg.m^{-3}$ correlate with a favourable temperature window in which diatoms proliferate, and concentrations < 3 $mg.m^{-3}$ indicate either newly upwelled water or mature water that has shifted towards flagellate dominance. Barlow (1982) proposed that maturing upwelled water has characteristic values between 1 and 20 $mg.m^{-3}$ and aged upwelled water ranges from 5-30 $mg.m^{-3}$. Comparative concentrations on the West Agulhas Bank were only found inshore with values of 2-5 $mg.m^{-3}$ throughout the coastal area in summer (Probyn *et al*., 1994) although concentrations can reach > 10 $mg.m^{-3}$ in subsurface layers (Mitchell-Innes *et al*., 1999). Highest concentrations over the mid-shelf and outer bank of the eastern region were 1-5 $mg.m^{-3}$ and frequently subsurface however, the inner reaches of the Central and East Bank showed extreme variability where stratification can result in concentrations of up to 40 $mg.m^{-3}$ (Probyn *et al*., 1994). Offshore waters generally had concentrations < 0.5 $mg.m^{-3}$.

1.2.2 Climate change impacts on eastern boundary upwelling systems

Long-term changes in global primary production have suggested decreases in primary production in most of the oceans but increases in primary production in coastal upwelling areas (Kahru and Mitchell, 2008). Demarcq (2009) showed a non-significant relationship between trends in biomass and SST in eastern boundary upwelling systems (EBUS) and a significant relationship between biomass and upwelling favourable winds. Of particular interest then, is how these winds affect primary production processes such as the vertical temperature structure and the vertical distribtution of Chl *a*. However, few current primay production models incorporate wind information. Instead, many assume a constant mixed layer depth assumed to be representative of the wind forcing. By considering wind, particularly

consecutive periods of wind, and its affect on subsurface structures and subsequently primary production, it may be possible to provide a clearer indication of how wind affects the upper ocean. For example, increased turbulence can have both positive and negative effects; increased nutrient supply to the upper layer and exsposing cells to longer periods lower light conditions may act to both enhance and suppress net primary production. In this way, changes in the wind forcing patterns can favour certain phytoplankton species over others. Thus incorporating wind data is a primary focus area of this thesis.

1.3 Relating the vertical distribution of Chl *a* to environmental variables

Vertical profiles of Chl *a* have been studied and frequently explained in terms of a four-parameter Gaussian curve (Lewis *et al*., 1983; Platt *et al.,* 1988; Sathyendranath *et al*., 1995). Taguchi *et al.* (1994) modelled the nonlinear relationships between the Gaussian curve parameters and environmental conditions using *generalized linear models* (GLMs). Their model was successfully able to predict three parameters of the Gaussian curve; background Chl *a* concentration, depth of the subsurface maximum, and the total biomass above the background concentration. Richardson *et al*. (2003) applied a similar approach to the southern Benguela system but were only able to successfully predict the depth of the subsurface maximum and the total biomass above the background concentration. Silulwane *et al.* (2001) and Richardson *et al.* (2002) used *self organizing maps* (SOMs), a type of artificial neural network, to determine characteristic Chl *a* profile classes from a database of Gaussian curve parameters. The database was created by fitting the curve to an archive of *in situ* profiles from the Benguela region. The profile class could then be predicted from their relationship between the Gaussian

curve parameters and a suite of easily measurable environmental variables including satellite data (SST and surface Chl *a*). Their approach however, is only semi-quantitative as it does not indicate the strength of the relationships and it is not a continuous parameterization of profiles (Richardson *et al*., 2002). Further, this method requires that profiles in the raw Chl *a* data that do not fit a Gaussian model be removed prior to analysis (e.g. ca. 15% in Silulwane *et al.*, 2001). Richardson *et al.* (2003) took the process of dynamic profiles further for the region by relating the four parameters of a shifted Gaussian model to real-time satellite-derived information (SST and surface Chl *a* concentration) and known inputs such as season, location and depth of the water column. The authors were able to predict the depth of the Chl *a* maximum and the height of the peak above the background Chl *a* but were not successful in determining significant relationships for the background Chl *a* and the width of the peak. Demarcq *et al.* (2008) used a series of 15 SOM-derived profile categories based on raw Chl *a* profile data and a simple light model to estimate the primary production in the Benguela and Agulhas systems. The authors used satellite-derived monthly SST and surface Chl *a* concentration and known variables (season, location and depth of the water column) to predict 1 of 15 characteristic profiles. This was performed using a two-step generalized modelling approach. The first step involved ascertaining the relationship between the surface observations and each of the 15 profiles using *generalized additive models* (GAMs) and the predictor variables of surface SST, Chl *a*, bottom depth, season and location. Location was categorized as either West Coast, West Agulhas Bank or East Agulhas Bank. Bottom depth was used as a proxy for distance from the shore. The second step involved assessing the relationship between profile number and each environmental predictor by inspecting the GAM plots and then parameterizing them

using piecewise linear regression or exponential transformation. These parameterizations were implemented as a GLM to predict the profile type from the satellite data.

Two important influences on primary production were omitted from the Demarcq *et al*. (2008) study; the influence of wind, and the evolution of processes over time. Wind plays an important role in the definition of many of the biomes presented by Longhurst *et al.* (1995) and is the primary forcing of upwelling in the Benguela system (Shannon, 1985; Pitcher *et al.*, 1992). There are many studies on the effects of synoptic wind stress and wind stress curl on primary production (for example, Blanke *et al*., 2002; Botsford *et al*., 2003; Botsford *et al*., 2006; Largier *et al*., 2006; Rykaczewski and Checkley, 2008; Schwarz *et al*., 2010; Albert *et al*., 2010). This is particularly important in the Benguela upwelling system where bloom development and decay occurs in concert with the atmospheric disturbances. Although highly variable there are recurring patterns in the passage of these disturbances.

1.4 **Statistical methodology to recognize patterns**

This thesis hypothesizes that these patterns in the daily wind field that relate to low level atmospheric processes can be used in conjunction with patterns of satellite-derived SST and Chl *a* to predict the vertical thermal structure of the water column and the vertical distribution of phytoplankton. Methods from the field of *pattern recognition* and *machine learning* will be applied to derive relationships among the variables based on regularities in the data. These patterns can be used to classify new data in a structured knowledge representation. Pattern recognition is an automatic approach to knowledge discovery by means of computer algorithms. A machine is built comprising these algorithms, which is able to tune the parameters of

an adaptive model from a set of training data. The training set is usually a vector of observed variables with a known outcome or response, which represents supervised training for a classification problem. The algorithm used can be represented as a function $y(x)$ where x is the vector of observed variables and y is the output label. The generalization or ability of the supervised model to correctly label new data can be tested on data that has been excluded from the training set. A fundamental concept used within machine learning is how it deals with the uncertainty that arises through noisy and sparse input data. Probability theory is a sound mathematical framework for the quantification of uncertainty (Bishop, 2006). There are two rules of probability (p) that form the basis of probabilistic induction: the *sum rule* (Eq. 1.1); and the *product rule* (Eq. 1.2).

$$p(X) = \sum_{Y} p(X, Y) \qquad (1.1) \quad .$$

$$p(X, Y) = p(Y|X)p(X) \qquad (1.2) \quad .$$

Here $p(X, Y)$ is the joint probability of X and Y, $p(Y \mid X)$ is the conditional probability of Y given that X has already occurred, and $p(X)$ is simply the probability of X or the marginal probability of X if Y has been marginalized or summed out of the equation as indicated in Equation 1.1. From the product rule and the property that $p(Y, X) = p(X, Y)$ the following conditional relationship can be derived (Bishop, 2006);

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad (1.3) \quad .$$

The denominator is the normalization constant and is calculated using the sum rule. It ensures that summing the conditional probabilities of X over all possible values or Y (Eq. 1.4) equals 1.

$$p(X) = \sum_{Y} p(X|Y)p(Y) \qquad (1.4) \quad .$$

Bayes' theorem (Eq. 1.3) is central to pattern recognition and machine learning. It allows decision or prediction models to make optimal selections based on all the given information, even though the information may be incomplete, complex or ambiguous. A Bayesian approach to determining probabilities differs to that of classical or *frequentist* approaches. The classical interpretation of probability is based on the frequencies of random repeated events. Any event that is not included in the training data will have a probability of zero. A Bayesian view of probability is to update an existing or prior probability. For example, if a coin is flipped three times and comes up heads each time, a classical estimation of the probability would be p(heads) = 1, whereas a Bayesian estimation will update a prior, for example, p(heads) = 0.5, after each toss of the coin.

1.5 **Thesis structure**

The overall aim of the thesis is to produce fine temporal and spatial resolution primary production estimates in three-dimensional space based on daily satellite surface measurements. It is hypothesized that complex relationships that evolve over time can be discovered automatically from an archive of data for these estimates. This thesis is broken up into several steps, chapter by chapter.

For practical purposes the input variables and classification labels are pre-processed into a new variable space to simplify the pattern recognition problem by improving computational speed and reducing the dimensions of the data. It is important to select methods that simplify the data, while maintaining as much useful information as possible. This is the focus of Chapter 2 in which the available data are first analysed in terms of its spatial and temporal variability. The data are then pre-processed in accordance with the variability to reduce its dimensionality.

Chapter 3 aims to test the hypothesis that subsurface phenomena such as temperature and Chl *a* profiles can be predicted in a probabilistic framework from satellite-derived surface data coupled with season and location. First a network is developed to predict surface Chl *a* values using the environmental data. This introduces the Bayesian network but also explores the potential to predict missing values often encountered in satellite SST and Chl *a* data. Initially the structure of the model is developed automatically, i.e. model building algorithms are applied that search for the network that maximizes the probability of the observed data. A second network is developed to predict the most likely profile given the environmental data. This structure is developed using an intuitive understanding of the variable relationships.

Chapter 4 takes this further by using time sequences of surface data of up to six days, testing the hypothesis that the relationships between physical atmosphere-ocean processes over a period of days and their influence on phytoplankton distribution can be generated automatically from data archives. To cope with the large amount of data, the study is narrowed down to the St Helena Bay region of the southern Benguela ecosystem. The method is evaluated by reconstructing observed or *in situ* data that is excluded from the model training data.

In Chapter 5 a primary production model is developed using the vertical distribution of biomass to estimate the subsurface light field and primary production at depth. The model is tested by comparing the model predictions to *in situ* estimates of these variables.

In Chapter 6 the models developed in Chapters 4 and 5 are combined by applying them as a complete model to the St Helena Bay region. The model takes the daily satellite data and various other information as input, fills in the missing data in the SST and Chl *a* sequences, predicts the most likely profile from these sequences and estimates production at each location. The model produces a time-series of daily depth-resolved primary production. The model development, application and results are evaluated by comparing individual model component outputs to previous findings in published literature and other available *in situ* data.

Chapter 7 concludes the thesis by discussing the novel contributions made in previous chapters in terms of a complete synthesis of information on the fundamental processes involved in a complex, dynamic and highly productive coastal ecosystem.

Chapter 2 - Data Acquisition, Pre-processing and Clustering

2.1 Introduction

As a parcel of water is advected from an active inshore coastal upwelling site downstream and across the shelf, it undergoes physical changes that impact both physically and physiologically on phytoplankton (Pitcher, 1988; Brink *et al*., 1995). Typically, coastal upwelling occurs close to the coast at shallow depths where upwelled water is cold (< 12^{o}C) and low in chlorophyll *a* (Chl *a*; < 1 mg.m^{-3}). As the surfacing water moves offshore, increased insolation causes stratification and allows considerable phytoplankton growth (Hutchings *et al*., 1984; Brown and Hutchings, 1987). These near-surface blooms are usually dominated by diatoms (Mitchell-Innes and Walker, 1991; Barlow *et al*., 2001). If conditions remain favourable for growth there will be a succession of phytoplankton from diatom to dinoflagellates and nanoflagellate dominance (Pitcher *et al*., 1996). Typically the nanoflagellates will manifest as a subsurface maximum as the parcel of water reaches the thermal front in the region of the shelf break (Barlow *et al*., 2001). At the thermal front the upwelled surface layer sinks below the warmer oceanic surface water, which is low in Chl *a* (Shannon and Nelson, 1996).

These well-studied biological processes are associated with various surface phenomena that can be recorded by remote sensors and other easily obtainable information (Pitcher *et al.,* 2006). With information on multiple variables that represent different aspects of these processes more precise depictions should be attainable. For example, if the sea surface temperature (SST) is 15^{o}C it can be

inferred that the parcel of water is probably somewhere on the shelf but it may represent for instance, typical winter surface temperatures or a spring time mixing layer temperature. If the bottom depth at the sample location is < 200 m and it is from the West Coast in summer, it can be inferred that the sample more likely represents maturing upwelled water that is likely to have high Chl *a* concentration. Further information on the concurrent wind will help explain how the Chl *a* is distributed in the water column, for example, a homogenous layer within a mixing layer or as a surface peak.

Such useful information may be continuous (e.g. SST) or categorical (e.g. month) and may have many dimensions. In order to utilize this information efficiently it is often necessary to simplify it. Pre-processing of data aims to reduce the dimensions of the data while retaining useful structures therein. This is the focus of this chapter. The variable data introduced above are analysed according to their spatial and temporal distributions with the purpose of discovering useful generalizations.

2.2 Data and methods

2.2.1 *In Situ* data

Region

The southern Benguela has been characterised according to shelf dynamics to include the West Coast, West Agulhas Bank and East Agulhas Bank. The West Coast is characterised as a typical eastern boundary shelf system with strong intermittent upwelling favourable winds producing nearshore upwelling (Shannon, 1985). Over the Agulhas Bank the east-west dichotomy is clear from both hydrological and biological characteristics (Probyn *et al.*, 1994). The East Agulhas

Bank has been described as a western boundary shelf system with an advectively controlled thermocline (Largier and Swart, 1987). Shelf-edge upwelling results from interactions with the Agulhas Current. Towards the western Agulhas Bank, wind mixing plays a more important role in the forcing and characteristics of both an eastern and western boundary system are observed. The West Agulhas Bank has similar wind-driven coastal upwelling typical of eastern boundary current systems but also has a seasonally stable two-layer structure typical of temperature shelf systems that is not as variable and susceptible to episodic events (Largier and Swart, 1987). Probyn *et al*. (1994) reviewed primary production on the Agulhas Bank and suggested a line running south from ca. 21°E can be used to divide the Agulhas Bank according to biophysical characteristics. This reference is used here to divide the southern region of the southern Benguela into the West Agulhas Bank (WAB) south of 34°S and west of 21°E, and the East Agulhas Bank (EAB) between 21°E and 25°E.

In contrast to the relative homogeneity of processes within the two distinct Agulhas Bank regions, the West Coast is comprises distinct regions in terms of coastal bathymetry, currents, wind stress and centres of upwelling (see Section 1.3). The West Coast therefore, can be subdivided according to these distinct characteristics. According to Shannon (1985) the West Coast (limited by the extent of the southern Benguela to the north and the Agulhas Current retroflection to the south) is subdivided into the Cape Peninsula upwelling area, the Cape Columbine or St Helena Bay upwelling area and the Namaqua upwelling area (Fig. 2.1). However, because the West Coast is distinct from the Agulhas Bank regions in terms of physical forcing of the upwelling system, the data for the sub-regions within the West Coast are pooled for analysis in this chapter.

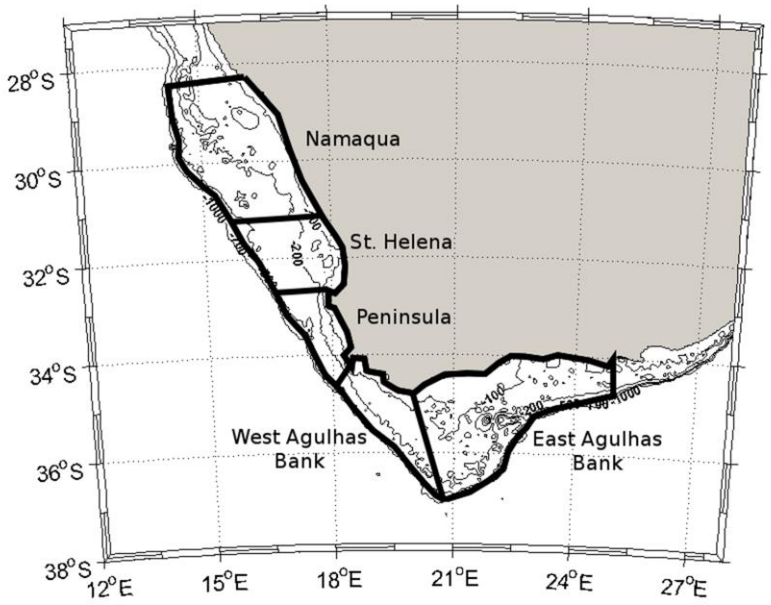


Figure 2.1 Sub-regions of the southern Benguela upwelling system. The West Coast is comprises the Namaqua, St. Helena and Peninsula sub-regions (Shannon, 1985). Bathymetry data are from ETOPO2v2 Global Gridded 2-minute database and are shown at 100 m bottom depth intervals.

Profiles

The Department of Environmental Affairs (DEA) has provided a record of over 6500 vertical profiles of phytoplankton fluorescence and temperature obtained from a thermistor and profiling fluorometer on a Seabird SBE CTD deployed from ships. These records date from 1988-2011 and cover the southern Benguela region.

The profile data consist of three batches based on their processing. The first batch, which spans 1988-2001 and the second batch from 2001-2008 were recorded with a Chelsea Instruments AquaTracka MK III. For the first batch a calibration equation was obtained by regressing the recorded voltages from the fluorescence sensor against discrete depth samples beginning a few metres below the surface and at

intervals of 10 m down to between 30 and 50 m depending on the profiling signature. Samples were filtered onto Whatman GF/F filters, which were placed in 90% acetone for 24 hours to extract the pigments. Chl *a* was measured fluorometrically using a Turner Designs Model 10-000R fluorometer.

The second batch was measured by reverse-phase high performance liquid chromatography (HPLC) method described by Barlow *et al*. (1997). A total of 402 HPLC-determined concentrations were regressed against their corresponding fluorescence voltages to obtain a calibration equation. The regression showed a good fit to the data (r^2 = 0.82). The calibration of the fluorometer readings was applied after they were binned into 1 m intervals. The third batch spans 2009-2011 and was obtained from a WETLabs fluorometer. A standard fluorometer calibration from WETLabs software was used to obtain Chl *a* concentrations. All the resulting Chl *a* and temperature profiles were smoothed using a five point moving average to reduce the noise.

The profile data consist of a depth vector and corresponding temperature or Chl *a* values that are not classified according to any state of the upper water column with respect to thermal structure or biological distribution. It is desirable that the number of profiles in the data set is reduced to a smaller set of natural "common" profile classes that represent the range and variability of the profiles. However, there is no clear indication of how many classes this should be. It is too simplistic to assume that the temperature and Chl *a* profiles follow along a regular succession of development from active upwelling to stabilization and growth followed by stratification and decline. The state of the system will depend on a number of interacting non-linear factors for example, the state of the system as upwelling

favourable winds begin and the strength, direction and duration of the wind. It is unlikely therefore, that the system will follow a predictable continuum. In addition, advective processes can introduce new water masses that may interrupt the growth or decline of the phytoplankton bloom (De Villiers, 1998). It is prudent therefore, to assume that biological processes may be in any state at a given time and that the profiles should be able to reflect the diversity of states rather than the continuum. Each profile class should then represent a collection of similar snap-shots of the system in a similar way to the collection of the samples.

A simple method for reducing the dimensions of the profiles is to classify or cluster them using an unsupervised (the true class is not known) clustering method. A *hard* clustering method maps each profile to only one cluster whereas *soft* clustering methods map each profile to one or more clusters with a probability of membership to that cluster. Both approaches usually require the number of clusters to be pre-specified by the user. There are algorithms available to select the optimal number of clusters by relating the within-cluster variability to the between-cluster variability. Developing algorithms for determining the optimal number of clusters is an ongoing focus of research in the field of pattern recognition. No one single method can always outperform the others (Kryszczuk and Hurley, 2010; Qin, 2012). For simplicity a hard clustering approach was chosen with a pre-specified number of clusters.

From published work (Brown and Hutchings, 1987; Mitchell-Innes *et al*., 2001; Silulwane *et al*., 2001) it can be intuitively deduced that three clusters will not adequately capture the variability of the system, and 25 clusters may not adequately differentiate between similar cluster as well as wasting computational effort in the models that discover relationships with the clusters. A popular and well established

method is the *k*-means clustering algorithm (MacQueen, 1967; Lloyd, 1982) which partitions the data into k clusters that represent the distribution of the input data optimally. The process involves describing a random set of k cluster centres μ_k (often a random subset of the input data), which represent a prototype with the same dimensionality as the input data $x_n \in \{x_1.,\ldots,x_N\}$, where x is a profile sample and N is the number of samples, and then optimizing the cluster centres using a simple objective function;

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2 \qquad (2.1)$$

where $r_{nk} \in \{0,1\}$ is a binary code indicating if the sample x_n belongs to the cluster k, and J is the objective function to be minimized. More specifically it is an iterative procedure involving two steps performed on the data set at each iteration. First, the sample x_n is compared to each μ_k and the cluster with the minimum sum of squares of Euclidean distances is chosen as the winning cluster (indicated by the vector r). Second, the winning μ_k is updated to minimize the sum of squares distance. At the end of the process μ_k is calculated as the mean of all the data point assigned to cluster k, hence *k*-means clustering. The process continues until convergence of J. Members of a cluster can be thought of as having short inter-point distances compared to distances to members of other clusters. One problem with this method is that J can converge to a local rather than a global minimum. Several iterations are therefore required. Before processing the temperature profiles, data from the upper 2 m and profiles with missing data above 50 m are removed as the algorithm ignores

profiles with missing values. The remaining temperature profiles are limited to the upper 50 m depth for clustering. Similarly the upper 2 m data are removed and a 40 m limit was set for the Chl *a* profiles. The shallower limit for the Chl a profiles allows more profiles from closer inshore to be used in training, which is considered unnecessary for the temperature profiles. These limits are believed to be sufficient to capture the characteristic physical and biological structures of the water column. Although there is some evidence of deeper Chl *a* maxima, the production is negligible due to low light (Carter *et al*., 1987).

Depth

The position of an upwelled water parcel relative to its origin provides useful biological information that can be approximated by the bottom depth of the water column. Bottom depth information was obtained from the ETOPO2v2 Global Gridded 2-minute Database, National Geophysical Data Centre, National Oceanic and Atmospheric Administration, U.S. Dept. of Commerce, http://www.ngdc.noaa.gov/mgg/global/etopo2.html. The data are based on satellite altimetry and are more appropriate for deep-water than close inshore. However, due to the dynamic nature of the upper water column the bathymetric detail is unlikely to be strongly related to phytoplankton distribution and thus the bathymetry is only used as an indication of distance offshore.

To investigate how SST, surface Chl *a* and wind data vary with bottom depth, samples of the variables are obtained by randomly selecting up to 50 values from each of the three main sub-regions consisting of the West Coast, West Agulhas Bank and East Agulhas Bank. Samples are taken from each day from 2002-2009 and between 0 m and 1000 m bottom depth, with each sample being an average of

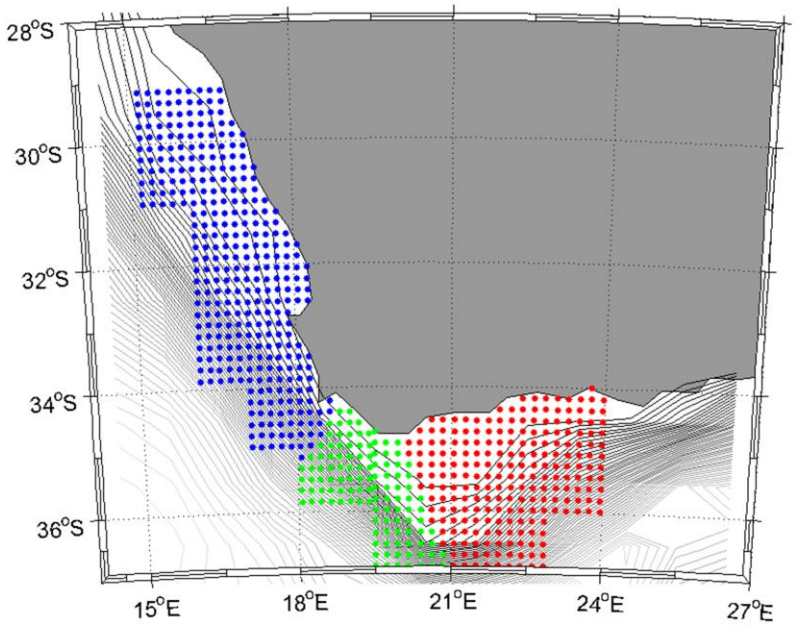


Figure 2.2 The 8 x 8 km pixel-averaged sampling positions used for evaluating variations in satellite derived SST, Chl *a* and wind data in the three sub-regions; West Coast (blue), West Agulhas Bank (green) and East Agulhas Bank (red).

an 8 x 8 km block to smooth the data and dampen the effects of outlier data (Fig. 2.2). The variability of the data is explored according to each bottom depth interval.

2.2.2 Satellite data

SST and Chl *a*

Temperature is an important factor in phytoplankton physiology and hence productivity (Eppley, 1972). In addition, the SST visible to satellites is useful for indicating upwelling, maturing upwelled water, oceanic water and other physical phenomena that affect the distribution of phytoplankton. SST has also been correlated to nutrient concentration (Waldron and Probyn, 1992; Silio-Calzada *et al*., 2008) and used to successfully predict other environmental parameters (Platt *et al*., 1995).

In studies of phytoplankton, surface Chl *a* concentration is used as a proxy for phytoplankton density. Chl *a* pigment absorbs wavelengths towards the blue end of the visible light spectrum and since the longer wavelengths are rapidly absorbed in seawater, the presence of phytoplankton tends to change the colour of seawater from blue to green. Chl *a* is measured by satellites typically as a ratio of green-blue wavelengths (although operational algorithms are more complex). However, other substances in the water column such as coloured dissolved organic matter (CDOM) also absorb in the blue spectrum and contribute to a greener colour.

SST and Chl *a* data were obtained from the moderate-resolution imaging spectrometer (MODIS) onboard the Aqua satellite (NASA), which provides global data from 2002 to the present. MODIS is the only instrument that provides coincident SST and Chl *a* maps derived from the same instrument. Aqua is in a near-polar sun-synchronous orbit and has a swath width of 2,330 km which enables it to view the entire surface of Earth every one-two days acquiring data at 250 m, 500 m and 1 km resolution. The data files were obtained from MODIS level 2 data available from the NASA Oceancolor web portal and processed using the standard SeaDAS version 6.4 software. The seasonal climatology (for the years 2002-2009) of the data is illustrated in Figure 2.3 by the months January (summer), April (autumn), July (winter) and October (spring).

A review of the quality of the MODIS Chl *a* data in terms of the uncertainty of the Chl *a* product was undertaken by Moore *et al.* (2009) who identified relatively high uncertainty in inshore and nearshore regions (68% and 60% respectively) of south-western Africa when compared to central basin water. To investigate the accuracy of the satellite data, satellite-derived SST and Chl *a* values are compared against their *in situ* counterparts obtained from the profile data. SST is compared against

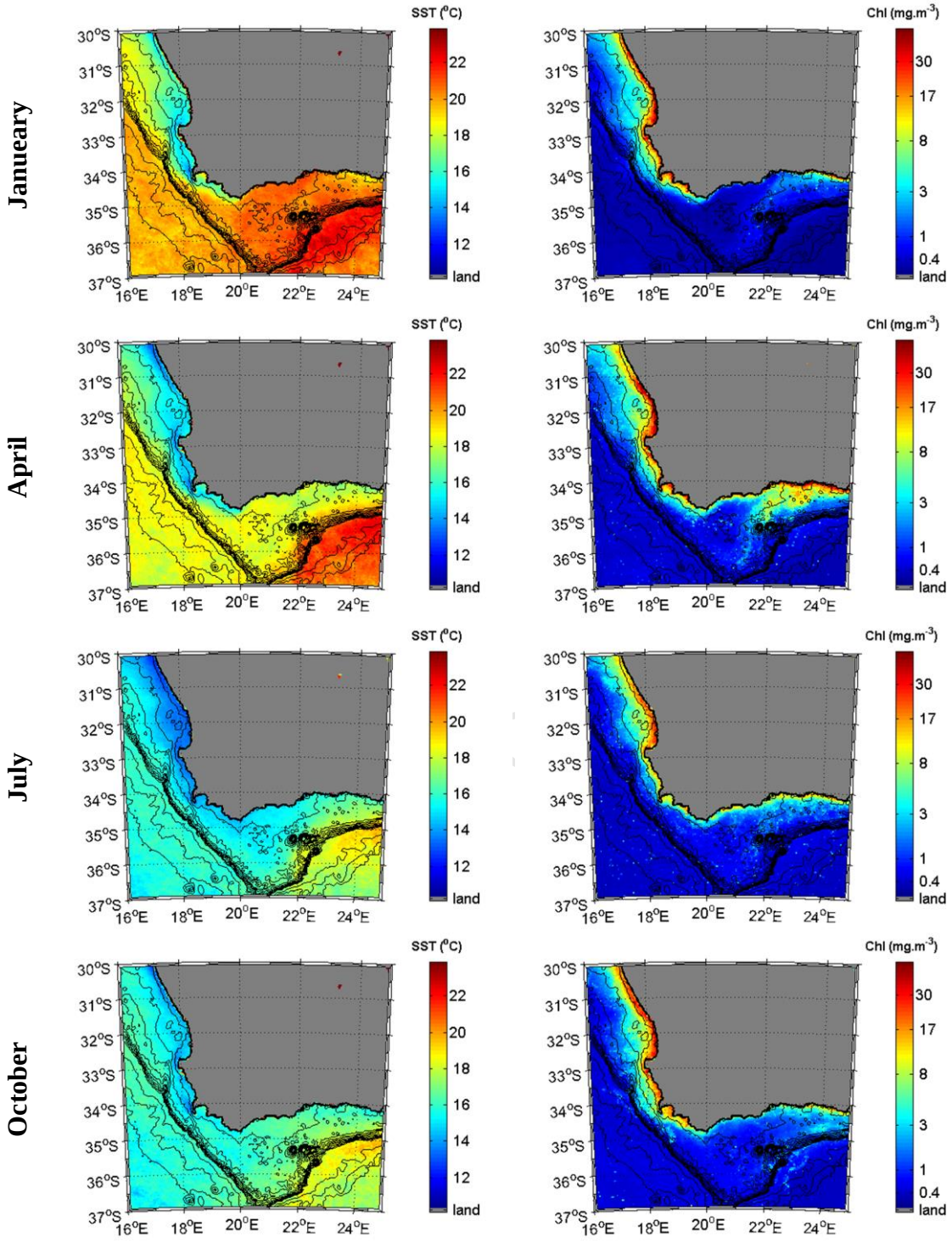


Figure 2.3 Monthly climatologies of MODIS SST (left column) and surface Chl *a* (right column) showing the seasons summer (January), autumn (April), winter (July) and spring (October). Depth contours show 100 m intervals from 0 m to 1000 m and 1000 m intervals thereafter.

the first available data point from the surface because the sensor observes only the thin surface layer of the ocean (Esaias *et al*., 1998). For the *in situ* Chl *a* data the depth to which the sensors sees is defined by the inverse of the Kd(490) value. According to Gordan and McCluney (1975) 90% of the blue-green signal originates from this depth. MODIS data also provides the Kd(490) value that is used to compare the mean of the water column Chl *a* concentration to the satellite data.

Wind

The dynamics of the upper water column are controlled by processes operating at the interface between the atmosphere and the ocean surface and by internal processes. During upwelling the most important is the momentum flux from surface wind stress. The surface stress curl and direction of wind are the major forces that control the upwelling process through Ekman pumping and Ekman transport respectively (Peterson and Stramma, 1991; Capet *et al*., 2004). In addition, heat fluxes and turbulence in the upper ocean are also modulated by surface winds. Changes in local wind have important consequences for the timing and position of the sequence of biological events following upwelling, mainly through the relative stability of the upwelled water (Andrews and Hutchings, 1980; Hutchings *et al*., 1984; Brown, 1986; Brown and Hutchings, 1987). The correspondence between the wind-field and upwelling is considered robust in both the spatial and temporal dimensions (Hardman-Mountford *et al*., 2003).

Daily surface wind stress data (at a 10 m reference height) were retrieved from the NASA SeaWinds scatterometer onboard QuikSCAT via IFREMER, Plouzané (France) at 0.5° x 0.5° resolution. QuikSCAT is in a sun-synchronous near-polar orbit that covers about 90% of Earth's surface each day. It crosses the equator on

its northward path at approximately 06:00 AM. The instrument is a microwave radar that has collected data on surface wind at 12.5 km^2 and 25 km^2 resolution under all weather and cloud conditions from July 1999 to November 2009. The pulses transmitted over a 1,800 km wide band are received as backscattered information on the state of the ocean surface roughness. The multiple near-simultaneous transmissions from slightly different angles are able to resolve both wind speed and direction. The resolution of the data means useful data are limited to a minimum distance of 25 km from the coast.

The QuikSCAT wind data has both U (meridional) and V (zonal) stress components. On the West Coast the V component is understood to be the most important in terms of upwelling whereas the U component is important along the South Coast (Nelson and Hutchings, 1983; Probyn *et al*., 1994). Cross-shore wind gradients tend to be strong near the coast due to the land-sea discontinuity of surface drag and the coastal orography (Edwards *et al*., 2001). These gradients are known to have a strong influence on upwelling intensity (Capet *et al*., 2001; Renault *et al*., 2009; Jin *et al*., 2009) but are not well captured by QuikSCAT (Croquette *et al*., 2007). Therefore only the wind stress vectors and not the wind stress curl are used. Daily wind stress data are sampled randomly for each day from 2002 to 2009 using the location grid in Figure 2.2. Data are categorized as either *inshore* (< 200 m bottom depth) or *offshore* (200-700 m bottom depth) and according to season.

2.3 **Results**

2.3.1 ***In Situ* data**

Profiles

Figure 2.4 shows the seasonal sampling and distribution of the collected profiles. The West Coast and West Agulhas Bank was frequently sampled in summer, autumn and spring but autumn sampling is largely confined to the inshore region (< 200 m bottom depth). In winter, sampling only occurred along two monitoring lines on the West Coast with the remainder of sampling carried out on the East Agulhas Bank. No sampling was done on the West Agulhas Bank in winter. The East Agulhas Bank was most frequently sampled in spring followed by autumn and summer.

Figure 2.5 and Figure 2.6 show the results of applying the *k*-means algorithm to the temperature and Chl *a* depth profiles respectively. After clustering, all profiles including those with depths shallower than 50 and 40 m, are labelled according to a minimum sum of squares between the profile sample and the cluster means. The temperature profiles are arranged approximately according to increasing surface temperature.

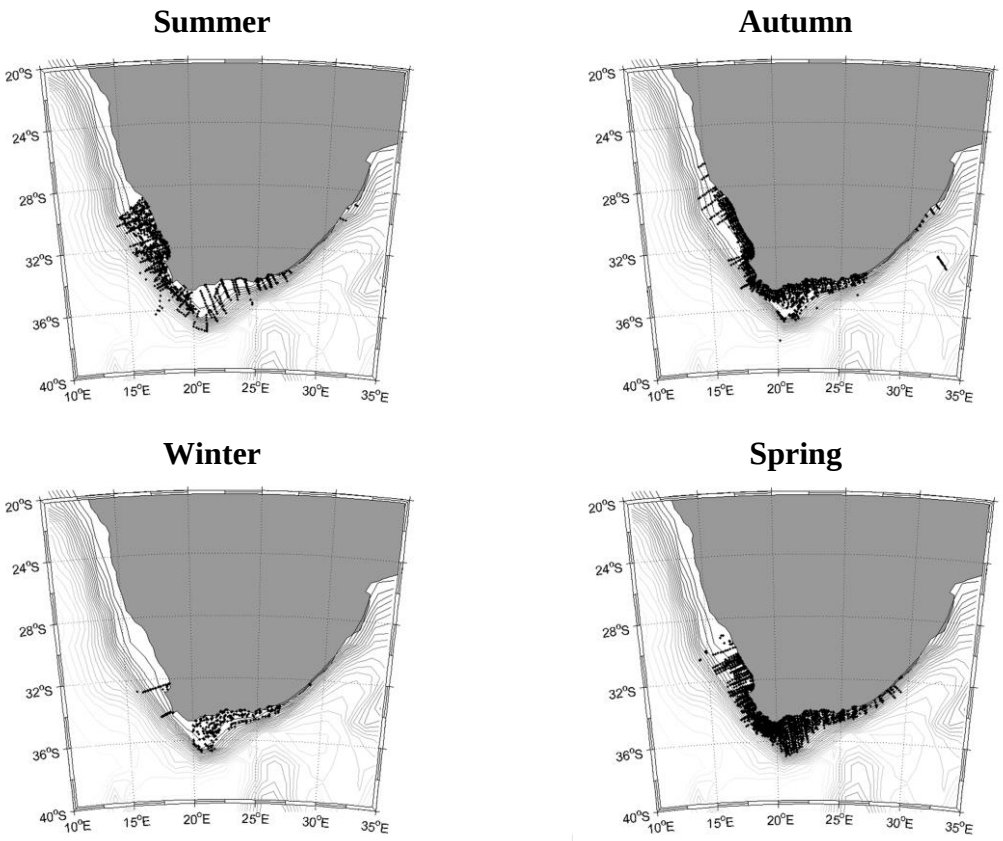


Figure 2.4 Seasonal distribution of the location of collected temperature and Chl *a* profiles around the coast of South Africa.

The range of temperature profiles includes shallow mixing layer depths of 10-15 m with cool surface water (Profile 2) to deep well mixed layers with warm surface water (Profile 11). Varying degrees of mixing layer depth, thermocline intensity and upper water column temperature are indicated. The Chl *a* profiles are arranged according to increasing integrated biomass within the profile (Fig. 2.6) and range from homogenous low Chl *a* throughout the water column (Profile 1) to surface blooms of varying concentration (Profiles 6, 8, 11) and deep Chl *a* maxima (Profiles 3, 5, 7 and 10).

The location of the 12 *k*-means temperature clusters are shown in Figure 2.7. Due to the strong seasonal bias in the sampling of the profiles it is difficult to analyze the seasonal occurrence of the profiles. Temperature Profile 1 shows the coldest

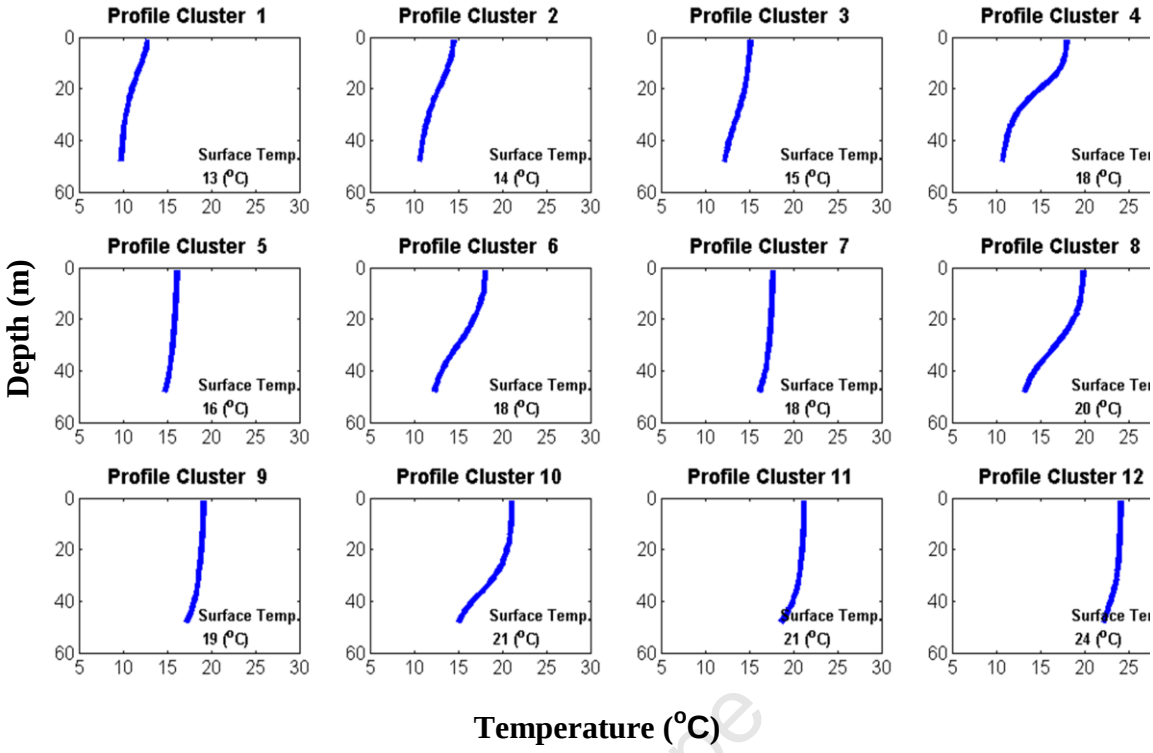


Figure 2.5 The *k*-means cluster centres for the temperature profile data.

surface water and is confined close inshore where the bottom depth is typically less than 100 m. It is frequently observed on the West Coast and occasionally on the Agulhas Bank where upwelling occurs. Profiles 2 and 3 indicate warmer surface water and are observed close inshore but further offshore than Profile 1. Profiles 4, 5, 8 and 10 depict a strong thermocline and progressively deeper mixing layers. They show warmer water at either the surface or at 40 m. They are observed throughout the regions, progressing further offshore from the inner-shelf to the shelf edge. Profiles 6, 7 and 9 indicate either a deep mixing layer to ca. 40 m or very weak stratification. They are observed throughout the regions and progressively in deeper water. On the West Coast they are mostly found in water > 200 m bottom depth whereas over the Agulhas Bank Profile 6 is observed close to the coast.

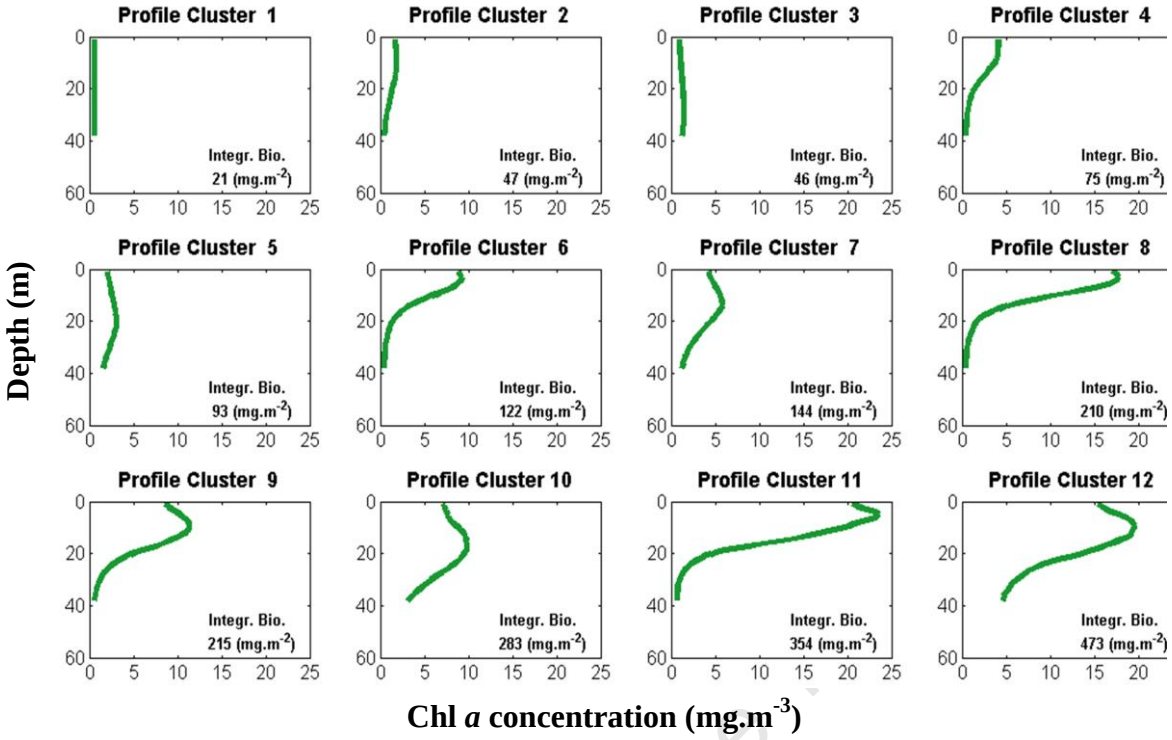


Figure 2.6 The *k*-means cluster centres for the Chl *a* profile data.

Profiles 11 and 12 represent the warmest water at the surface and 40 m depth. They are observed over the shelf edge and in deeper water.

The location of the Chl *a* profiles are shown in Figure 2.8. Profile 1, which has the lowest integrated biomass, is observed throughout the regions but on the West Coast it is mostly observed either close inshore (bottom depth ca. < 50 m) or close to the shelf edge. Profiles 2 and 3 also have low biomass but Profile 3 has a subsurface peak deeper than ca. 20 m. Where Profile 2 is usually found close to the coast, Profile 3 is typically observed across the West and East Agulhas Bank and at the shelf edge on the West Coast. Profiles 4, 6, 8 and 11 have near-surface peaks that increase in biomass and are progressively found closer inshore in fewer numbers and become more confined to the West Coast upwelling centres. Profile 7 has a moderate subsurface peak at ca. 15-20 m and is frequently observed at all

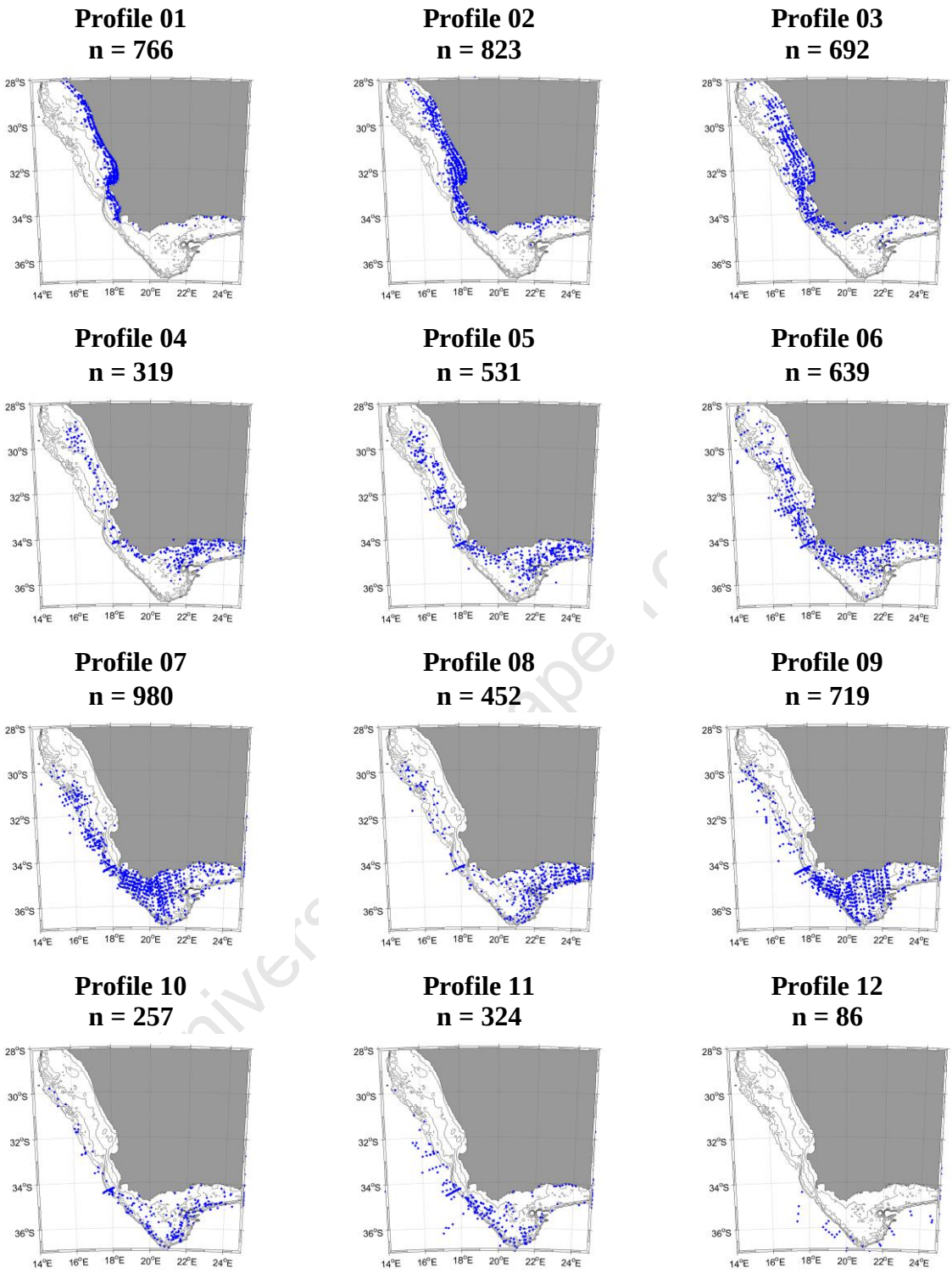


Figure 2.7 Location of the 12 temperature profile clusters.

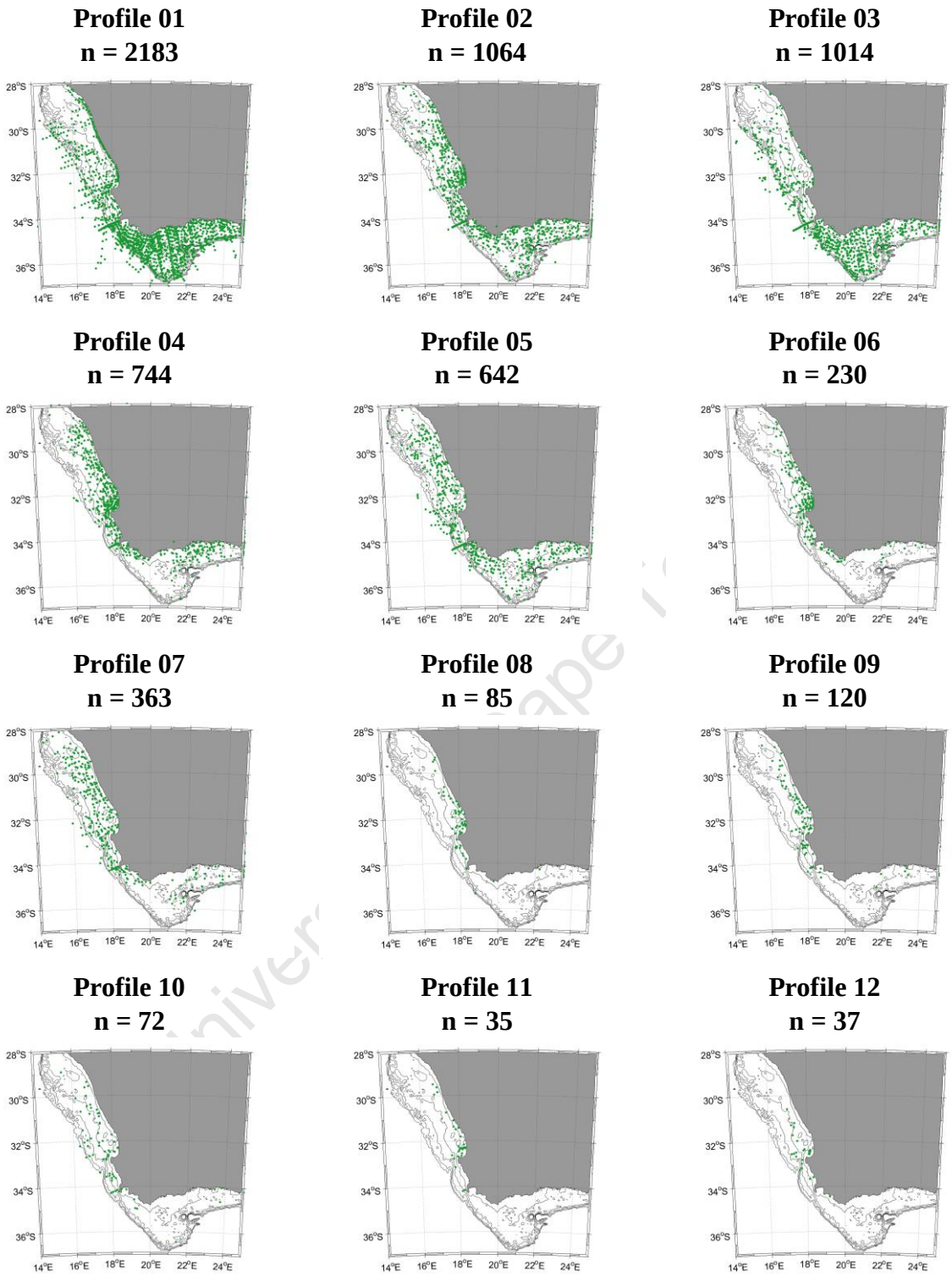


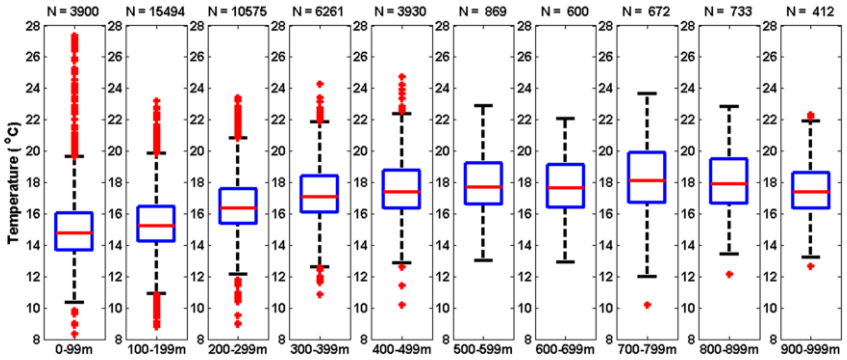
Figure 2.8 Location of the 12 Chl *a* profile clusters.

bottom depths on the West Coast and East Agulhas Bank but closer inshore on the West Agulhas Bank. Profile 10 has a larger subsurface peak at ca. 15-20 m and is occasionally observed in the region of Profile 7 on the West Coast. Profile 12, which has the largest integrated biomass and a peak at ca. 10-15 m depth, is occasionally observed downstream of the Cape Peninsula and Cape Columbine upwelling cells and in St Helena Bay.

Depth

Figures 2.9, 2.10 and 2.11 shows box and whisker plots of the SST and Chl *a* data within 100 m bottom depth intervals on the West Coast, West Agulhas Bank and East Agulhas Bank respectively. The box indicates the lower quartile, median and upper quartile values. The whiskers indicate the extent of the data (within 1.5 times the interquartile range) and the red dots indicate the outlier data (> 1.5 times the interquartile range). The high number of outliers in some depth intervals indicates that very high biomass events do occur but that their frequency is far less than those occurring within the inter-quartile range.

SST



Surface Chl *a*

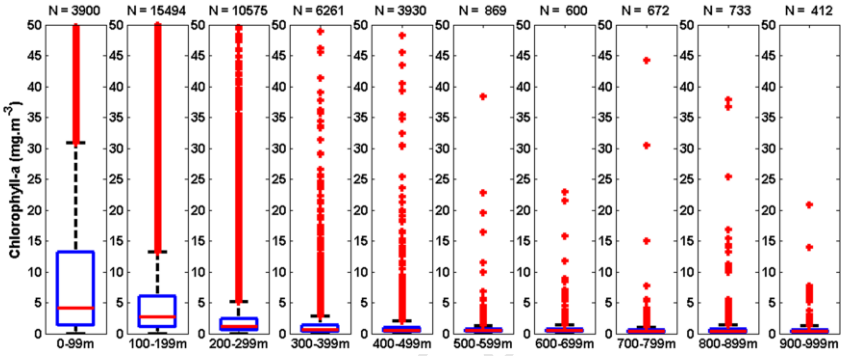
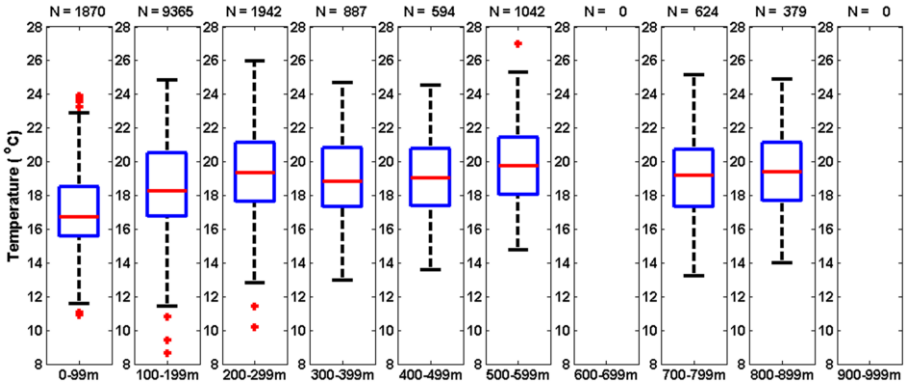


Figure 2.9 Box and whisker plots of SST and surface Chl *a* within 100m bottom depth intervals for the West Coast. The blue box indicates the lower and upper quartile limits and the red line is the median value. The whiskers indicates the extent of the data (within 1.5 times the interquartile range) and the red dots indicate the outlier data (> 1.5 times the interquartile range).

Cooler water and higher Chl *a* concentration is consistently found closer inshore in all regions. On the West Coast this occurs progressively inshore of 500 m bottom depth whereas on the West and East Agulhas Bank it occurs inshore of 200-300 m bottom depth.

SST



Surface Chl *a*

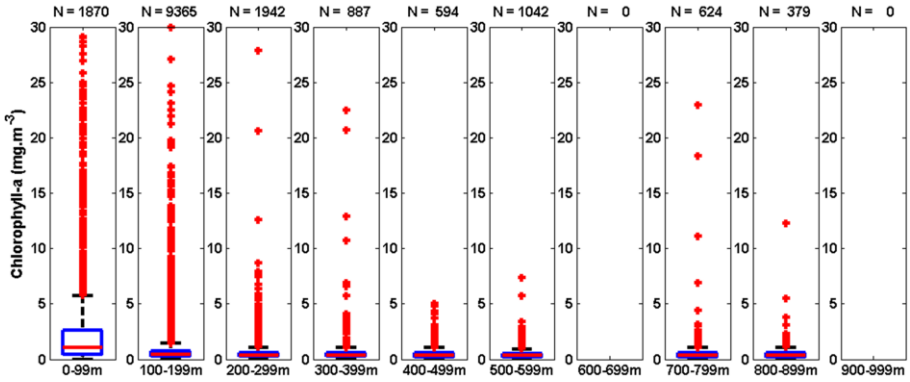


Figure 2.10 Box and whisker plots of SST and surface Chl *a* with 100 m bottom depth intervals for the West Agulhas Bank. The blue box indicates the lower and upper quartile limits and the red line the median value. The whiskers indicates the extent of the data (within 1.5 times the interquartile range) and the red dots indicate the outlier data (> 1.5 times the interquartile range).

On the West Agulhas Bank (Fig. 2.10) the SST is shown to change substantially from the coast to 300 m bottom depth whereas relatively high Chl a concentrations are mostly confined within the 100 m bottom depth limit.

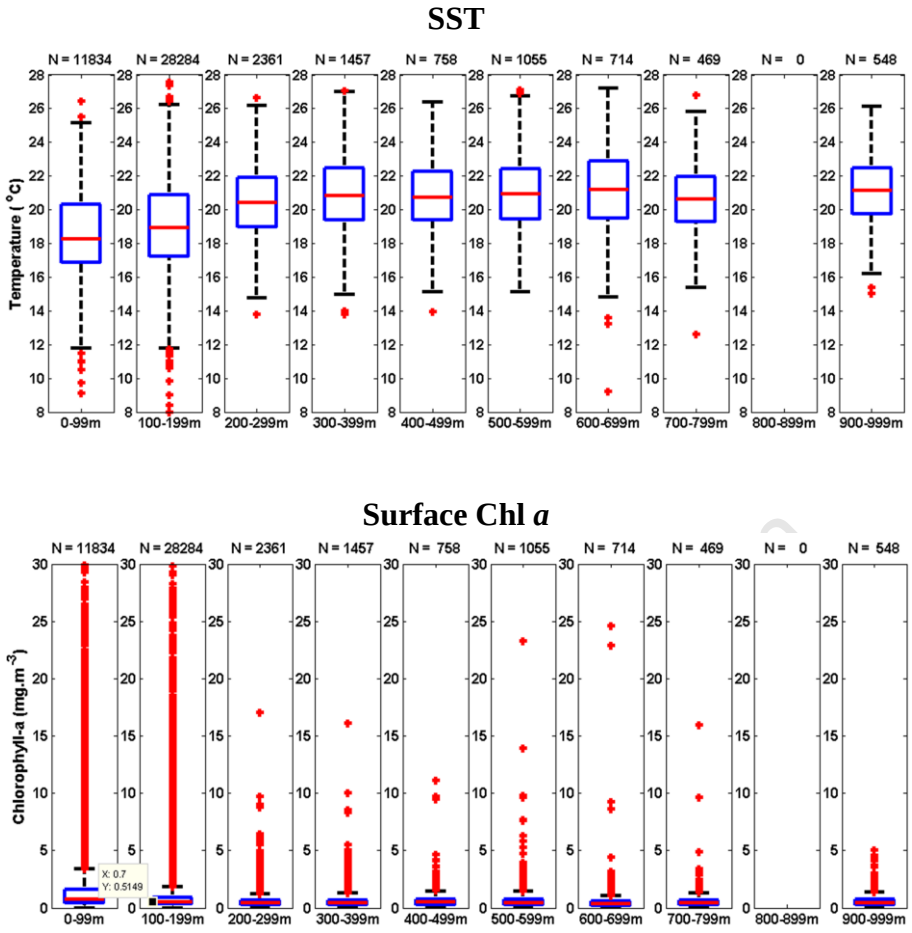


Figure 2.11 Box and whisker plots of SST and surface Chl *a* with 100m bottom depth intervals for the East Agulhas Bank. The blue box indicates the lower and upper quartile limits and the red line the median value . The whiskers indicates the extent of the data (within 1.5 times the interquartile range) and the red dots indicate the outlier data (> 1.5 times the interquartile range).

The East Agulhas Bank (Fig. 2.11) indicates warmer water and lower Chl a biomass inshore (< 100 m bottom depth) than the other two regions but a similar change over the first three bottom depth intervals.

2.3.2 Satellite data

SST and Chl *a*

The results of the regression of the SST and log-transformed Chl *a* satellite data against the corresponding *in situ* data are shown in Figure 2.12. The temperature data show good correlation with a correlation coefficient of 0.86. Some of the variance of the *in situ* data can be attributed to using the 3 m depth values from the profiles instead of the surface values, which was necessary due to frequent missing data in the upper few metres. The Chl *a* data shows a correlation coefficient of 0.68 and the orthogonal regression shows MODIS data tends to underestimate the *in situ* data at low values and and overestimates the *in situ* data at high values (in accordance with Moore *et al*., 2009). Orthogonal regression is used because errors are expected in both the satellite data and the measured data and the regression relates the two variables without assuming that the predictor is accurate The data

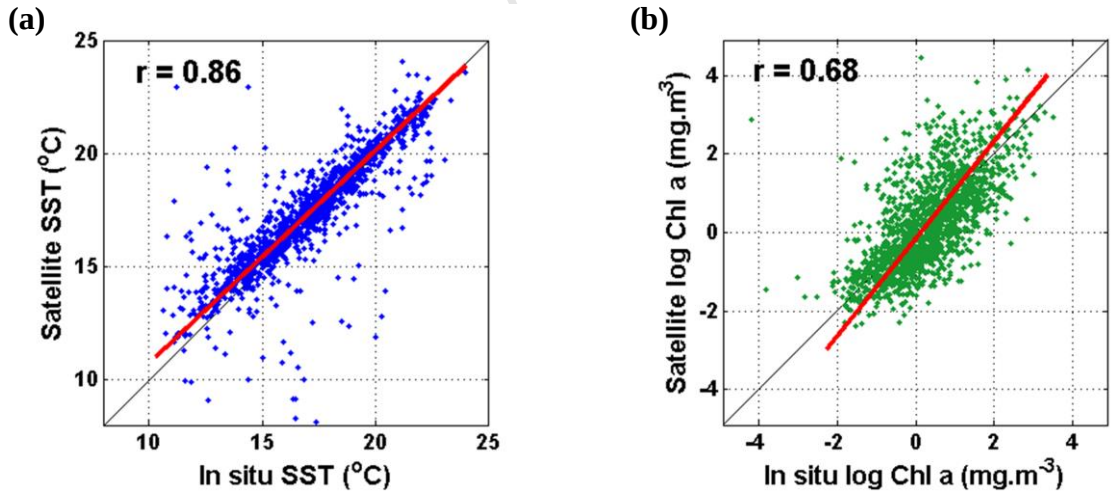


Figure 2.12 Orthogonal regression of the *in situ* data, obtained from the (a) temperature and (b) Chl *a* profiles, against the concurrent MODIS satellite data. The correlation coefficient (*r*) is also shown. The *in situ* SST data are the 3 m depth value and the Chl *a* data are the mean of the layer from the surface to 1 optical depth (the inverse of the satellite-derived attenuation coefficient Kd490).

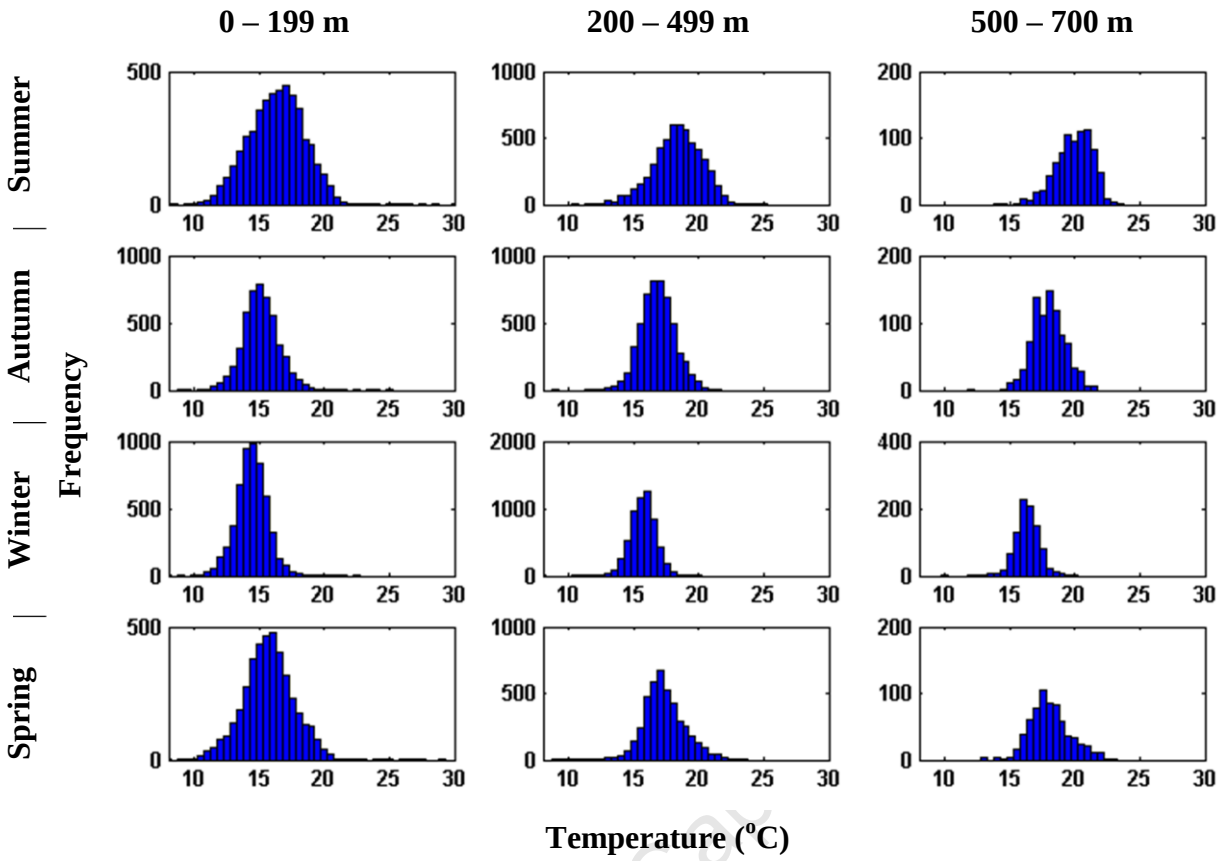


Figure 2.13 Frequency plots of the satellite SST distributions according to season and bottom depth intervals for the West Coast.

tested here shows an average relative error (the mean absolute difference between the *in situ* value and the MODIS estimate) of 169%, which is higher than the 68% found by Moore *et al*. (2009) in similar high-biomass coastal waters.

To investigate the variability of the system from remotely sensed data, seasonal observations of SST and surface Chl *a* are plotted for bottom depth intervals < 200 m (inshore), 200-499 m (shelf) and 500-700 m (offshore). The West Coast SST data (Fig. 2.13) appear normally distributed for all three depth regions. In accordance with expectations, summer and spring show a broader range of temperature particularly inshore. Seasonal change is evident in all regions where winter has the coldest temperatures and summer the warmest. The inshore region has a fairly

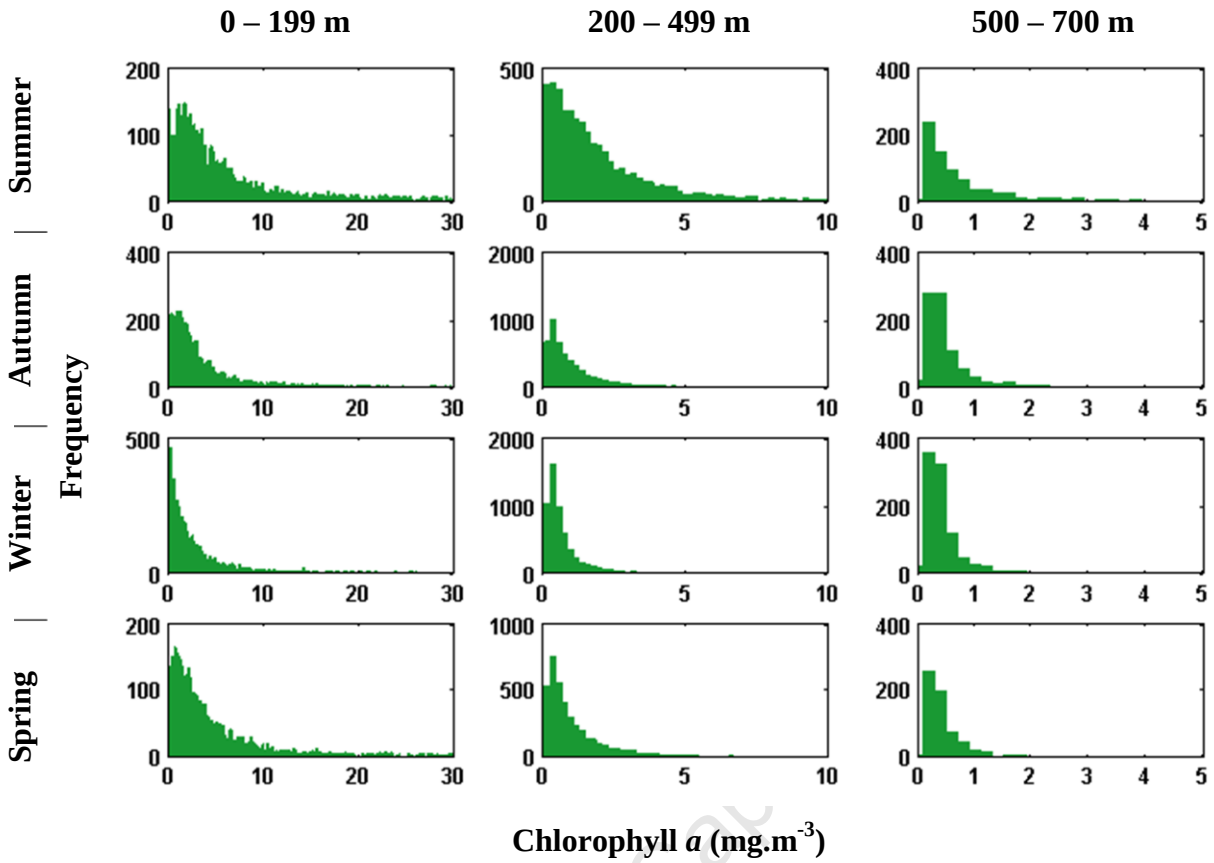


Figure 2.14 Frequency plots of satellite Chl *a* distributions according to season and bottom depth intervals for the West Coast.

consistent lowest temperature (ca. 10°C) whereas the highest temperature fluctuates seasonally (ca. 18-22°C) suggesting upwelling of cold water occurs year round.

The Chl *a* data display log-normal distributions (Fig. 2.14). Seasonal changes in the Chl *a* distributions are evident. Summer and spring have relatively high values more frequently than autumn and winter. In winter the most frequent interval is the lowest (0-0.2 $mg.m^{-3}$). Over the shelf in summer higher values are more common than other seasons whereas offshore the distributions are very similar.

The West Agulhas Bank SST data show approximate normal distributions for most seasons and depth regions however, the distribution in spring appears bimodal (Fig. 2.15). The spring distribution is probably the result of intrusion of Agulhas

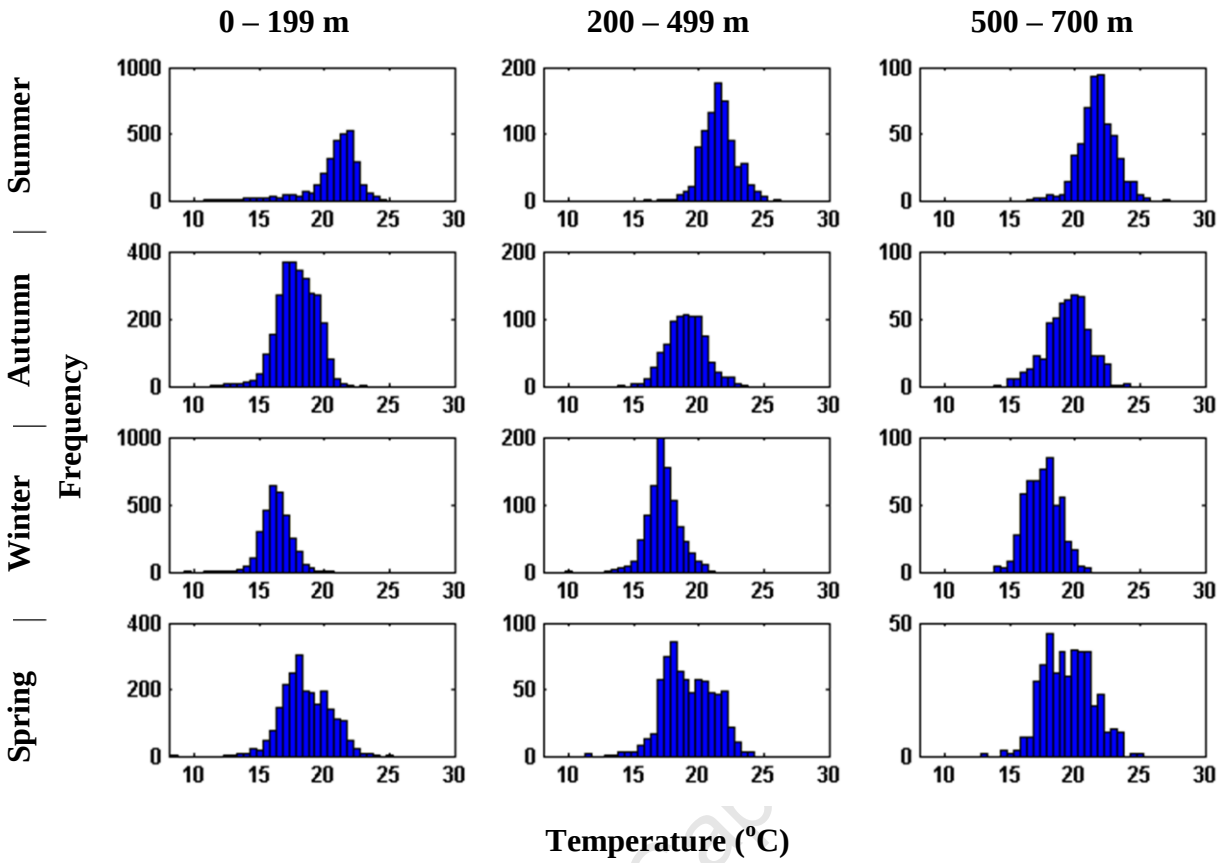


Figure 2.15 Frequency plots of satellite SST distributions according to season and bottom depth for the West Agulhas Bank.

Current water over the bank. There is a clearer seasonal variation in SST distribution than the West Coast. Autumn and spring shows the broadest range of temperatures particularly inshore.

Chl *a* values remain low throughout the seasons (Fig. 2.16). In contrast to the West Coast, relatively high values are more common in autumn and winter than summer and spring.

The East Agulhas Bank SST data (Fig. 2.17) is normally distributed. Unlike the West Coast and West Agulhas Bank the seasonal SST data have similar ranges. Inshore the cold temperature limit of the distributions is similar in autumn, winter and spring but substantially higher in summer.

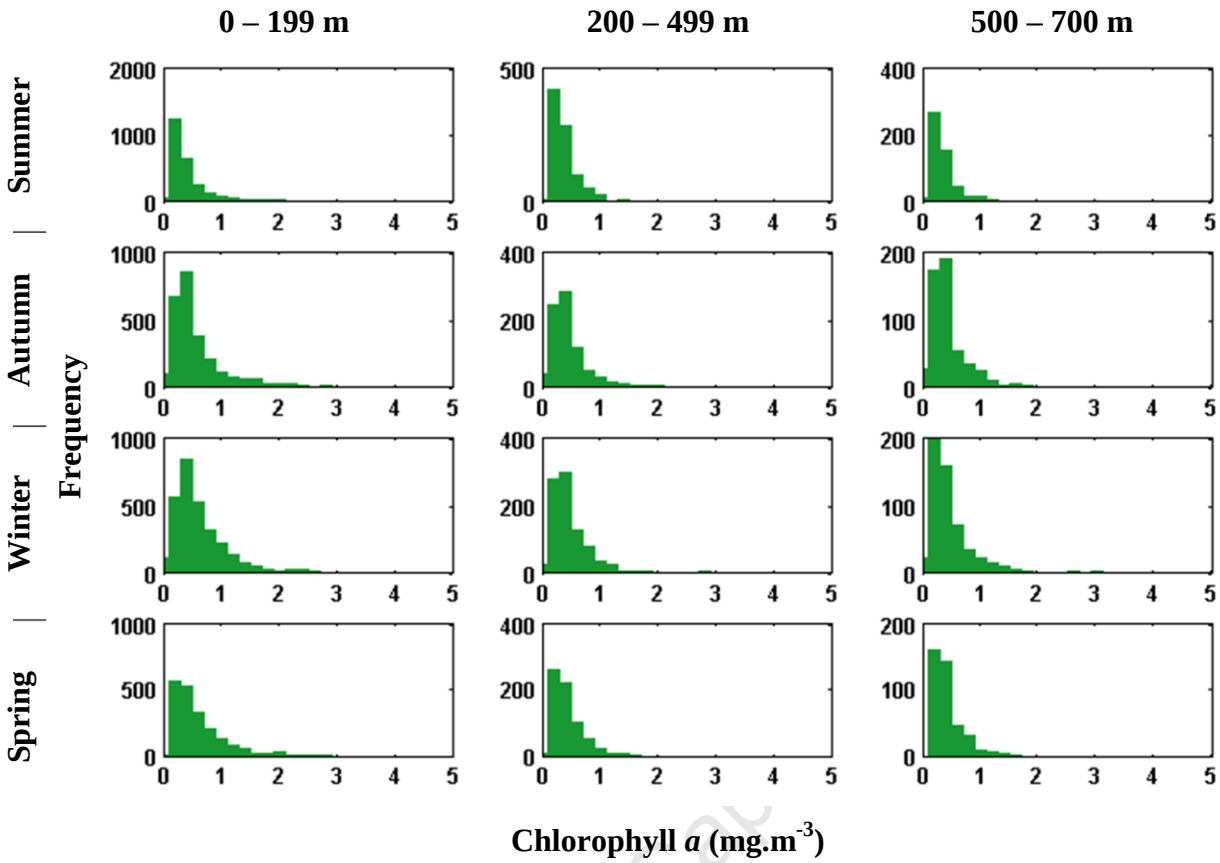


Figure 2.16 Frequency plots of satellite Chl *a* distributions according to season and bottom depth for the West Agulhas Bank.

Chl *a* data (Fig. 2.18) show similar low values in all seasons and across all depths. Similar to the West Agulhas Bank, the lowest intervals are most frequent in summer while autumn and winter show that slightly higher values are more common. The statistics of the SST data are presented in Table 2.1 and that of the log-transformed Chl *a* data in Table 2.2. The SST data shows that inshore on the West Coast temperatures are substantially colder that the Agulhas Bank in all seasons. The West Agulhas Bank is colder than the East Agulhas Bank and shows a significant increase in variability in summer. Both the West Coast and West Agulhas Bank have highest variability in summer and spring whereas the East Agulhas Bank has highest variability in autumn and spring. However, further offshore both the West

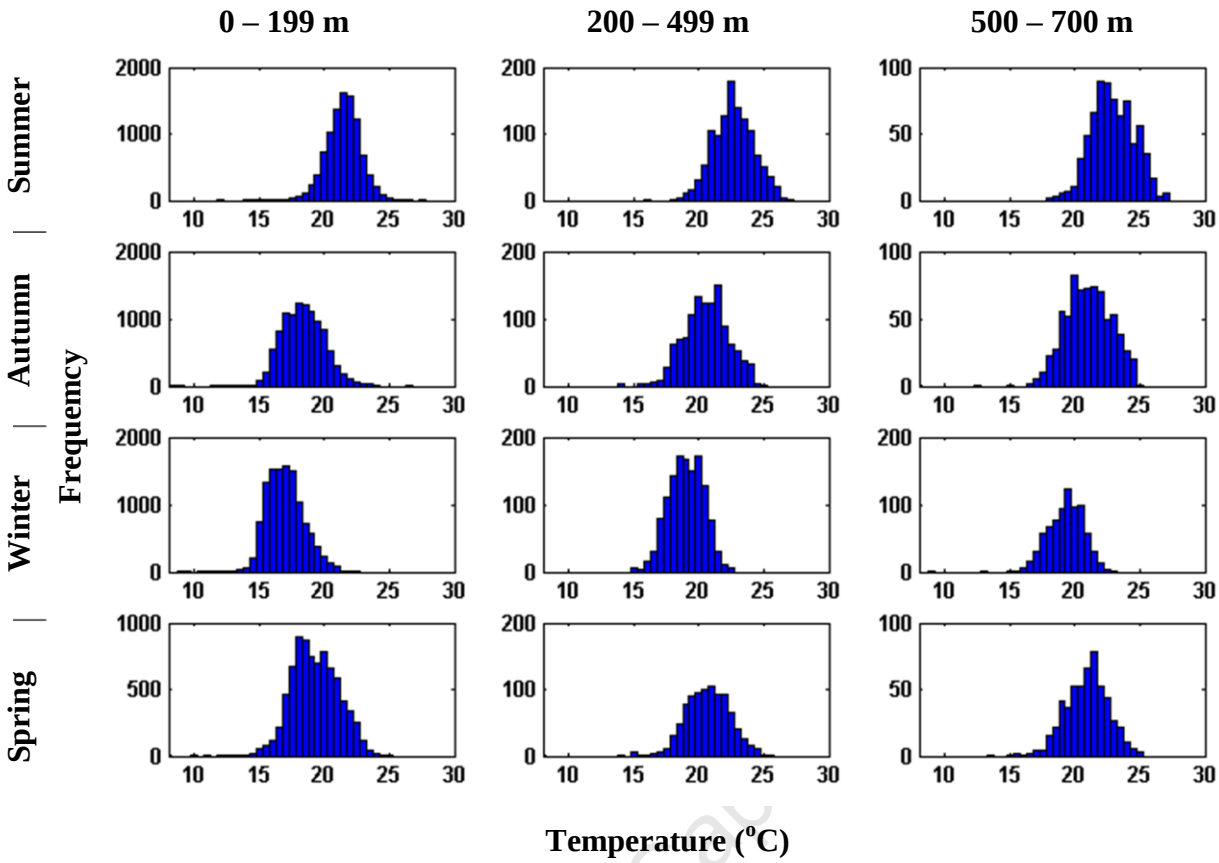


Figure 2.17 Frequency plots of satellite SST distributions according to season and bottom depth for the East Agulhas Bank.

Agulhas Bank and the East Agulhas Bank have highest variability in autumn and spring.

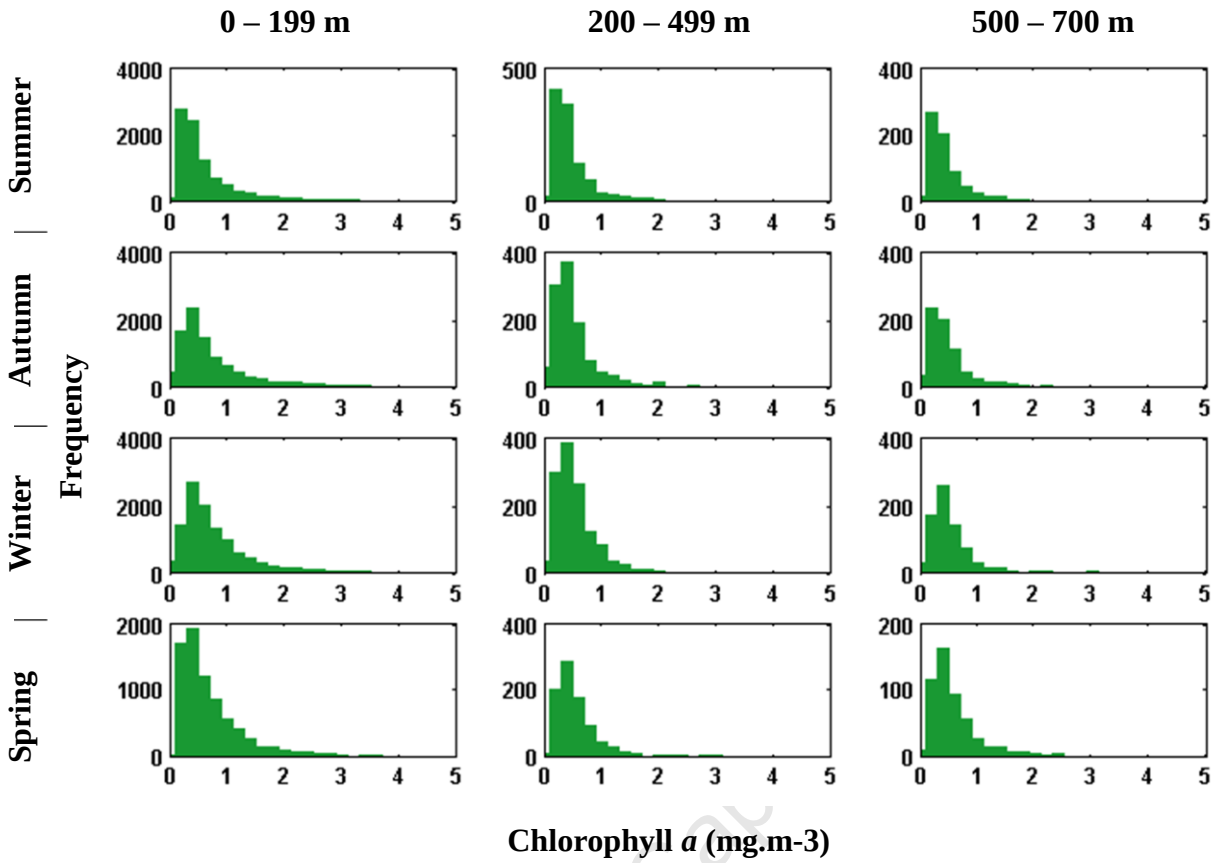


Figure 2.18 Frequency plots of satellite Chl *a* distributions according to season and bottom depth for the East Agulhas Bank.

The seasonal mean Chl *a* and standard deviation (Table 2.2) vary considerably among regions and also among depth intervals within regions. The West Coast has maximum Chl *a* values in summer for all depth sub-regions but inshore variability is highest in winter and highest in summer further offshore. Slightly higher values of mean surface Chl *a* are observed in winter over the West Agulhas Bank but highest variability changes from summer to autumn to winter over the depth intervals. Over the East Agulhas Bank highest variability is continuously observed in autumn but the highest mean Chl *a* is observed in autumn inshore and in winter and spring further offshore.

Table 2.1 Statistics (mean and standard deviation) of the seasonal MODIS SST (°C) for three depth categories of the three sub-regions.

	West Coast		West Agulhas Bank		East Agulhas Bank	
	Mean	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**
			Inshore			
Summer	**16.35**	**2.18**	**20.86**	**1.92**	**21.47**	**1.32**
Autumn	**15.09**	**1.38**	**17.85**	**1.48**	**18.35**	**1.64**
Winter	**14.42**	**1.21**	**16.25**	**1.10**	**16.93**	**1.41**
Spring	**15.74**	**2.01**	**18.57**	**1.85**	**19.29**	**1.81**
			Shelf			
Summer	**18.34**	**1.85**	**21.51**	**1.23**	**22.64**	**1.52**
Autumn	**16.84**	**1.31**	**18.93**	**1.45**	**20.70**	**1.76**
Winter	**15.67**	**1.00**	**17.33**	**1.20**	**19.04**	**1.37**
Spring	**17.29**	**1.59**	**19.20**	**1.81**	**20.63**	**1.76**
			Offshore			
Summer	**19.63**	**1.52**	**21.82**	**1.58**	**22.83**	**1.73**
Autumn	**17.90**	**1.24**	**19.68**	**1.67**	**21.12**	**1.90**
Winter	**16.25**	**0.85**	**17.82**	**1.26**	**19.38**	**1.35**
Spring	**17.97**	**1.43**	**19.59**	**1.84**	**21.07**	**1.71**

Table 2.2 Statistics (mean and standard deviation) of the seasonal MODIS surface Chl *a* ($mg.m^{-3}$) for three depth categories of the three sub-regions. Statistics were obtained from the log transformed Chl *a* data.

	West Coast		West Agulhas Bank		East Agulhas Bank	
	Mean	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**
			Inshore			
Summer	**4.33**	**3.43**	**0.48**	**3.31**	**0.55**	**2.72**
Autumn	**2.66**	**3.40**	**0.54**	**2.63**	**0.78**	**3.19**
Winter	**2.75**	**3.75**	**0.59**	**2.34**	**0.71**	**2.41**
Spring	**3.55**	**3.57**	**0.59**	**2.66**	**0.61**	**2.39**
			Shelf			
Summer	**1.55**	**3.11**	**0.35**	**2.06**	**0.40**	**2.09**
Autumn	**0.87**	**2.74**	**0.41**	**2.12**	**0.45**	**2.23**
Winter	**0.61**	**2.37**	**0.43**	**2.00**	**0.48**	**1.98**
Spring	**0.94**	**2.81**	**0.39**	**2.03**	**0.49**	**1.97**
			Offshore			
Summer	**0.73**	**2.88**	**0.32**	**1.91**	**0.38**	**2.18**
Autumn	**0.41**	**2.15**	**0.37**	**1.98**	**0.41**	**2.47**
Winter	**0.37**	**1.83**	**0.42**	**2.11**	**0.47**	**1.98**
Spring	**0.43**	**2.34**	**0.36**	**1.75**	**0.51**	**2.10**

The three sub-regions display unique characteristics in the SST and surface Chl *a* data. Generally, the West Coast has cooler temperatures and higher Chl *a* values with Chl *a* peaks in summer and spring. The East Agulhas Bank is warmer and has higher Chl *a* values than its western counterpart but in contrast to the West Coast, peak Chl *a* on the Agulhas Bank occurs in autumn and winter. However, these characteristic patterns overlap for example, cold water of ca. 16°C may indicate typical winter temperatures inshore on the West Agulhas Bank but also indicates

maturing upwelled water on the West Coast in summer. Similarly, Chl *a* concentrations of ca. 1 mg.m^{-3} indicate typical inshore concentrations over the East Agulhas Bank but also shelf waters on the West Coast. Fortunately, the SST and Chl *a* information will be used in conjunction with other information such as region and depth and should therefore be discretized in such a way as to be useful for all regions rather than to discriminate among regions. This is done by considering both the distribution and statistics of the data discussed above as well as published research. The selected intervals are presented in Table 2.3.

Table 2.3 Discrete states of satellite-derived SST and Chl *a*.

SST (^{o}C)	Chl *a* (mg.m^{-3})	State
< 13.5	0 - 0.39	1
13.5 – 14.99	0.4 – 0.99	2
15 - 16.49	1 – 2.99	3
16.5 – 17.99	3 – 7.99	4
18 - 19.49	8 – 16.99	5
19.5 – 21.99	17 – 29.99	6
> 22	> 30	7

Wind

Most important to primary production in the southern Benguela is the variability of the wind associated with synoptic atmospheric phenomena (Nelson and Hutchings, 1983). Figure 2.19a,b illustrate the passage of a cyclonic and anticyclonic feature, which drive the characteristic winds in the region, and the resulting zonal (U wind) and meridional (V wind) components. This example is to illustrate both the short-term temporal variability of the wind captured by the QuikSCAT data and the small scale alongshore variability. The six-day sequence shows a semi-stationary summer

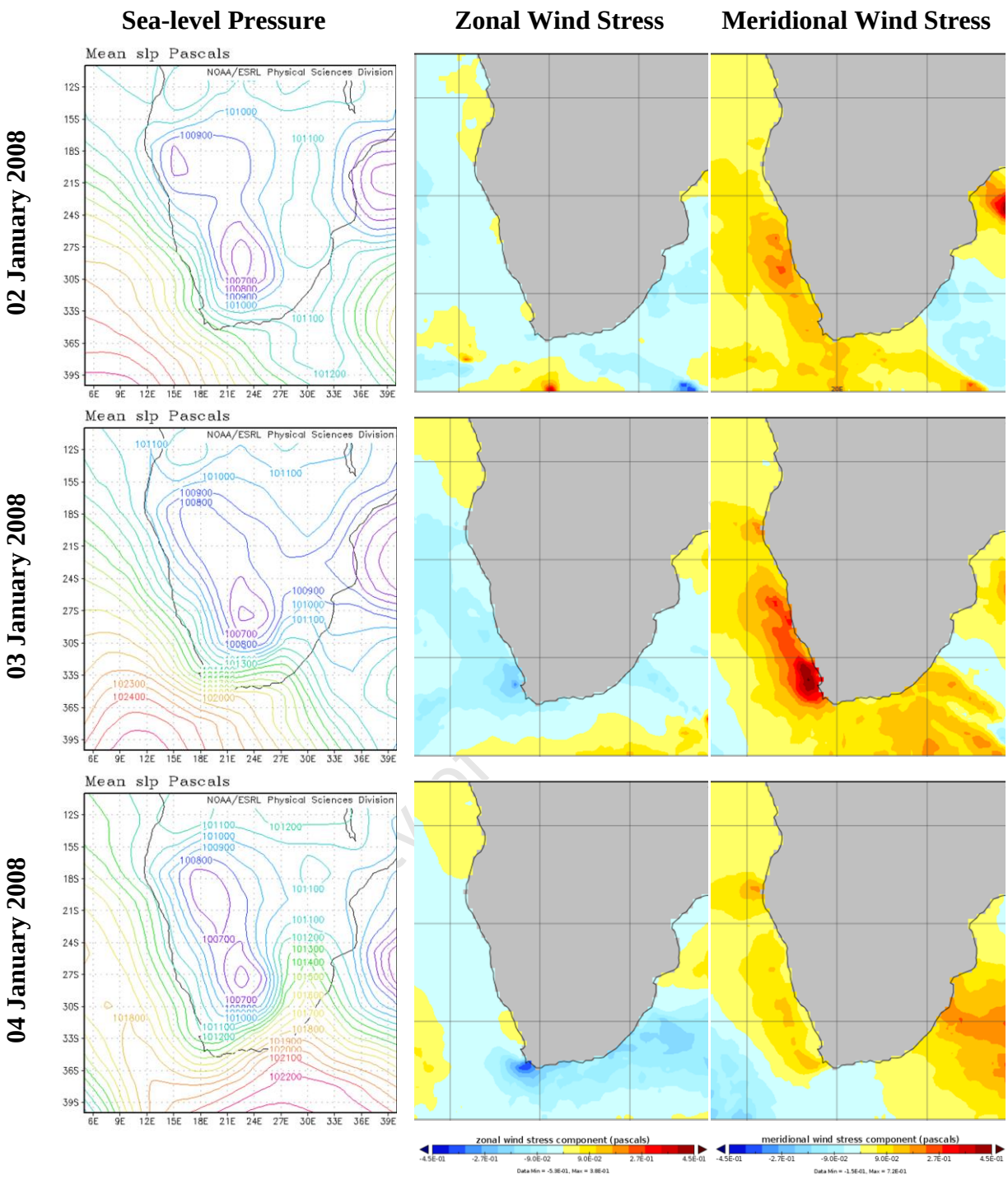


Figure 2.19a The eastward tracking of an anticyclone from 2-4 January 2008 shown by the NCEP Reanalysis sea-level pressure (left panel), QuikSCAT zonal wind stress (centre panel) and meridional wind stress (right panel).

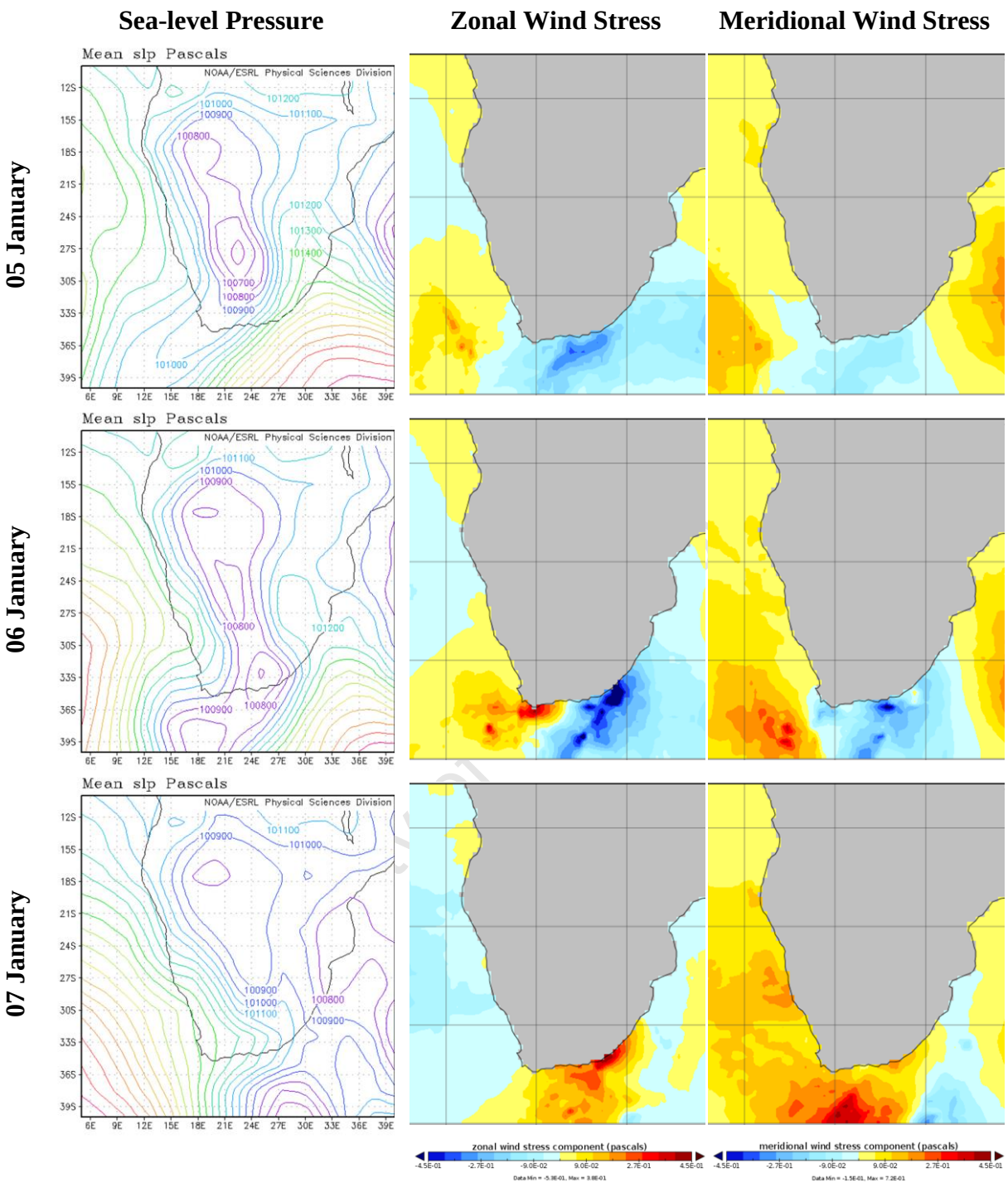


Figure 2.19b The eastward tracking of a cyclonic depression from 5-7 January 2008 show by the NCEP Reanalysis sea-level pressure (left panel), QuikSCAT zonal wind stress (centre panel) and meridional wind stress (right panel).

heat low over the arid region of southern Africa and the offshore pressure gradient over the West Coast which forces the prevailing summer south-easterlies. The northern extent of an eastward tracking anticyclonic feature (2-4 January) is followed by a cyclonic depression (5-6 January) to the south of the continent. Both the NCEP Reanalysis plots of sea-level pressure and the QuikSCAT wind stress data indicate an intensification of easterly wind at the southern tip of the continent (2-4 January), which switches from south-easterly to north-easterly as the anticyclone passes. From 4-5 January north-westerly winds, associated with frontal systems develop over the southern tip and are replaced by southerly winds on 6 January. The southern inshore area of the West Coast has very different winds with intense south-easterly wind developing on 3 January in between two days of moderate south-westerly. Further north along the coast the wind remains south-easterly over this period.

To investigate the wind data further, the sampled daily zonal and meridional wind stress is plotted for each region and season, and for the inshore (< 200 m bottom depth) and shelf (200-500 m bottom depth). Figure 2.20 shows the wind data distribution and the mean (red line) for the West Coast. In all seasons and for both depth intervals the most frequent wind stress is the weakest. There is little difference between the inshore and shelf region although over the shelf winds are slightly stronger, particularly the zonal component in winter and the meridional component in summer. The meridional wind stress consistently shows the largest wind stress. Northerly winds are most common in winter but southerly winds dominate all seasons and are strongest in summer and spring. The mean zonal component shows only a slight shift from easterly in summer and spring to near neutral in autumn and winter.

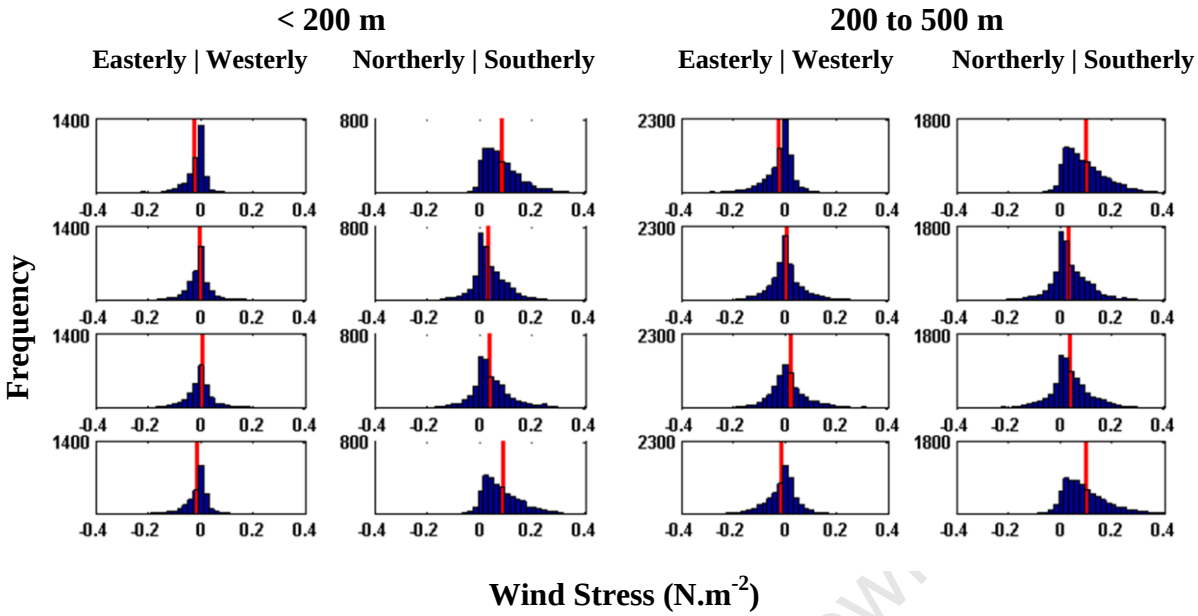


Figure 2.20 West Coast wind stress distribution for summer (top row) to spring (bottom row) using daily sampled data from 2002-2009. The red line indicates the mean wind stress.

Over the West Agulhas Bank (Fig. 2.21) the seasonal variability is most evident in the zonal wind component. Strong westerly wind occurs in winter and autumn while strongest easterly wind is observed in summer and spring. During summer and spring, mean meridional wind stress shifts towards easterly wind.

Over the East Agulhas Bank (Fig. 2.22) the wind is very similar to the West Agulhas Bank. The easterly wind component is slightly stronger.

To simplify the wind data discrete intervals are based on a simple interpretation of calm winds that are assumed to result in a strong stabilization of the water column, moderate winds that mix the upper layer and strong winds that create a deep mixing layer. The limits of the intervals are also selected to ensure there are sufficient data in each interval. The same intervals are applied to the zonal and meridional wind stress components. The results are shown in Table 2.4.

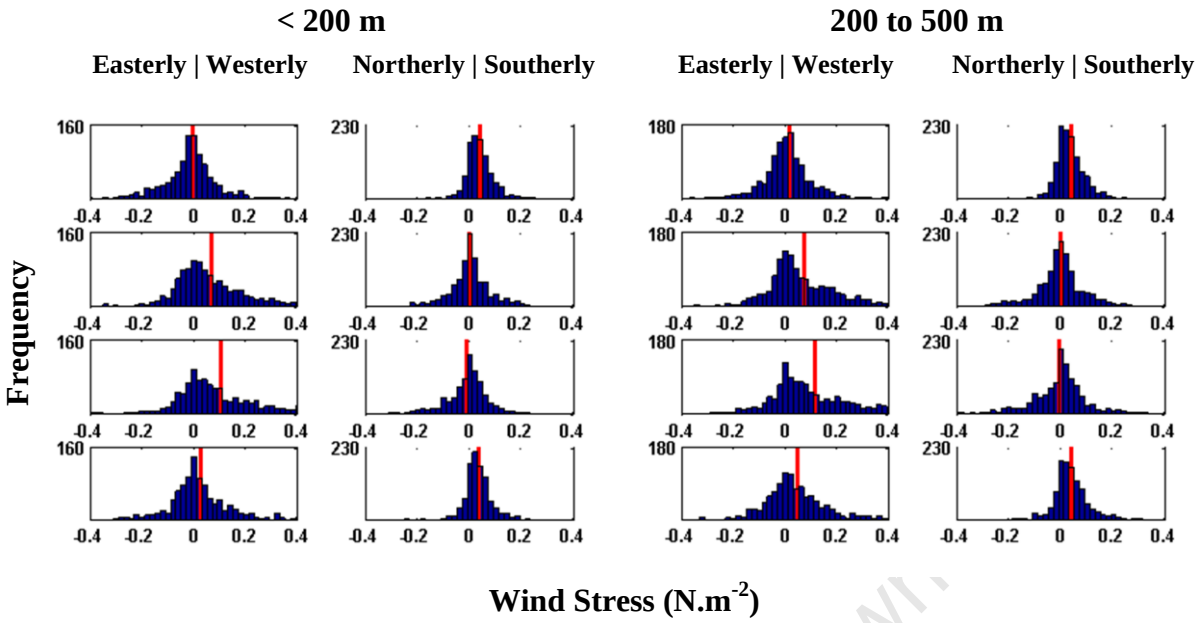


Figure 2.21 West Agulhas Bank wind stress distribution for summer (top row) to spring (bottom row) using daily sampled data from 2002-2009. The red line indicates the mean wind stress.

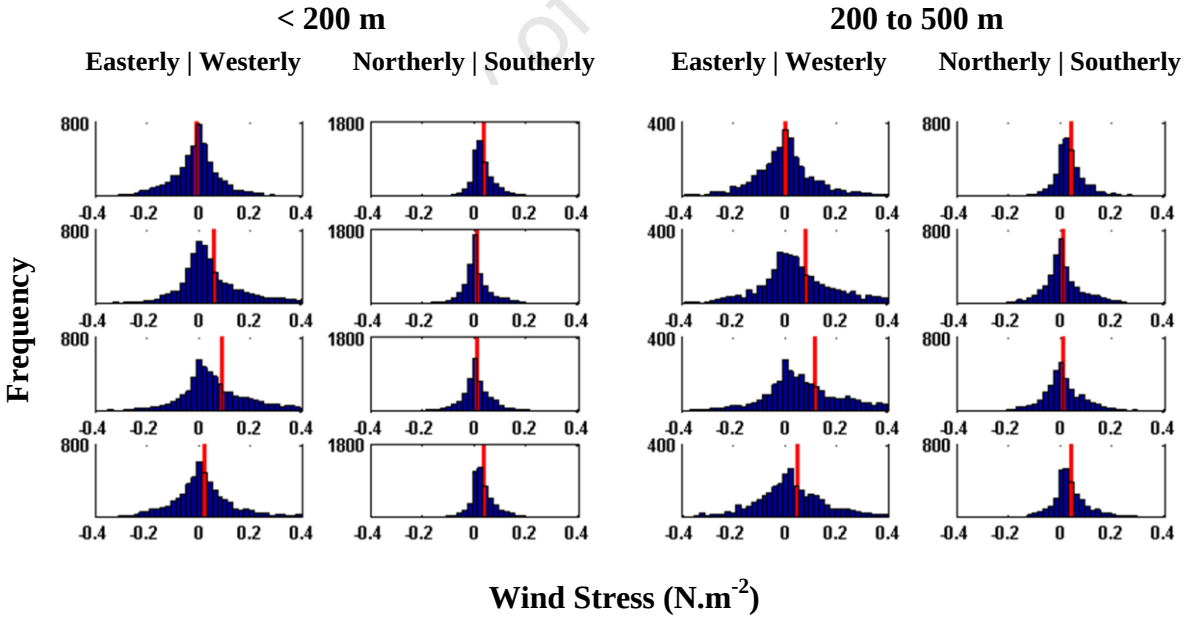


Figure 2.22 East Agulhas Bank wind stress distribution for summer (top row) to spring (bottom row) using daily sampled data from 2002-2009. The red line indicates the mean wind stress.

Table 2.4 Discrete states of the QuikSCAT wind stress vectors.

U Wind Stress ($N.m^{-2}$)	V Wind Stress ($N.m^{-2}$)	State
< -0.150	< -0.150	1
-0.150 to -0.049	-0.150 to -0.049	2
-0.050 to -0.001	-0.050 to -0.001	3
0 to 0.049	0 to 0.049	4
0.050 to 0.149	0.050 to 0.149	5
>= 0.150	>= 0.150	6

2.4 Discussion

2.4.1 *In situ* data

Profiles

The 12 temperature clusters created using the *k*-means algorithm provide a set of profiles that highlight the variability and also the unique sub-regions of the southern Benguela. The profile clusters are either specific to a region or depth, or to specific depths in different regions. In accordance with research in the regions, profiles with warmer water are found progressively offshore and profiles with either a clear shallow or deep mixing layer are found across the shelf (Nelson and Hutchings, 1983; Hutchings *et al*., 1984; Brown and Hutchings, 1987; Probyn *et al*., 1994; Weeks *et al*., 2006). However, it should be noted that no temperature profiles emerge among the 12 cluster means that clearly depict strong stabilization without a mixing layer. Strong stabilization is expected during longer periods of calm wind following upwelling. In addition, although temperature Profile 1, which has the coldest surface water, highlights the upwelling along the West Coast, it does not

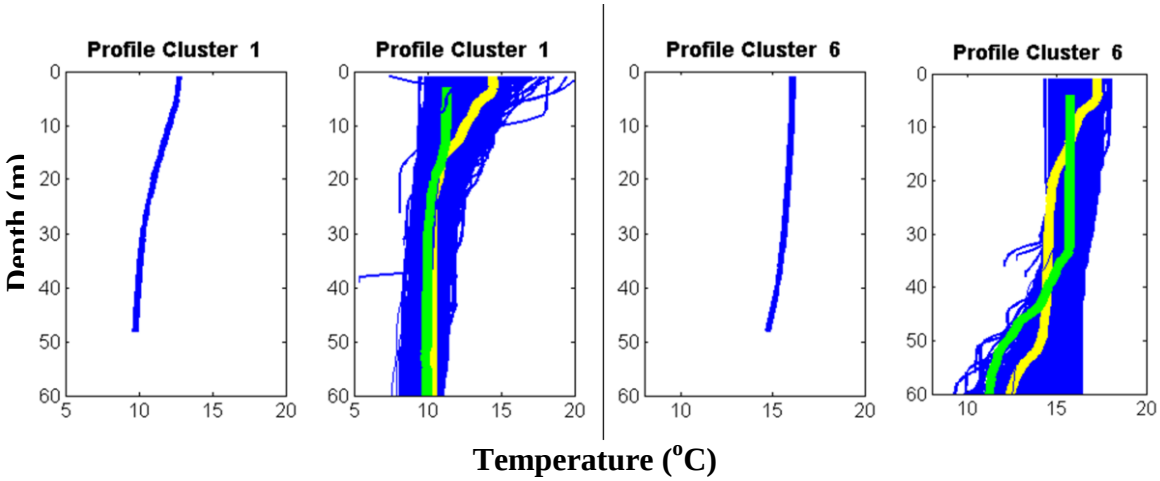


Figure 2.23 Two of the temperature profile clusters showing the cluster means and the individual profiles in the clusters. Two individual profiles are highlighted in each cluster (yellow and cyan profiles).

isolate the Cape Peninsula and Cape Columbine cells (Fig. 2.7). In Figure 2.23 the individual temperature profiles belonging to Clusters 1 and 6 (Profile 6 suggests either deep mixing or weak stabilization) are shown and two individual profiles in each cluster are highlighted. The individual profiles in each cluster illustrate a problem that can be encountered with the *k*-means unsupervised clustering algorithm; clustering is based only on Euclidean distances with no regard for specific structural features. In this case, Figure 2.23 shows that although the member profiles of each cluster have similar temperatures, their structures can be quite different. Thus, for example, Profile 1 can represent both active upwelling with near-surface temperatures ca. 10°C, and slightly more mature upwelled water where the surface layer has warmed, and is likely to be found downstream of the cells.

The Chl *a* profiles show that high biomass blooms are mostly found on the West Coast and that moderate surface and subsurface blooms do occur in regions of the West and East Agulhas Bank where upwelling has been reported (Chapman and Largier, 1989; Lutjeharms *et al*., 1989). As with the temperature profiles, the

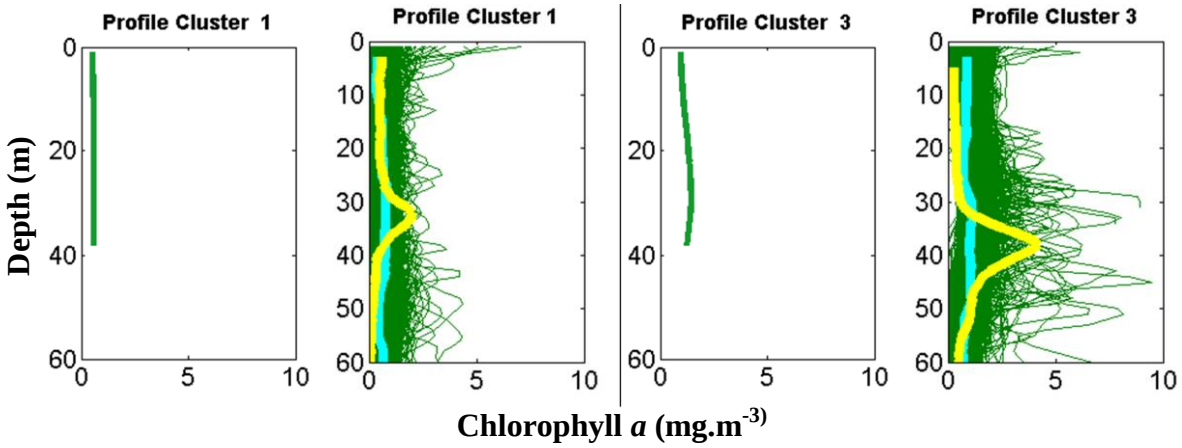


Figure 2.24 Two of the Chl *a* profile clusters showing the cluster means and the individual profiles in the clusters. Two individual profiles are highlighted in each cluster (yellow and cyan profiles).

locations of certain profiles do not seem to fit published literature as well as they should. For example, Profile 1 is assumed to indicate active upwelling water low in Chl *a* biomass and hence, its location close inshore is expected (Hutchings, 1984; Brown and Hutchings, 1987). It is also frequently located over the Agulhas Bank, which may indicate oceanic water, but here, particularly over the West Agulhas Bank, profiles should commonly depict a low subsurface peak at ca. 30-40 m (Carter *et al*., 1987; Probyn *et al*., 1994). The expected subsurface peak could be depicted by Profile 3, which suggests a deep low-biomass peak and is frequently observed over the bank (Fig. 2.8). It is also observed close inshore on the West Coast. Figure 2.24 examines the individual profiles in Clusters 1 and 3. In Cluster 1, the profiles include a homogenously low-biomass profile (cyan profile) expected to depict upwelling or oceanic water, and also a profile with a deep peak at 40-50 m (yellow profile) as expected over the Agulhas Bank. Similarly, the highlighted profiles in Cluster 3 show homogenously low biomass throughout the water column and also a deep peak at ca. 40 m. The yellow profiles and the cyan profiles in Clusters 1 and 3 seem to show similar structures although they are mapped to different clusters.

These temperature and Chl *a* clusters, and possible others, indicate that the *k*-means algorithm does have a weakness; it cannot determine some important features that can be useful for characterizing different water masses.

To define all the possible temperature and Chl *a* biomass structures will lead to many more than 12 clusters, which can result in an intractable model. Producing clusters according to specific profile structures, such as mixing layer depth or depth of the Chl *a* maximum, may improve the accuracy of the models developed in this study, particularly when the profiles are related to wind, SST and surface Chl *a*. Nevertheless, the *k*-means clustering algorithm does reproduce the broad range of structures, which can be used confidently to evaluate the coastal dynamics.

Depth

Over much of the coastal region of the southern Benguela the shelf depth delineates the surface manifestation of SST and Chl *a* (Fig. 2.3). Brown (1992) showed that there were significant differences between the inshore (0-200 m) and offshore (200-500 m) Chl a integrated to 30 m depth, for the period 1971-1989 along the southern Benguela, including the Namaqua cell to the East Agulhas Bank. The author attributes the higher inshore biomass to upwelling processes, which dominate the West Coast throughout the year but are more seasonal and hence less pronounced on the South Coast. The West Coast shows substantial changes in SST and Chl a from the coast to 500 m bottom depth (Fig. 2.9), after which both variables are more consistent.

Largier et al. (1992) identified three distinct across-shore regions on the West Agulhas Bank that functioned independently based on currents and temperature. They noted an inner-shelf (< 100 m bottom depth) dominated by wind-forcing, a mid-

shelf (100-200 m bottom depth) characterized by strong stratification, and an outer shelf (200-500 m bottom depth) dominated by oceanic forcing such as the Agulhas Current.

Probyn et al. (1994) considered the outer shelf (100-200 m bottom depth) of the East Agulhas Bank, or more specifically the shelf-edge inshore of the Agulhas Current, of particularly importance relative to the West Coast as it is an additional area of dynamic upwelling (Chapman and Largier, 1989). Although Brown (1992) indicated a significant difference between mean SST and Chl a values in the 0-200 m and 200-500 m bottom depth intervals, Figure 2.11 also highlights the distinction between the < 100 m (mean SST = 18.2, mean Chl a = 0.74) and 100-200 m (mean SST = 18.9, mean Chl a = 0.51) bottom depth regions for both SST and Chl a. These three intervals are then in accordance with the West Agulhas Bank described by Largier et al. (1992).

The analysis of eight years of daily satellite data supports the findings of many published research papers and showed that the coastal shelf can be subdivided according to bottom depth. The first 100 m bottom depth interval may be too broad to isolate the active upwelling of deeper water (for example, the locations of temperature Profile 1, when indicating upwelling, is in < 50 m bottom depth; Fig. 2.7) but satellite data (particularly QuikSCAT wind data) is poorly represented so close to the coast and having a narrower interval will not result in different associated satellite data. The 200-500 m bottom depth interval may also be inappropriate for regions where the shelf is narrowest for example, between the Cape Peninsula and Cape Columbine upwelling cells, as inshore water will quickly advect into deeper bottom depth water compared to regions where the shelf is broad. These areas are

relatively small compared to the overall shelf region and the uncertainty produced when using bottom depth information should also be small.

From published literature and supported by the results discussed above, the inshore region can be subdivided into two zones from the coast to 100 m bottom depth, and from 100-200 m bottom depth. The shelf is described by the bottom depth interval 200-500 m. In addition to these three intervals, the shelf edge at about 500-700 m is included due to the location of specific temperature and Chl a profiles (see Figs. 2.7 and 2.8). An interval for depths > 700 m is included to represent the open ocean. Each of these characteristic zones is labelled categorically where each label represents a "state" value. A summary is given in Table 2.5.

Table 2.5 Discrete interval limits of bottom depth and the assigned state value.

Depth Interval	State Value
< 100 m	1
100-199 m	2
200-499 m	3
500-699 m	4
> 700 m	5

2.4.2 Satellite data

SST and Chl *a*

The regression of the satellite SST against the *in situ* data shows a good positive correlation (r = 0.86; Fig. 2.12). The variance of the SST satellite data can be attributed to the depth from which the *in situ* measurements are taken for comparison (often 3 m below the surface) and the differences in time between the

satellite overpass and the sampled data. For example, samples were often taken along a cruise path beginning at day break and ending near sunset. The difference between the satellite Chl *a* concentration and the *in situ* data may also be attributed to the mismatch between sampling and overpass time but the daily variability of Chl *a* is expected to be small. In highly productive coastal water the MODIS algorithm is known to perform poorly relative to the open ocean (Moore *et al.,* 2009) and this probably accounts for most of the uncertainty. The results of the models developed in later chapters may be affected by this relationship. A regional algorithm using local empirical data should improve the coefficient of determination.

Optimal discrete intervals cannot be chosen for the SST and satellite Chl *a* data as their distributions are Gaussian and thus there are no characteristic signals to extract. Intervals need to be selected that can represent the variability of all the regions and depths from the available satellite data. For example, Brown and Hutchings (1987) and Mitchell-Innes *et al*. (1999) considered SST ca. 11°C to indicate newly upwelled water on the West Coast on the West Agulhas Bank respectively. However, the comparison of the *in situ* and MODIS data seldom shows temperatures lower than 13°C (Fig. 2.12). Thus, including more intervals below 13°C will produce very few labels during training of the models developed in later chapters. This may interfere with the models' ability to predict profiles representing newly upwelled water and water in an early stage of maturation. The few cases of SST < 13°C is surprising as the 13°C SST isotherm it has been suggested to indicate the offshore boundary of intense upwelling processes (Hagen *et al*., 1981; Boyd and Agenbag, 1985; Lutjeharms and Valentine, 1987). Few values < 13°C occur in the satellite data analysis (Figs. 2.3, 2.5 and 2.7) and the satellite versus *in*

situ comparison (Fig. 2.12). The infrequent samples in the satellite data analysis may occur due to the sampling strategy used (Fig. 2.2); the small area of the upwelling cells relative to the area of the shelf will result in only a few of the random samples occurring in these areas. In the satellite SST versus *in situ* data comparison (Fig. 2.12), the few cases of temperature profiles with surface values < 13°C may be decreased further by poor match-ups between the profile and concurrent satellite data, and thus the profiles are excluded from the comparison. As a result, newly upwelled water may be poorly represented in the training data. Other intervals correspond well with the literature. For example, Mitchell-Innes and Pitcher (1992) suggested SST in the range 12-15°C is indicative of rapid diatom growth and Barlow *et al*. (2005) showed that optimal conditions for high biomass growth was in the range 12-17°C. Optimal temperature for dinoflagellates appeared to be 13-16°C (Barlow *et al*. 2005). Barlow *et al*. (2005) noted water of 18-20°C extended from the south coast into the West Coast region. Weeks *et al*. (2006) observed a warm event over the inner continental shelf where SST reached 20-22°C showing that high temperatures can be attained inshore. SST > 23°C was associated with the mid-shelf on the West Agulhas Bank and particularly low Chl *a* (< 0.2 $mg.m^{-3}$; Weeks *et al*., 2006).

The regression of *in situ* Chl *a* and satellite surface Chl *a* shows a poor fit of the data (Fig. 2.12). For this reason the intervals are required to be large enough to allow for the large error. These intervals seem to fit the literature well. For example, Barlow (1982) indicated Chl *a* concentrations in recently upwelled water, maturing upwelled water and aged upwelled water are < 1, 1-20 and 5-30 $mg.m^{-3}$. Inshore, Barlow *et al*. (2005) found elevated levels of Chl *a* in the region of 2-15 $mg.m^{-3}$ and < 1 $mg.m^{-3}$

offshore. Weeks *et al*. (2006) found that concentrations < 3 $mg.m^{-3}$ coincided with the Cape Columbine and Cape Peninsula upwelling cells but concentrations < 1 $mg.m^{-3}$ corresponded with very cold upwelled water. Over the Agulhas Bank ranges of Chl *a* were 1-5 $mg.m^{-3}$ but periods of strong stratification can result in concentrations of up to 40 $mg.m^{-3}$ (Probyn *et al*., 1994). Very low Chl *a* concentrations (< 0.5 $mg.m^{-3}$) were associated with Agulhas Current water, according to expectations.

Wind

The effect of wind on the upper water column depends on its strength, direction and duration and also on the state of the water column prior to the wind event. This creates the complex problem of choosing optimal bins for the wind vector data. Strong alongshore winds on the West Coast may create an Ekman layer that is thicker than the water column depth producing net Ekman transport that is more aligned to the wind direction (Nelson, 1992). More moderate winds therefore could produce more rapid upwelling. Ideally the relationship between wind strength, duration and direction, and upwelling should be resolved to be able to indicate reasonable intervals for discretizing the wind stress data. However, the wind data should also indicate the likely structure of offshore water, which will be related to the strength of the thermocline. Therefore, analysing the wind data in terms of both inshore and offshore processes is likely to be complicated. For simplicity, similar to the satellite SST and Chl *a* data discretization, intervals are selected based on the distribution of the data. Intervals are chosen intuitively according to the frequency distributions (each interval should have sufficient observation to be competitive with other intervals) and the limits of the data. Unfortunately there is little published

literature on the effects of actual wind stress values on the upper water column with which to corroborate the intervals.

2.5 **Conclusion**

The southern Benguela has been described as three sub-regions having characteristics of eastern and western boundary systems. These sub-regions, consisting of the West Coast, West and East Agulhas Bank are characterised according to the thermal and biological dynamics of the water column over the continental shelf and slope. The coastal alignment and its interaction with atmospheric systems produce regions with unique physical and biological characteristics. Within each of these regions characteristic signals tend to be related to season and bottom depth (used here as a proxy for distance offshore) but are not necessary co-varying. In accordance with these signals, data relevant to the vertical distribution of phytoplankton are analysed with the purpose of simplifying or discretizing the signals. The discretized data are easier to incorporate into probabilistic models and are more likely to produce robust relationships.

Discretizing the data is likely to have a compounding effect on the accuracy of the model as detail is inherently smoothed over. In this section the data are explored and described in order to highlight the necessary detail to be retained while reducing complexity. Most variables are discretized while considering the affect of season, region and bottom depth. This means that before developing and testing the models developed in later chapters, relationships among the data should already exist. The results show that although the data have been simplified the full range of the data has been included and no data have been discarded.

Chapter 3 - Developing a Static Bayesian Network

3.1 Introduction

Several methods have been developed for describing non-uniform biomass profiles in the oceans (Platt *et al*., 1988; Longhurst *et al*., 1995; Sathyendranath *et al*., 1995). However, these methods tend to produce profile categories that are fixed for large spatial and temporal scales and may not be representative of the smaller scale variability in chlorophyll *a* (Chl *a*) profiles. Recently, more flexible approaches using a suite of environmental variables have been used to estimate the shape of Chl *a* profiles. Techniques such as self-organizing maps (SOMs), a type of artificial neural network particularly adept at pattern identification (Kohonen, 1997; Hewitson and Crane, 2002; Richardson *et al.,* 2003), have been used to identify limited sets of characteristic Chl *a* profiles from archives of vertical Chl *a* traces (Silulwane *et al.*, 2001; Richardson *et al.*, 2002). Such studies have had some success e.g. r^2 of parameter values ranging from 15-74% (Richardson *et al*., 2003) in predicting the subsurface shape of Chl *a* profiles from predictors that can be estimated from satellites (sea surface temperature (SST), surface Chl *a*) or are known (water column depth, season and location).

This chapter is reproduced from a published paper (Williamson *et al.,* 2011) that introduced Bayesian networks as a method to explore complex relationships among *in situ* Chl *a* profiles and other environmental variables. Therefore, the data preprocessing and analysis is somewhat different to Chapters 2 and 4. It presents a modified approach from earlier methods (Richardson *et al.,* 2003; Demarcq *et al.,* 2008) to estimate vertical distributions of Chl *a* that can be used with a simple light

model to derive primary production. In place of self–organizing maps (SOMs) and general additive and linear models (Demarcq *et al*., 2008) that assume a continuous series of profiles, *k*-means clustering and Bayesian networks (Pearl, 1985) are used; these are better suited to predicting profile categories from complex relationships. K-means clustering has the advantage over SOMs for clustering profiles because it produces a greater range of profiles; SOM nodes tend to be focused around the more frequent profiles. Bayesian networks produce an easily understandable network which encodes the relationships among the variables. Bayesian networks are directed acyclic graphs whose nodes (variables) are connected by arcs (arrows; Fig. 3.1). The direction of the arrows indicates causal relationships and encodes the conditional dependencies between the variables. Each variable has a *conditional probability table* (CPT) which is be parameterized from empirical data by counting the frequencies of the variable states for a particular configuration of its parents' states. Thus, the graph and the CPTs represent the joint probability distribution in an efficient manner (Heckerman, 1997). They are widely applied in fields as different as robotics and medical diagnosis. This approach is illustrated by applying it to an area off southern Africa that includes the dynamic upwelling of the southern Benguela region and the warm stratified conditions of the Agulhas Bank.

Two experiments are presented. First, a Bayesian network is developed to relate the potential causes (surface data, including SST and wind) to their effect (surface Chl *a* concentration). The second experiment relates an archive of shipboard vertical Chl *a* profiles to satellite derived surface variables with the aim of allocating to each satellite pixel the most probable Chl *a* profile class. These profile classes can then be used in a primary production model to produce regional estimates of integrated Chl *a* and primary production. The detailed spatial and temporal resolution of the

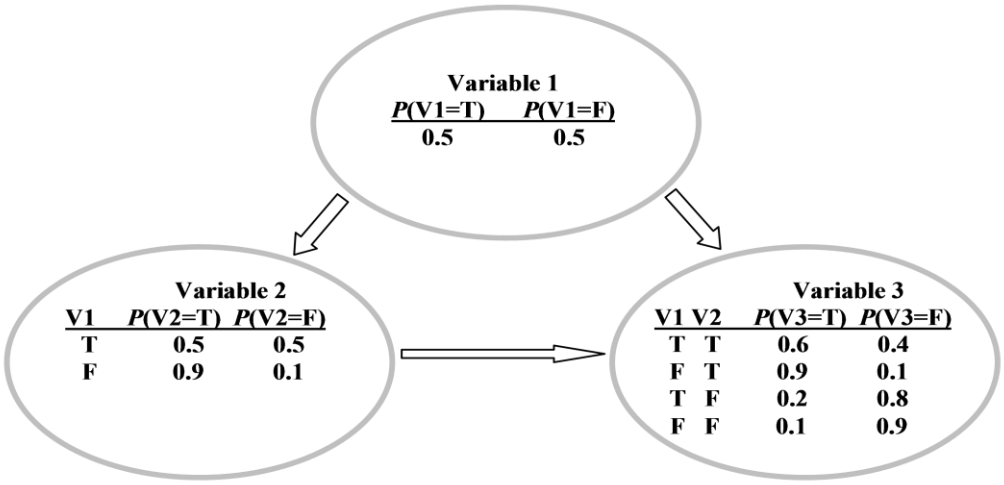


Figure 3.1 A Bayesian network (graphical model) of three variables. Each node (ellipse) represents a variable and the arrows represent the direct dependencies. The dependencies of each variable (model parameters) are encoded in the conditional probability tables.

primary production estimates will be useful for analysing long-term ecosystem-scale changes e.g. regime shifts.

3.2 Data and methods

3.2.1 Remote sensing data

Archives of daily weighted averages of SST and sea surface Chl *a* concentration are obtained from the *www.rsmarinesa.org.za* website. The website produces three-day weighted composites (day-1 weight=1, day weight=3, day+1 weight=1) from MODIS level 2 data available from the NASA Oceancolor web portal. The daily composites are produced at 1 km^2 resolution. Composites are used as they fill in much of the missing data due to frequent cloud cover over the region.

Surface wind data consists of the Blended Sea Winds product from the National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Centre.

The data contain globally-gridded, 25 km^2 resolution surface vector winds at daily time resolution. Blended multiple-satellite observations fill in temporal and spatial data gaps and compare well with *in situ* observations (Zhang *et al*., 2006; Bentamy *et al*., 2009). Depth was estimated from the latitude and longitude of each profile using data provided by the US National Geophysical Data Centre.

3.2.2 Discretizing continuous data

Bayesian *learning*[1] and inference with continuous data can be simplified by reducing the data into an appropriate number of discrete intervals. The four satellite-derived surface variables (U and V wind stress, SST and surface Chl *a*; CHL) and $Season$, $Region$ and $Depth$ are arranged into 4-6 discrete categories for each variable (Table 3.1), based on similar frequencies within each category, but modified by our understanding of the data and processes involved. Four seasons were defined with a one month lag from conventional seasons (e.g. austral summer: January-March), because of the lag in ocean response to atmospheric forcing.

[1] Learning refers to the process where algorithms are trained on data and are then able to classify or infer from new data

Table 3.1 Discrete states of the variables used in the Bayesian network learning. The upper and lower bounds of the states were obtained from the data range and current understanding of the relationships between the variables and upwelling processes leading to primary production.

Season	*Region*	*Depth* (m)	*U/V Wind* ($N.m^{-2}$)	*SST* (°C)	*CHL* ($mg.m^{-3}$)
Summer	**North of Orange R.**	**< 100**	**< -0.15**	**< 12.5**	**< 0.5**
Autumn	**Namaqua Cell**	**100 to 199**	**-0.15 to -0.049**	**12.5 to 15.49**	**0.5 to 0.9**
Winter	**St. Helena Bay**	**200 to 499**	**-0.05 to -0.001**	**15.5 to 16.99**	**1 to 2.9**
Spring	**Table Bay**	**500 to 699**	**0 to 0.049**	**17 to 18.99**	**3 to 9.9**
	West Agulhas B.	**>= 700**	**0.05 to 0.149**	**> = 19**	**> = 10**
	East Agulhas B.		**>= 0.15**		

Five sub-regions are defined (Fig. 3.2), based on their physical oceanographic characteristics with the sixth sub-region being outside these five. The West Coast has several patterns of seasonal upwelling from north to south that includes the Namaqua upwelling cell, the Cape Columbine upwelling cell in the St Helena Bay sub-region and the Cape Peninsula upwelling cell in the Table Bay sub-region. The West Agulhas Bank has some upwelling and a persistent deep thermocline, and the East Agulhas Bank has a shallow thermocline. *Depth* is used to indicate each profile's position relative to the shore and continental shelf; five depth states relate to inshore, inner-shelf, outer shelf, continental slope and open ocean. Wind speed is arranged as U (east-west) and V (north-south) vectors, each with six states ranging from strong in one direction to weak, to strong in the opposite direction (on the West Coast positive U is onshore and favourable to downwelling while positive V is

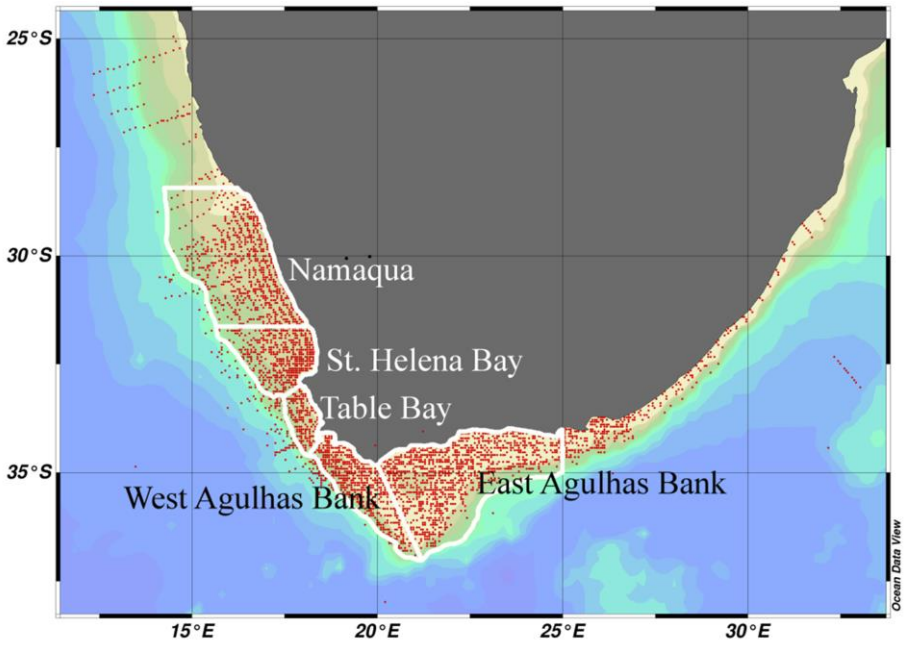


Figure 3.2 Distribution around southern Africa of the ~7300 Chl *a* profiles used in the study. Five sub-regions, characterised by local hydrographic dynamics, are demarcated by white lines and a sixth region is considered outside the white lines.

longshore and favourable to upwelling). *SST* and surface Chl *a* (*CHL*) concentration have five states each.

3.2.3 Identifying characteristic profiles

The second experiment relates a large number of highly variable Chl *a* profiles to the satellite surface variables. *K*-means clustering is used to group all vertical Chl *a* profiles into a small number of typical groups (clusters) having less variability within than among clusters. A total of six clusters is chosen to limit the size of the CPT for the *Profile* node and hence, reduce the size of the model. Each cluster is represented by the mean of the profiles in the cluster. Individual profiles are first smoothed using a three-point running mean and then interpolated at 1 m intervals from 0-60 m depth. Clustering is performed on all available smoothed profiles

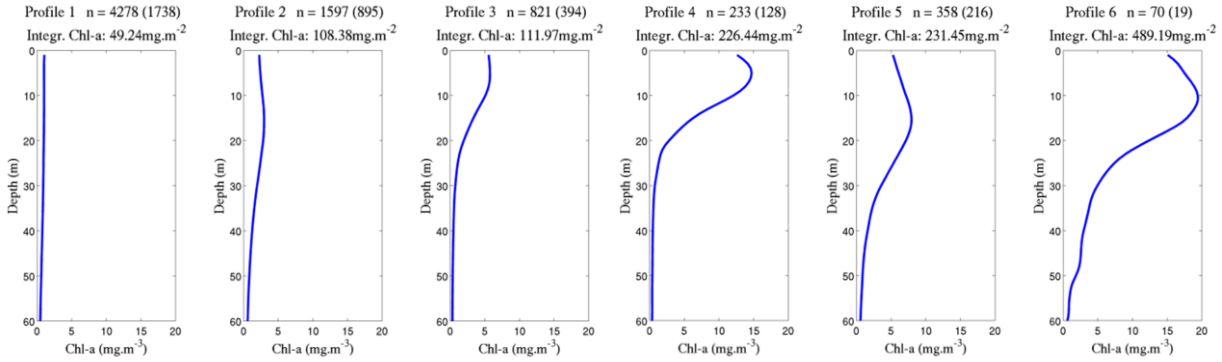


Figure 3.3 Six mean Chl *a* profiles resulting from *k*-means clustering. The integrated concentration of Chl *a* and the number of raw profiles (n) included in each cluster are shown (total = 7357). Numbers in brackets refer to the number of profiles that have corresponding SST and surface Chl *a* data.

collected along the southern African coastline, with no assumptions regarding their shape. For profiles in < 60 m bottom depth, a missing data code is used for intervals to 60 m to equalize the number of input rows across the dataset. Thus, input data consists of the Chl *a* at 60 1 m depth intervals (columns) by the 7357 profiles (rows). Clustering is performed using the Statistics Toolbox for MATLAB® vR2008a software (The MathWorks, Inc).

Figure 3.3 shows the six profile cluster means obtained from 7357 Chl *a* profiles collected along the southern African coast. The procedure has arranged the Chl *a* profiles according to natural patterns in the data, ranging from Profile 1 with little vertical structure, low surface Chl *a* and low integrated Chl *a*, to Profile 6 with a clear subsurface peak and high surface and water column integrated Chl *a*, with a gradation of clusters between them. The profiles are arranged in sequence from lowest to highest integrated Chl *a* concentrations. Some of the characteristic profiles have highest Chl *a* values near the surface, while others have sub-surface peaks. The number of profiles within each cluster differs by two orders of magnitude with a

Table 3.2 Mean surface Chl *a* concentration and integrated Chl *a* values for each profile cluster.

	Profile 01	Profile 02	Profile 03	Profile 04	Profile 05	Profile 06
Surf. Chl–*a* ($mg.m^{-3}$)	**1.0**	**2.3**	**5.5**	**12.6**	**5.2**	**15.1**
Integr. Chl–*a* ($mg.m^{-2}$)	**49.2**	**108.4**	**112.0**	**226.4**	**231.5**	**489.2**

maximum of 4278 in Cluster 1 and a minimum of 70 in Cluster 6. Table 3.2 gives the surface and integrated Chl *a* values of each characteristic profile.

3.2.4 Experiment 1: Predicting surface Chl *a* from satellite data

In the first experiment a Bayesian network is automatically generated or *learned* using only satellite surface data coupled with information on season, region and depth. The Bayesian Network Structure Learning toolbox for Matlab is used to extract the relationships among variables. Bayesian variables or *nodes* are pre-arranged into five logical levels according to our understanding of the cause-effect relationship among variables. The independent variables, *Season*, *Region* and *Depth* are assigned to level 1 and the "predicted" variables, *SST* and surface Chl *a* (*CHL*) concentration, are assigned to level 5. Arrows between the nodes represent dependent relationships as derived by the network learning algorithm. The algorithm searches all possible graph structures and returns one with the maximum likelihood given the data, according to Bayes' theorem. The data represents about 20 million pixels from one year of daily observations. Figure 3.4 shows the network derived from the training dataset.

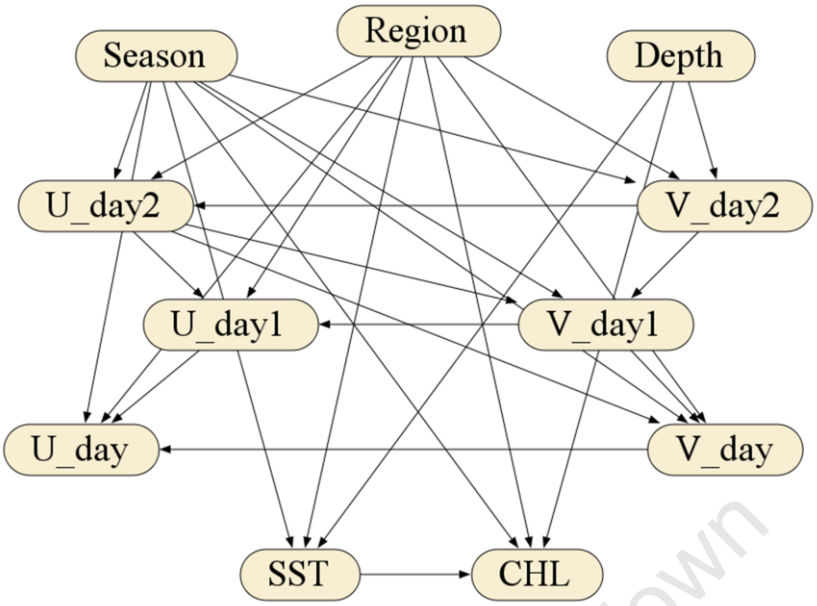


Figure 3.4 Network structure *learned* from the satellite data by the Bayesian Network Structure Learning Toolbox. Arrows connecting variables indicate dependent relationships; their absence indicates conditional independence. U and V are the wind vectors for 2 days ("*_day2*") and 1 day ("*_day1*") prior and the same day ("*_day*") as the Chl *a* data. Level 1 is at the top of the diagram and level 5 at the bottom (*SST* and *CHL*).

In this experiment, the importance of wind vectors one day prior and two days prior to the day of observation is examined, giving a total of eleven variables in the network. The network structure indicates that *Season* and *Region* are important factors in understanding most of the ocean and atmospheric variables, i.e. if we know the season and region we can infer something about the probable wind, SST and surface Chl *a*. Only the U wind vector on the day of observed SST has a direct connection to *SST* and no wind vectors are directly related to surface Chl *a* (*CHL*). In addition to *Season* and *Region*, *CHL* is directly dependent on *SST* and *Depth*. After deriving a network structure, Netica v4.16 software (Norsys Software Corp.,

Vancouver, BC.) is used to parameterize it with a subset of the data, and to calculate conditional probabilities for each node state. The network is then evaluated by testing its ability to "predict" the most likely surface Chl *a* (CHL) concentration from a subset of the data excluded from training.

3.2.5 Experiment 2: Relating vertical profiles to surface variables

In the second experiment the dataset is comprises individual profiles and their environmental data. Although Bayesian network algorithms are capable of handling missing data, for simplicity this network was developed using only fully observed data, i.e. each profile has a full set of variable values. Chl *a* profiles are first associated with other environmental variables according to time and space co-ordinates. The total 3390 profiles that remain that are used to parameterize and test the network. The data are divided into two subsets, one for training (parameterizing), the second for testing the accuracy of predicting the Chl *a* $Profiles$ 1-6. The Bayesian relationships are again parameterized using the *Netica* software.

Whereas in the first experiment the network structure was learned from the data and then parameterized, in this experiment the structure was developed from our knowledge of the ocean processes before parameterizing. In the southern Benguela the coastline is roughly aligned either north-south or east-west. The longshore wind (V wind vector) drives upwelling on the West Coast, whereas strong westerly winds (U wind vector) cause deep mixing over the Agulhas Bank sub-regions. In addition, the width of the shelf varies substantially among sub-regions. The Chl *a* profiles change due to physical and biological processes as the surface water mass moves offshore over the shelf during upwelling. Hence, wind, region and depth (a proxy for distance offshore) are considered to be direct causes of profile shape. Surface Chl

a, although unreliable as a profile predictor on its own, nevertheless adds valuable information on water column characteristics. The resulting network is shown in Figure 3.5. The ability of the Bayesian approach to predict profile shape is tested on a subset of data excluded from that used to train the network.

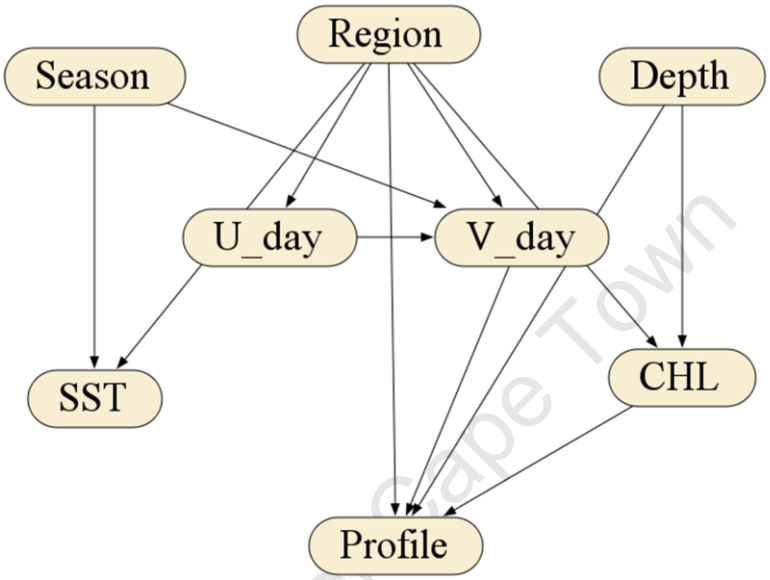


Figure 3.5 Bayesian network constructed from our understanding of the relationships between surface variables and subsurface Chl *a* profiles. The north-south wind prior to the observed profile (*V_day*), *Region*, *Depth* and surface Chl *a* (*CHL*) are indicated by arrows as having a direct influence on profile shape (*Profile*).

3.3 **Results**

3.3.1 Experiment 1: Predicting surface Chl *a* from satellite data

In the first experiment the Bayesian network is constructed automatically. The network parameters are obtained from a random subset of the empirical data and the remaining subset is used to test the model on predicting the surface Chl *a* state. The predictive skill is illustrated by a confusion matrix (Table 3.3). The matrices have the

correct or observed state on the y-axis and the predicted state on the x-axis, therefore the correctly predicted states fall along the main diagonal. Hence, the predicted distribution is indicated along the rows. *CHL* State 1 (< 0.5 $mg.m^{-3}$) is most accurately predicted (80% correct) followed by State 3 (1-3 $mg.m^{-3}$) at 63% correct. States 2, 4 and 5 (0.5-1 $mg.m^{-3}$, 3-10 $mg.m^{-3}$ and > 10 $mg.m^{-3}$, respectively) are correctly predicted ca. 50% of the time. The average predicted correct is 61%. If we include the adjacent state with the correct state then 95% of the test data are either correctly categorized or categorized adjacent to the correct state.

Table 3.3 Analysis of the accuracy of the network for "predicting" the surface Chl *a* state (column 6) from test data (n= ~10^7). Row values indicate the percentage distribution of "predictions"; shaded diagonal cells indicate the percent correct "predictions" of the observed state. The overall error rate was 36%.

Predicted states (%)					Observed
State 1	State 2	State 3	State 4	State 5	
80	**15**	**4**	**0**	**0**	**State 1**
22	**55**	**21**	**2**	**0**	**State 2**
6	**16**	**63**	**13**	**2**	**State 3**
2	**3**	**3**	**54**	**12**	**State 4**
2	**1**	**1**	**33**	**54**	**State 5**

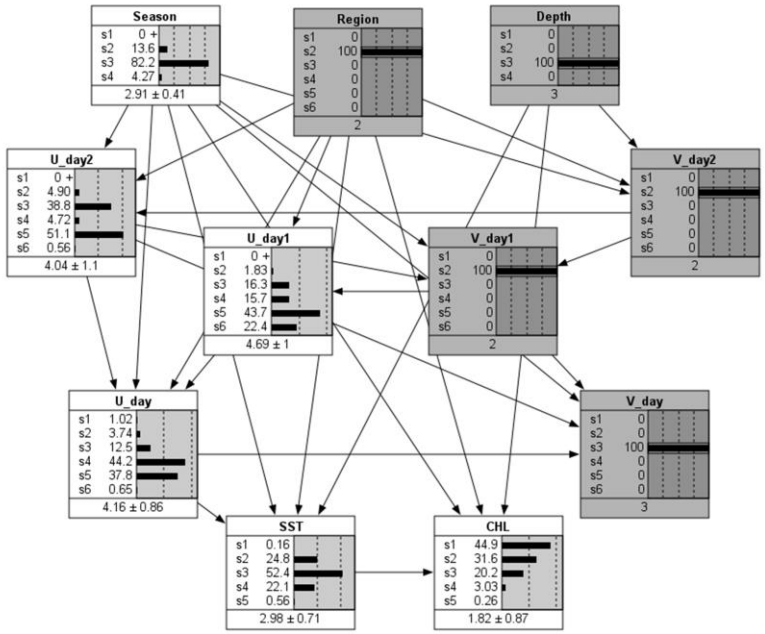
To illustrate how the Bayesian network can be used in an oceanographic context, wind conditions were specified in a particular sub-region and depth, and the network was used to predict the other associated surface variables. Figure 3.6a shows the most likely variable states for the Namaqua sub-region (*Region* State 2) at 200-500 m

depths (*Depth* State 3) after three consecutive days of northerly wind. These conditions are most likely to occur when *Season* is in State 3 (winter), *SST* is in State 3 (15.5-17 ^{o}C) and *CHL* is in State 1 (< 0.5 mg m^{-3}). Figure 3.6b depicts the likely situation after three consecutive days of strong southerly wind at the same location. This is most likely to occur with *Season* in State 1 or 4 (summer or spring), *SST* predominantly in State 3 or 4 (15.5-19^{o}C) and *CHL* in State 3 (1-3 mg m^{-3}).

3.3.2 Experiment 2: Relating vertical profiles to surface variables

In the second experiment the network was constructed by intuitive understanding of dependencies among the variables including the *in situ* Chl *a* profiles. The model is tested on the predictive accuracy of Chl *a* profile category. Overall an average of 22% correct predictions are achieved (Table 3.4). *Profile* 1 is predicted most accurately (88%), followed by *Profile* 2 (16%), while *Profile* 6 is not predicted from the testing data. *Profile* 1 is predicted the most although the percentage of predictions declines steadily for *Profiles* 2-6.

(a)



(b)

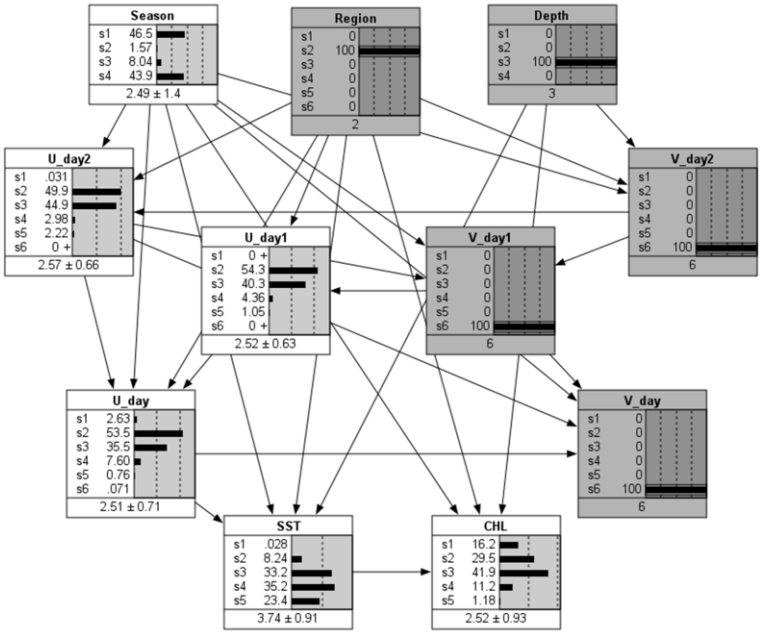


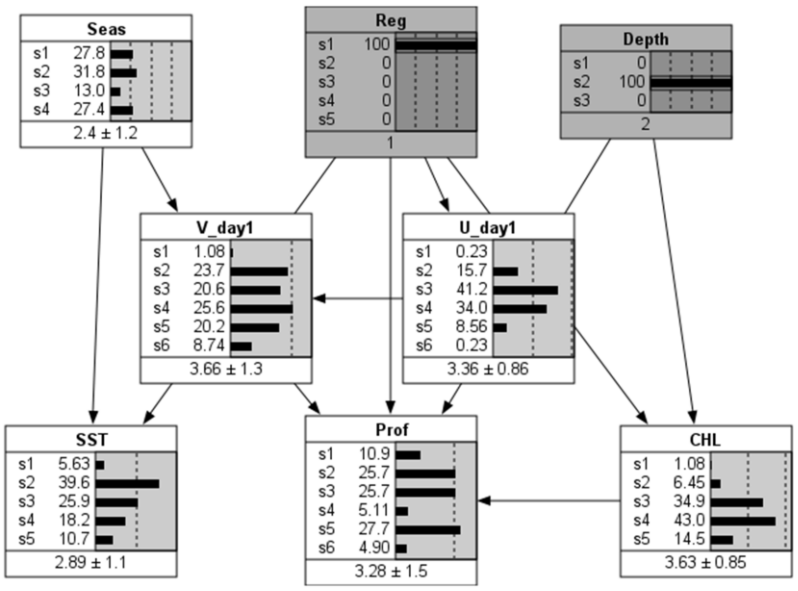
Figure 3.6 Comparison of the effects of north and south wind vector states for *Region*=2 (Namaqua cell) and *Depth*=3 (outer-shelf) on the remaining variables; 3 consecutive days of (a) moderate northerly wind and (b) strong southerly wind. The specified variable states (100%) and the % probabilities of the states of other variables are indicated in the histograms.

Table 3.4 Analysis of network accuracy in predicting a profile from test data (n=565). Row values indicate the percentage distribution of predictions; shaded diagonal cells indicate the percent correct predictions of the observed state. The overall error rate was 47%.

Predicted profiles (%)						Observed
Profile 01	**Profile 02**	**Profile 03**	**Profile 04**	**Profile 05**	**Profile 06**	
88	**8**	**2**	**1**	**1**	**0**	**Profile 01**
76	**16**	**4**	**1**	**3**	**0**	**Profile 02**
63	**18**	**8**	**8**	**3**	**0**	**Profile 03**
61	**6**	**16**	**12**	**5**	**0**	**Profile 04**
57	**19**	**8**	**10**	**6**	**0**	**Profile 05**
44	**17**	**5**	**28**	**6**	**0**	**Profile 06**

The network can also be used to predict unknown variables when only some are observed. For example, in Figure 3.7a, if we specify *Region* as State 1 (Namaqua cell) and *Depth* as State 2 (100-200 m) it can be inferred with high probability that surface *CHL* will be in State 3 or 4 and the Profile will be in States 2, 3 or 5. However, if *Region* is changed to State 4 (West Agulhas Bank) at the same depth range the most likely surface Chl *a* (*CHL*) concentration is < 0.5 mg.m^{-3} and *Profile* State is 1 (Fig. 3.7b). The network can also be used as a diagnostic tool. For example, Figure 3.8a shows that *Profile* State 6 and *Region* State 2 (St Helena Bay) are probably found on the inner-shelf, as the result of weak south–westerly winds. The same profile in *Region* State 5 (East Agulhas Bank) is most probably on the outer shelf and the result of strong south-easterly winds (Fig. 3.8b).

(a)



(b)

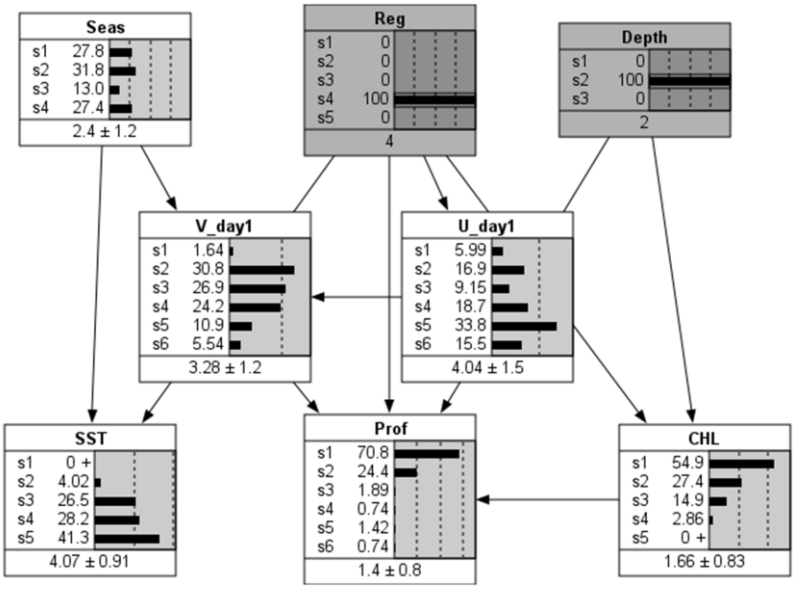
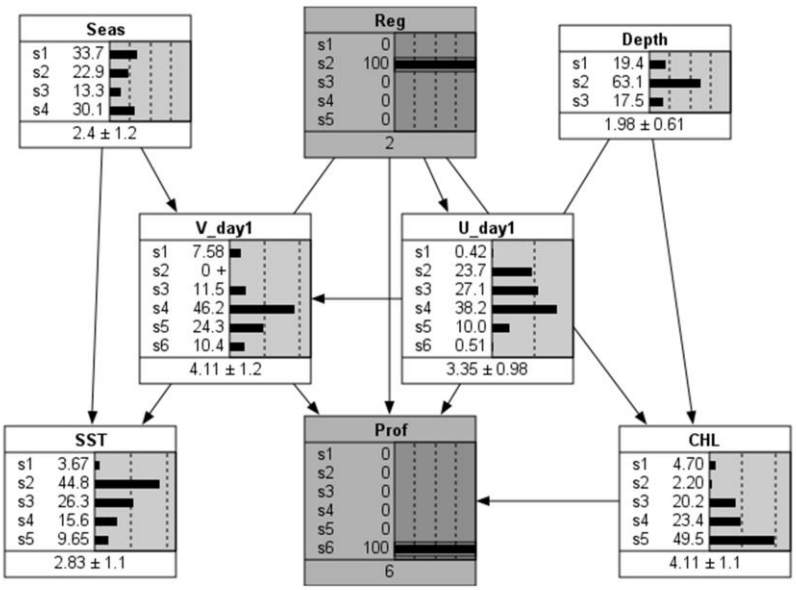


Figure 3.7 Example analyses using the parameterized network of surface variables and sub–surface profiles. Some variable states are specified (100% values in the histograms) and the percent probabilities of their influence on dependent variables compared: (a) *Reg*=1, *Depth*=2, compared to (b) *Reg*=4, *Depth*=2.

(a)



(b)

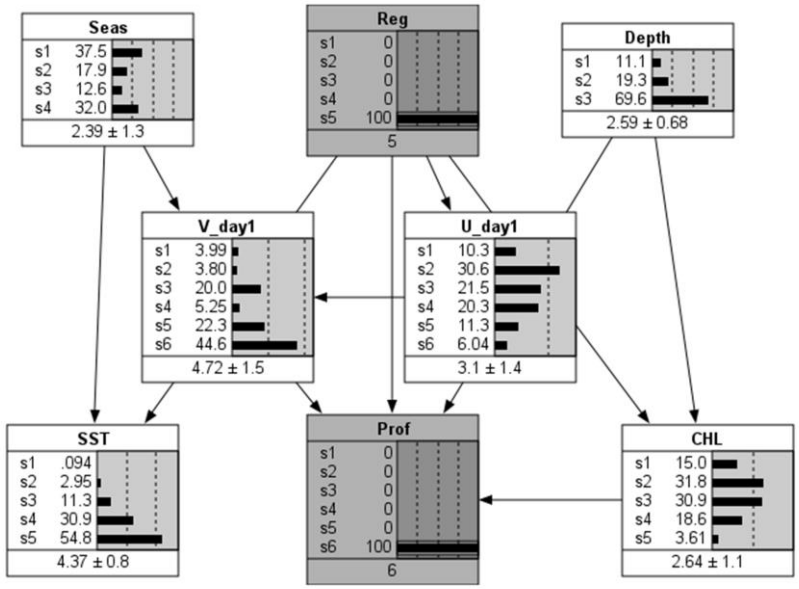


Figure 3.8 Example analyses using the parameterized network of surface variables and subsurface profiles. Some variable states are specified (100% values in the histograms) and the percent probabilities of their influence on dependent variables compared: (a) *Prof*=6, *Reg*=2 compared to (b) *Prof*=6, *Reg*=5.

3.4 **Discussion**

This chapter illustrates the ability of Bayesian models to extract patterns from large data sets and represent them as easily understood graphic networks. The derived network can then be parameterized using one subset of data and its predictive accuracy tested on another. In the first experiment, in which the structure was learned from a large set of satellite and positional data only (Fig. 3.2), the model scored an error rate of 36%. While this may seem high, Table 3.2 indicates that the majority of incorrect predictions were made on states adjacent to the correct one. This experiment also illustrates the method's ability to characterise the most likely region, season and conditions for particular SSTs and Chl *a* concentrations.

In the second experiment subsurface Chl *a* profiles were predicted from a much smaller set of surface satellite data in which three successive days of wind data and same-day SST and Chl *a* data coincided with the shipboard profile. Here the network structure was developed from the current understanding of the processes involved and a subset of observations was used to parameterize the Bayesian model. This model was then tested on the remaining subset of data. The error rate for predicting Chl *a* profiles was 47%, largely due to incorrectly predicting *Profile* 1 when *Profiles* 2-6 were observed (Table 3.4). *Profile* 6 was never predicted; the network did not capture causal conditions specific to this profile due to its relatively rare occurrence. Such issues may be resolved by including temperature profiles in the network, which may result in more robust relationships with causal variables and improved predictions.

The diagnostic power of Bayesian networks was explored by specifying a profile type and examining probable causes (Fig. 3.8a,b). For example, it can be deduced that

weak southerly winds over the St Helena Bay sub-region (Fig. 3.7a) facilitate stratification and subsequent phytoplankton blooms in the bay. In the East Agulhas Bank sub-region (Fig. 3.7b) strong winds are needed to break down the persistent shallow thermocline and introduce new nutrients into the upper layers. This is more likely to occur where the warm Agulhas Current (indicated by warmer SST) impinges on the shelf edge.

This approach has potential for improving remotely-sensed phytoplankton biomass and production estimates and understanding the spatial and temporal variability of dynamic regions at appropriate scales. Only a few variables known to influence the development and decline of phytoplankton blooms are used. Other potentially useful variables include photosynthetically active radiation (PAR) and sea surface height (SSH). Future work could include development of dynamic Bayesian networks which incorporate observational time sequences preceding the profile observation itself. The predicted Chl *a* profiles can then be used with a light algorithm, such as those of Platt *et al*. (1988) or Kyewalyanga *et al*. (1992), to estimate primary production at the same time intervals. The approach here is similar in principle to that of Demarcq *et al*. (2008) but has the advantage of not assuming a continuous distribution of profile types. Further, wind data are included because wind is the primary forcing of the physical and biological processes. The good spatial and temporal resolution and robust estimates provided by this approach may be useful in the analysis and understanding of ecosystem-scale changes such as regime shifts (van der Lingen *et al*., 2006; Jarre *et al*., 2006).

Chapter 4 - Classification Using Time Series Data

4.1 Introduction

In the previous chapter information on the prior evolution of processes that drive or describe upwelling and bloom development were omitted. For this reason the predicted profile was based on limited information. The short-term wind-driven upwelling of the Benguela system occurs at intervals of several days and it is intuitive to use information on the changing conditions prior to the day of the predicted profile. The problem then requires a model that can classify a sequence of information according to a set of profile classes. Whereas the clustering of the profiles into several classes using the *k*-means algorithm is an unsupervised learning approach, the class labels of each profile and the associated sequence of events can be used in a supervised manner. Thus, a supervised training method can be used to determine or *discover* the relationship between the profile class and the changing environment.

Before deciding on an appropriate method some of the fundamental underlying processes need to be considered. The general response of the sea surface temperature (SST) and chlorophyll *a* (Chl *a*) surface variables to wind forcing, irrespective of their actual values, is expected to be consistent because the wind is mainly responsible for mixing or stratification of the surface layer. For example, cooling SST might be expected across the region during the onset of strong equatorward winds, signifying the start of an upwelling event inshore or surface layer mixing further offshore. There may be no change in SST if the wind remains constant for a long enough period of time during which the mixing depth reaches a

thermocline. For phytoplankton there will probably be an increase in surface Chl *a* concentration over the shelf at the onset of quiescent periods as the surface layer begins to stratify and phytoplankton growth increases. A decline might occur near the end of the quiescent phase as the algae sink. The sequential regularities suggest that there are likely to be reoccurring patterns in the surface data that can be related to the sub-surface structures. In addition to these relationships that are dependent on wind, advective processes should also be considered, as regularities in the data describing surface layer processes might also include intrusions of water masses that are not driven by wind. For instance, any sequence being considered is taken at a point that is considered to be static, and may be a series of daily snapshots of a water mass moving across this location. At any time an adjacent water mass with different surface characteristics may move through the location. Some sequences are therefore expected to have abrupt changes if the water mass changes. These observations of the processes imply that there will be relevant information in shorter sequences closer to the event (profile record) and also longer sequences. For example, a six-day sequence may consist of three days of relatively low Chl *a* followed by three days of high but declining Chl *a* signifying that a bloom filament with a high concentration of Chl *a* has advected into the area. Therefore a shorter sequence may be associated with a profile with more certainty than the complete sequence. The primary goal is to discover recurring sequences of varying length that can be used to classify new sequences according to the physical and biological sub-surface structures.

Algorithms designed to recover recurring sequences (or patterns) or can be classified as those that search for patterns in a single sequence (for example, DNA strands or streaming data) and those that search among multiple sequences. The

problem here falls into the latter category. Agrawal and Srikant (1994) presented their *apriori* algorithm designed to discover association rules among sales items (I) in very large databases (D) where the association rule is in the general form $M \Rightarrow N$ where $M \subseteq I$ and $N \subseteq I$. For example, a customer who buys items M also buys items N. They extended their work (Agrawal and Srikant, 1995) to include sequential patterns in the transaction database. For example, a customer who buys a set of items m_j is likely to buy the set of items n_k at a later stage, where j and k are arbitrary indices to I or subsets of I. The authors provide a generalized form of their algorithm in Srikant and Agrawal (1996) termed *generalized sequential patterns* (GSP) that generates or *discovers* all sequential patterns that satisfy a minimum user-specified support. The support is simply the number of times a pattern occurs in the database. The algorithm makes multiple passes over the data. During the first pass the support of each individual item is recorded, where an item is an attribute-value pair. For example, the attribute may be SST and its particular state (or discrete temperature interval) in the sequence may have the value 3 and is written as (SST, 3). Those items that have minimum support are then used to generate a candidate set of sequences of its length plus one. So if the item (SST, 3) is frequent it will be used to generate all possible two-length sequences starting with that item; [(SST, 3), (SST, 1)], [(SST, 3), (SST, 2)] etc. The process is repeated until all sequences that satisfy the minimum support are discovered.

4.2 **Pre-processing wind data**

In Chapter 2 data were pre-processed by considering the variability of the data taken from single day observations. In this chapter time sequences of data are

considered. For SST and Chl *a* data the time sequence is simply a sequence of individual one-day states. For example, a set of SST states {4, 4, 3, 2, 2, 1} indicates cooling surface water over six consecutive days. Seven states were considered sufficient to capture the variability of SST and Chl *a* in the southern Benguela (see Table 2.4). In Chapter 2 wind stress is considered as having separate u and v vectors and each is discretized manually. There is little published data on the effect of absolute wind stress values on subsurface structures so the intervals are somewhat arbitrary and are selected to indicate low, moderate, strong and very strong wind stress. Here, an automated approach is taken to extract typical wind scenarios. The reason for this is that combining all possible sequences from the six states chosen in Chapter 2 (see Table 2.5), leads to 216 possible three-day wind scenarios. Many of these wind scenarios may be redundant. The *k*-means approach has been shown to produce clusters that cover the range of the data input (see Section 2.4.1). The method is applied to one- to three-day wind vector sequences. Different numbers of states were tested on the one, two and three day sequences to capture the range of data efficiently. A selection of 10, 12 and 16 classes were chosen respectively. The results of the clusters mean vectors are shown in Figures 4.1-4.3.

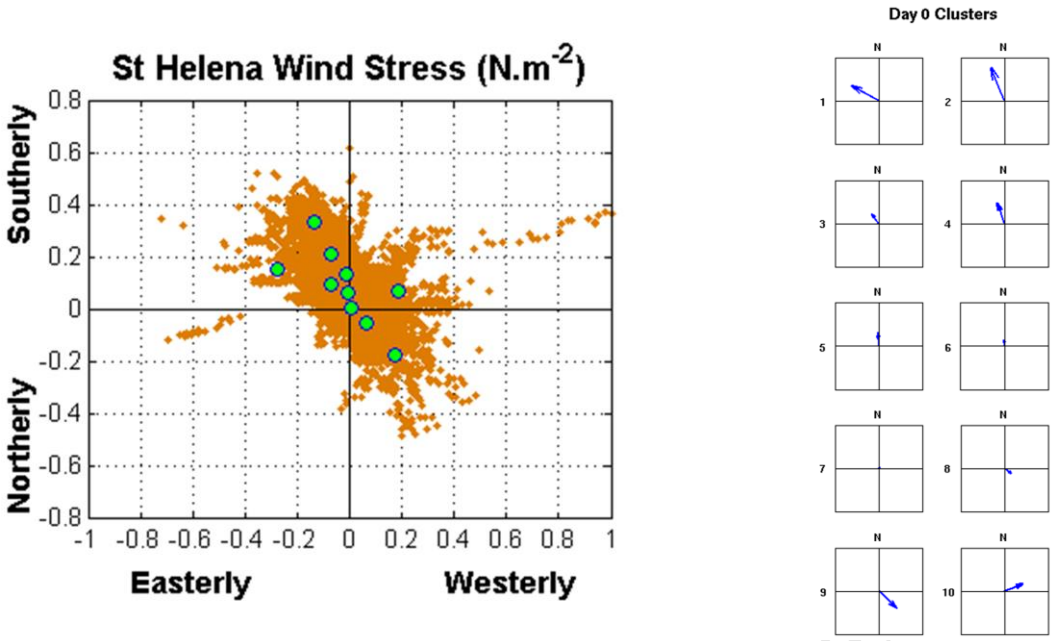


Figure 4.1 The 10 *k*-means wind stress clusters for a single day. The left figure shows the cluster means (green dots) plotted over the daily QuikSCAT winds (orange dots). The right figure shows the individual cluster mean wind vectors.

The results show the most common synoptic wind scenarios. For example, the wind vectors for one-day wind (Fig. 4.1) are mostly southerly winds ranging from calm near-zero wind stress (Cluster 7) to moderate southerlies (Cluster 4) to very strong east-south-easterly and south-south-easterly wind stress (Clusters 1 and 2 respectively). Two and one cluster is selected for north-westerly and south-westerly wind respectively. The southerly and north-westerly winds tend to dominate the West Coast (Andrews and Hutchings, 1980), where the north-westerly winds occur mostly in winter in association with frontal systems. The north-westerly winds are frequently followed by south-westerly wind (Preston-Whyte and Tyson, 1988).

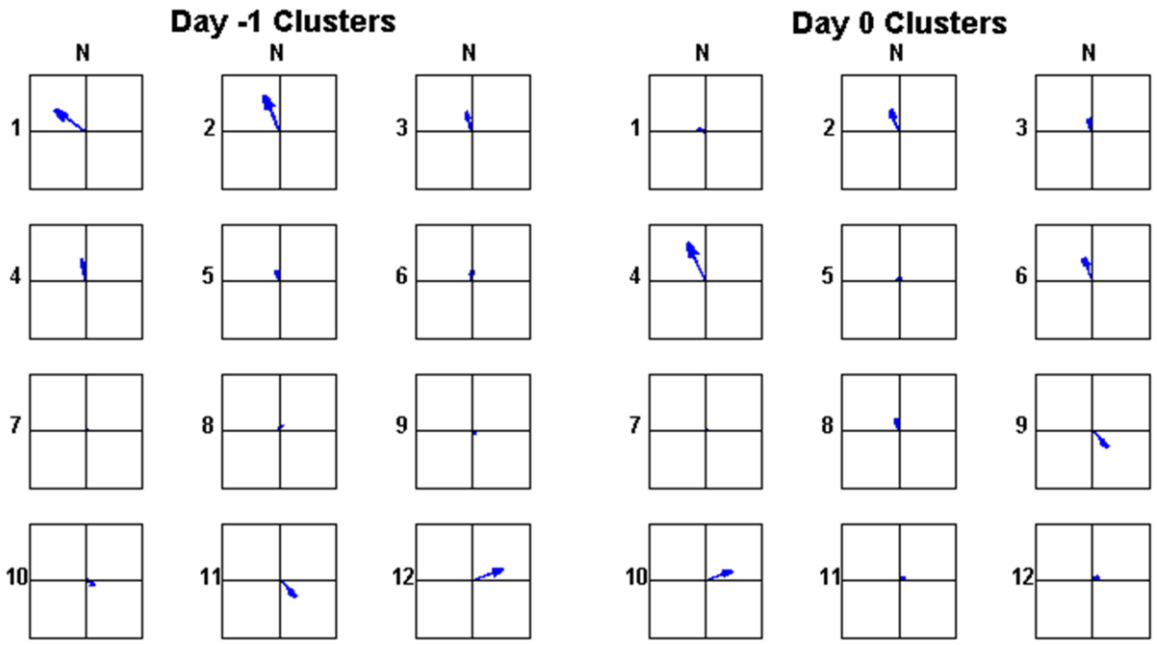


Figure 4.2 The 12 *k*-means cluster mean wind vectors for two consecutive days. For example, Cluster 1 indicates a strong south-easterly wind followed by a weak south-easterly wind.

The two-day wind clusters (Fig. 4.2) show strong south-easterly wind followed by moderate south-easterly wind (Cluster 1), and also moderate southerly wind preceding strong south-easterly wind (Cluster 4). Moderate south-westerly wind follows weak north-westerly (Cluster 10) and moderate north-westerly wind precedes weak south-easterly wind (Cluster 11). There is also a two-day period of calm (Cluster 7).

Figure 4.3 shows the 16 clusters for three-consecutive days of wind. There are different scenarios not seen with the two-day clusters. For example, two consecutive days of moderate to strong south-easterly winds are shown in the three consecutive days of moderate to strong south-easterly wind (Cluster 2). Further, strong south-westerly wind is seen following strong north-westerly wind prior to weak south-

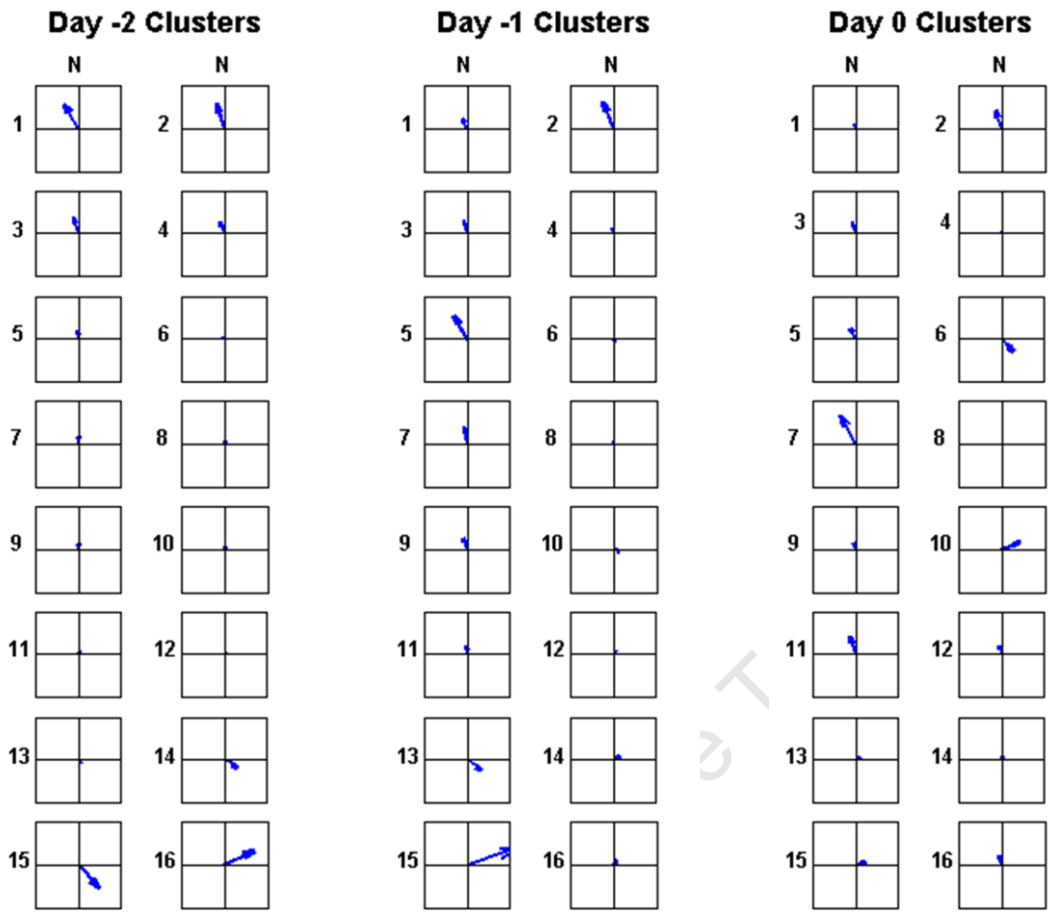


Figure 4.3 The 16 *k*-means cluster mean wind vectors for three consecutive days. For example, Cluster 1 indicates a strong south-easterly wind followed by a moderate south-easterly followed by a weak south-easterly wind.

westerly wind (Cluster 15). There are also sequences of increasing, decreasing or constant south-easterly winds shown in Clusters 7, 1 and 3 respectively, and three-day periods of calm (Cluster 8).

The resulting *k*-means clusters reduces the dimension of the wind data but still show a range of synoptic wind scenarios that can be used for predicting physical and biological changes in the surface waters. Whereas in Chapter 2 wind interval limits

are somewhat arbitrarily selected, here only wind patterns that are actually observed are represented.

4.3 Classification with sequences

4.3.1 Developing a classifier

The goal of classification is to develop a model that can associate items, sets of items or sequences of items, of new data observations to the most likely class label. A classifier is used to weight evidence that supports the associates of the item with the target class so that a new observation can be classified. Liu *et al*. (1998) integrated two data-mining techniques to produce a classifier they showed to be more accurate that the highly rated C4.5 (Quinlan, 1993) classification system. They combined *classification rule-mining* methods that aim to *discover* a small set of rules to classify new data, with *association rule-mining* methods that aim to discover all rules in a database that satisfy some user-defined *minimum support*[2]. The authors adapt an association rule-mining algorithm developed by Agrawal and Srikant (1994) to mine all *class association rules* (CARs). For classification the CARs are in the form $X \Rightarrow y$ where $y \in Y$ is a class label from the set of class labels Y, $X \subseteq I$ and I is the set of all items. To evaluate the evidence for the CAR, it is said that the CAR holds in D with confidence c if c% of cases in D that contain X are labelled with class y. For instance, if 100 of 1000 cases have the rule X and 90 of these cases associate rule X with class y, then the rule has a confidence value of 0.9. The rule $X \Rightarrow y$ has support s in D if s% of the cases in D contain X and are labelled with class

[2] The support of a sequence is the percentage of data sequenes in the archive that contain the pattern sequence.

y. The support value will then be 0.09. The confidence and support are two parameters most often used in rule selection and classification scoring. The authors integrate the two approaches by selecting each subset of the association rules that is associated to one class. This subset then becomes the class association rules. By generating all rules none are overlooked. A classifier is built based on those rules that satisfy the minimum support. By integrating the two methods they were able to produce a number of advantages over previous classification algorithms; the generated rules are easy to understand, whereas other algorithms manipulated the data before searching for rules; all rules are discovered and the weakest rules are discarded, instead of finding only a few rules. Their algorithm scales well with database size.

A number of short-comings of Liu *et al*. (1998) have been addressed by Liu *et al*. (2003). Firstly, in the earlier work their classifier is binary in that a new data case can only belong to one class. Further, it does not account for noisy data and skewed class distributions, which can bias the support and confidence parameters. Although one of the advantages of the rule-mining algorithm is that it discovers all rules, it should be noted that a minority class may have support for a class association rule that is below the minimum threshold, although it may be a very confident class rule, and thus will not be selected. To address the first issue the authors propose a scoring algorithm using the CARs termed *Scoring Based on Associations* (SBA). The objective is to score a new case by assigning a probability estimate of the new case belonging to a particular class. By generating all rules an excess of information is created. This presents a challenge because the same items may be associated with different classes and multiple rules may be associated with a single class that

may either be accurate, or simply noise. The problem is which rules to value above others.

To address the class imbalance, a class-specific minimum support and minimum confidence is proposed by the authors. An overall minimum support (called t_minsup) is specified and then adjusted to each class according to;

$$minsup(C_i) = t\_minsupp \times \frac{f(C_i)}{|D|} \quad (4.1)$$

where $f(C_i)$ is the number of C_i classes in the training data. $|D|$ is the total number of cases in the training data. This gives the rules in frequent classes a higher minimum support, and vice versa for infrequent classes. For minimum confidence the following formula is used;

$$minconf(C_i) = \frac{f(C_i)}{|D|} \quad (4.2)$$

Thus, rules for a class that do not have a confidence greater than the probability of the class itself, are excluded. Each of the generated rules will be assigned a support and confidence, which should be used in scoring the data.

4.3.2 Incorporating sequence data

Hu *et al*. (2007) adapted these classification methods for time sequential data, where the value of an attribute may change over time and form part of the classification process. They consider a *sequence* as an ordered list of events $\alpha = < e_1, e_2, \ldots, e_n >$

where each event e_i may contain an attribute only once. A data case d has the form $d = \{\alpha, y\}$ where $y \in Y$ and Y is the set of all labels. Hu *et al*. (2007) also applied the concept of multiple minimum supports that reflect the different frequencies of items or class labels. Generating the candidate sets in Hu *et al*. (2007) follows the GSP algorithm of Srikant and Agrawal (1996) with minor modifications. Tseng and Lee (2009) showed that building a classifier from sequential patterns extracted from each class always outperformed a classifier built from patterns extracted from the entire database. They showed this in their data mining method, which they termed *Classify-By-Sequence* (CBS). Therefore, the database is partitioned according to the class labels and the frequent sequences of each class are mined. In other words, frequent sequences are discovered for each individual class label.

4.3.3 Rule scoring algorithms

Liu *et al*. (2003) showed that scoring by all matched features can perform better than scoring using the best matched feature. Scoring the data is fairly simple once all the sequences from the class association rules have been discovered. For a new unclassified sequence, all classification rules of various lengths need to be discovered. Each of these rules may have different class support and point to different class labels with various confidence. The score for each class is calculated as the sum of scores for each discovered class association rule and the class with the highest score is chosen as the best prediction. Two scores are generated for each class, the *weighted confidence* and the *weighted support* (Hu *et al*., 2007), and are given respectively;

$$\sum conf \times \frac{supp}{\min(d)} \tag{4.3}$$

and

$$\sum \frac{supp}{\min(d)} \tag{4.4}$$

respectively, where *min*(d) is the minimum support threshold of either the item or the class. Although not discussed or implemented in previous work, here the score for each variable is normalized. The reason for this is that the wind variable has 12 possible states whereas the SST and Chl *a* variables have only seven. Thus, there are many more potential wind sequences and the support for any pattern is likely to be lower. The scores are normalized according to;

$$norm_{score} = \frac{score - \min(score)}{\max(score) - \min(score)} \tag{4.5}$$

The confidence of a class rule is essential as it represents the probability estimate. The support reflects the reliability of the rule, since a rule with poor support covers few of the data cases and may be overfitted. The *min(d)* parameter tends to weight in favour of a class that has fewer data cases and therefore is likely to have lower support for those sequences. For example, a rule $X \Rightarrow y_i$ that occurs more frequently in a larger class dataset D_{yi} will have greater support than a rule with the same sequence $X \Rightarrow y_j$ in a smaller class dataset D_{yj} even though the proportion of rules in each class dataset may be the same. However, the smaller dataset D_{yj} will have a smaller *min(d)* that will weight the score less negatively. Liu *et al*. (2003) presented a more complex weighted average scoring method, which is also tested in

this chapter. Their approach aims to achieve a balance between the number confidence of positive class rules and negative class rules (those class rules covering the sample data case but excluding the class being evaluated). Their equation is;

$$\frac{\sum_{i\in POS} w^{i}_{positive} \times conf^{i} + \sum_{j\in NEG} w^{j}_{negative} \times conf^{j}_{positive}}{\sum_{i\in POS} w^{i}_{positive} + \sum_{j\in NEG} w^{j}_{negative}} \qquad (4.6)$$

where:

POS is the set of positive class rules that can cover the data case.

NEG is the set of negative class rules that can cover the data case.

$w^{i}_{positive}$ is the weight for the positive class rule $i = conf^{i} \times supp^{i}$.

$w^{j}_{negative}$ is the weight for the negative class rule $j = \frac{conf^{j} \times supp^{j}}{k}$ where k is a constant that reduces the impact of the negative class rule j.

$conf^{i}$ is the original confidence of the positive class rule.

$conf^{j}_{positive}$ is the confidence after converting the negative class rule j to a positive class rule, i.e. $conf^{j}_{positive} = 1 - 'the confidence\ of\ rule\ j'$.

Both methods described above are tested during the implementation of the classification model on the profile data, which is discussed below.

4.3.4 Classifying surface variables to profile classes

There are differences between the structure of the database used by Liu *et al*. (2003) and Hu *et al*. (2007), and the data structure used here. These differences determine the structure of the candidate rule-generation algorithm. Firstly, their data elements (transaction at time t) consist of numerous items. For example, they consider three elements with various numbers of items from the available attributes in each element such as <[(a,1),(b,2)] [(a,2)] [(a,1),(b,3),(c,4)]>. In the present study, an element could consist of three items; wind, SST and surface Chl *a*. However, as will be indicated later, the combinatorial potential is far too large for the number of available data cases. Instead, each sequence is made up of items of only one attribute. Secondly, the authors are interested in sequences in the database that have no particular reference to a start time reference. A two-element sequence can occur anywhere along the data sequence. In this study, the purpose is to find sequences along a six-day time-series which is referenced to the date of the profile. A six-day series is chosen because this is the approximate time taken for phytoplankton bloom development in the Benguela system (Hutchings *et al.,* 1984). Hence, sequences with one element will always be at the position t_0. These two attributes and the relatively small number of data cases (all the cases can easily be held in computer memory) make the candidate generation algorithm more straightforward.

In this chapter two methods are tested to generate the class association rules. Although both are based on Hu *et al*. (2007), they differ in that Hu *et al*. (2007) began at the start of the sequence t_1 and generate all sequences of length m from t_1. Those sequences that have minimum support in the class subset are recorded along with their confidence. This is repeated for t_2-t_{N-1}. Here, sequences of predetermined

length are used and two different length options are tested. Given the length of the sequence N = 6 and that the profile class is recorded at t = 6, in the first method sequences of length m = {$t_{1:4}$, $t_{2:5}$, $t_{3:6}$, $t_{4:6}$, $t_{5:6}$} are generated. In the second method m = { $t_{1:3}$, $t_{2:4}$, $t_{3:5}$, $t_{4:6}$, $t_{5:6}$}. For generated sequences of length > 4 the minimum has to be set very low to retain any of the sequences. This occurs as a result of very few repeated patterns of length > 4 for many of the database subsets.

Generating multiple sequences allows some flexibility in the classification given the potential for noisy data. For every six-day sequence of observations, five sub-sequences are generated, each with its own confidence. Those sub-sequences containing noisy data are expected to have reduced confidence if they occur frequently within other profile classes or infrequently within a single class. These methods are tested using a leave-one-out approach, where the model is trained on all the data cases except one. The trained model is then used to predict the profile for the omitted case and the highest scoring profile is recorded. The process is repeated for each data case. This method was chosen because there are few training samples relative to the large number of potential patterns and the method maximizes the number of training cases. The results are then visualized using a *confusion matrix* that compares the predicted class against the actual class. Both methods are tested on the temperature and Chl *a* profiles using various combinations of wind, SST and Chl *a* data. The two scoring methods of Liu *et al*. (2003) and Hu *et al*. (2007) are also tested. All the methods and the relevant equations are coded using the MATLAB® programming language. The best results are presented below for each of the profile sets.

4.3.5 Results

Figure 4.4 shows the result for the temperature and Chl *a* profiles as confusion matrices. The best results (highest percentage correctly predicted) are obtained using shorter sequences of length three and the scoring method of Hu *et al*. (2007). Best results for the temperature profile classes are obtained using both wind and SST sequences. For the Chl *a* profiles, best results are obtained using wind, SST and surface Chl *a*. Overall 33% of the temperature and wind sequences are classified correctly according to temperature profiles (the correct profile is associated with the sequences) and 18% of the wind, SST and Chl *a* sequences are correctly classified according to Chl *a* profiles. Only temperature Profiles 5, 9 and 11 and Chl *a* Profile 1 achieve over 50% accuracy. Temperature Profile 4 receives the fewest correct predictions and most often is confused with Profiles 5 and 7. Some of the profile confusions can be explained by the similarity between certain classes. For example, Profiles 1 and 3 have a similar mixing layer depth and thin thermocline layer below and Profiles 5, 7 and 9 have similar structures. The hard clustering of the profiles, which means a profile can only belong to one class, does not allow for gradients between classes. Further, the surface sequence data are also hard partitioned. This means it is highly likely that certain sequences will be frequent for more than one profile class. The algorithm selects the class based on highest score, therefore the classification will always reflect the highest confidence or probability.

Chl *a* Profiles 2 and 3, which represent low biomass mixed to ca. 20 m and a low surface bloom respectively, are frequently predicted incorrectly. Profile 3 is structurally similar to Profile 5 but with a lower surface biomass and is consistently predicted instead of Profile 5. Similarly, Profile 2 is predicted instead of Profile 4. In these cases it is assumed that the surface Chl *a* should distinguish among these

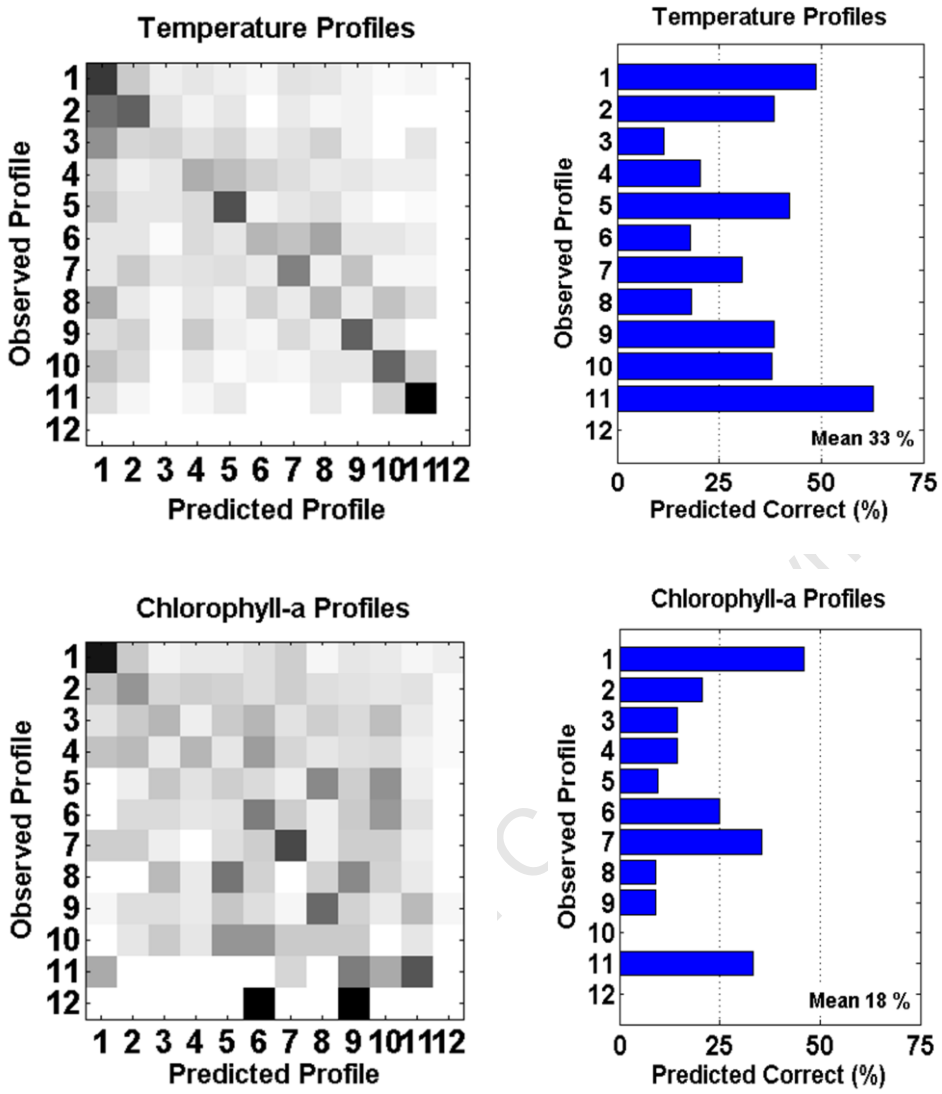


Figure 4.4 The best results obtained from the classification of sequence data for temperature and Chl *a* profiles. The temperature profile classifier obtained best results using a 3 sequence approach and scoring using the method based on Hu *et al*. (2007). The Chl *a* profiles were classified best using a 4 sequence approach and the scoring method of Liu *et al*. (2003). For both classifiers including wind data provided improved results to only SST or Chl *a* sequences. The model was trained randomly on 75% of the data and tested on the excluded 25%. Sequences with missing values were excluded. The grey scale shading indicates the relative distribution of the predicted states for an observed state.

profiles and that the error observed is probably the result of the weak correlation between satellite surface Chl *a* and the *in situ* profiles (r = 0.68, Fig. 2.16). It may also result from the selected intervals used to discretize the surface data.

One particular problem that is likely to have a strong influence on the outcome of the tests is the missing data in the SST and Chl *a* sequences. Of the 1004 temperature and Chl *a* profiles from the St Helena Bay area, 71% of the cases contain missing data in the sequences as a result of cloud cover and other limitations on satellite sensors. Better results are obtained by excluding sequences with missing values but then the model is based on far fewer samples than there are potential sequence combinations. For this reason a method to accurately interpolate or predict the missing values is required for a more robust classifier.

4.4 **Predicting missing values from sequences**

4.4.1 Interpolation

Interpolation constructs new data points within the range of a discrete set of known points. This discrete set of known points can be approximated by a simple function for the interpolation of the unknown data. One of the simplest methods is linear interpolation where the estimated value is taken along a straight line between two points. Although simple, it is often not very accurate and errors tend to increase with the distance between points. Polynomial interpolation generally provides a smoother fit to continuous data but may result in a high-degree (n-1) function and may exhibit oscillatory artefacts at the end points. Spline interpolation (Schoenberg, 1946) uses low-degree polynomials in each of the intervals so that the data function is represented by a set of polynomials, which can reduce the interpolation error of a

single polynomial. These methods are tested on 10,000 random samples of data that have a complete sequence of the input or observation variables. All three of these methods can only operate between two observed values, so the first and last value of the sequences must exist. To test the accuracy of these methods, either three or four values are randomly removed between the first and last values prior to testing and the interpolation methods above are used to estimate the missing value. Interpolation is performed on the raw (undiscretized) SST and log-transformed Chl *a* data and the interpolated value discretized and compared to the true discrete observation. The methods are applied to data subdivided by season and the results are presented in Table 4.1 as confusion matrices. The matrices have the correct or observed state on the y-axis and the predicted state on the x-axis, therefore the correctly predicted states fall along the main diagonal. The grey shading indicates the relative frequency distribution (darker grey indicating a higher frequency) of the predicted states for a particular observed state. Hence, the predicted distribution is indicated along the rows. Percentages of correct estimates and of estimates that fall adjacent to the correct state are shown in the figures.

Table 4.1 A comparison of interpolation methods for SST (above) and Chl *a* (below) sequences. The randomly removed observed state *s* = {1,…,7} on the y-*a*xis is compared against the corresponding interpolated value on the x-*a*xis. The correct interpolation runs on the diagonal. The grey scale shading indicates the relative distribution of the interpolated states for an observed state.

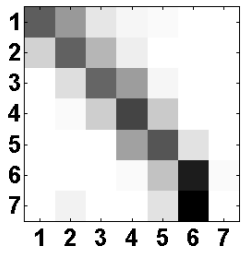
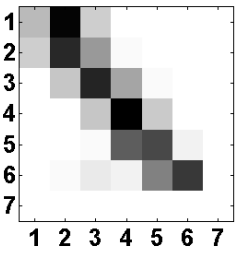
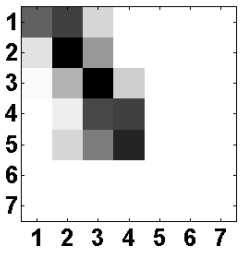
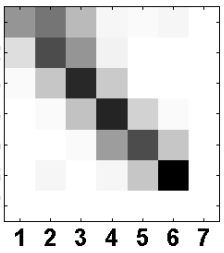
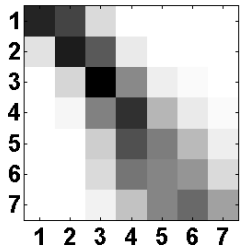
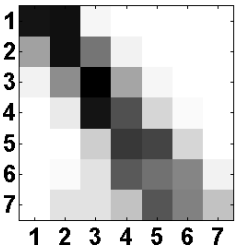
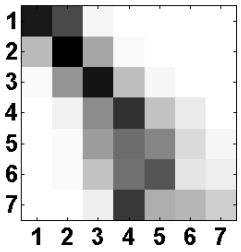
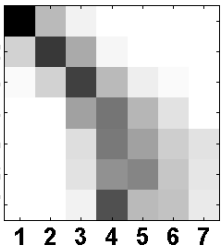
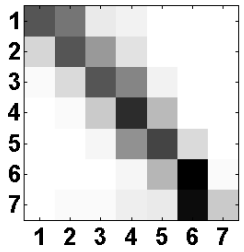
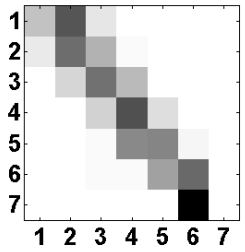
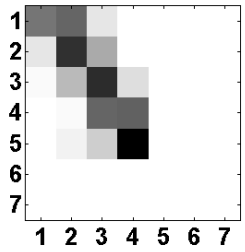
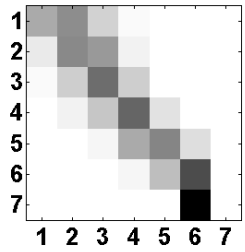
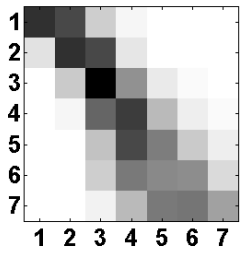
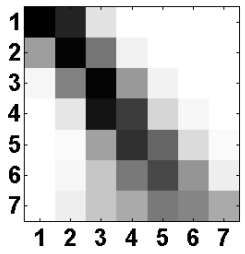
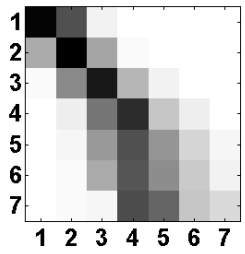
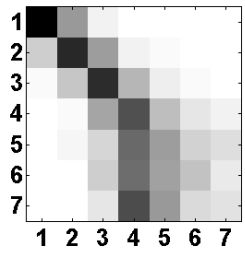
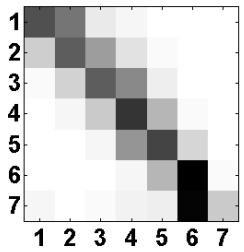
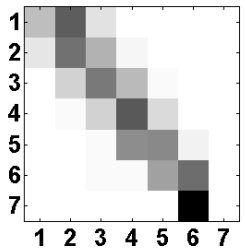
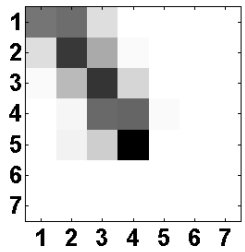
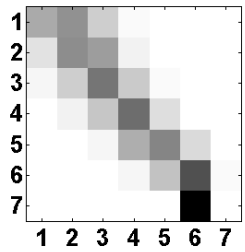
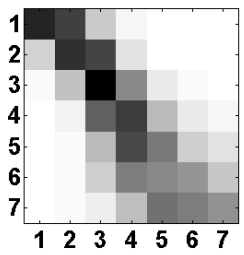
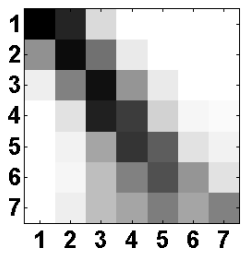
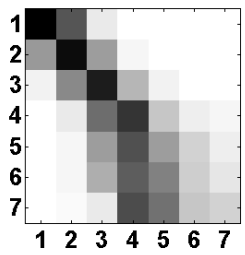
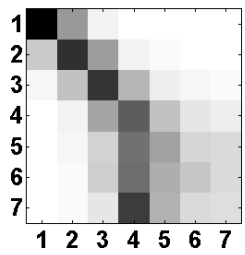
	Summer	**Autumn**	**Winter**	**Spring**
	Linear Interpolation			
SST	% Corr. 59 % Adj. Corr 37	% Corr. 60 % Adj. Corr 38	% Corr. 61 % Adj. Corr 37	% Corr. 60 % Adj. Corr 35
				
Chl *a*	% Corr. 47 % Adj. Corr 42	% Corr. 46 % Adj. Corr 47	% Corr. 53 % Adj. Corr 40	% Corr. 50 % Adj. Corr 38
				

Table 4.1 continued

	Summer	Autumn	Winter	Spring
Cubic Interpolation				
SST	% Corr. 57 % Adj. Corr 38	% Corr. 58 % Adj. Corr 40	% Corr. 61 % Adj. Corr 37	% Corr. 55 % Adj. Corr 40
				
Chl *a*	% Corr. 45 % Adj. Corr 43	% Corr. 45 % Adj. Corr 47	% Corr. 52 % Adj. Corr 41	% Corr. 51 % Adj. Corr 36
				
Akima Spline Interpolation				
SST	% Corr. 56 % Adj. Corr 39	% Corr. 56 % Adj. Corr 41	% Corr. 59 % Adj. Corr 39	% Corr. 53 % Adj. Corr 40
				
Chl *a*	% Corr. 43 % Adj. Corr 43	% Corr. 44 % Adj. Corr 46	% Corr. 50 % Adj. Corr 42	% Corr. 49 % Adj. Corr 36
				

All three interpolation methods give similar results. Interpolation performs better on the SST data than the Chl *a* data. Linear interpolation produces the best results and the Akima spline interpretation produces the poorest. In summer the linear interpolation struggles to predict the highest temperature state correctly, which the other two methods also do in summer, autumn and spring. This suggests that very warm water may appear suddenly such as during intrusions of oceanic water, filaments or warm-cored eddies onto the shelf (Shannon, 1985). The interpolations also do not accurately predict very cold events in autumn, winter and spring but they do in summer. This may be because strong persistent south-easterlies in summer cause predictable upwelling, whereas during the other seasons the southerly winds may be interrupted more frequently and less consistently in terms of periodicity. With the Chl *a* data the linear method is again marginally better than the other two methods. All three methods perform well on the lower concentrations but their accuracy steadily decreases as the state values increase. This may be due to the initial slow growth of phytoplankton when at low concentrations and the more rapid and erratic growth at higher concentrations. It may also be the result of more isolated aggregates of phytoplankton retained in near-shore currents and filaments over the shelf where higher concentrations of phytoplankton are usually observed. The results of the interpolation show that the changes in SST and Chl *a* are not consistent, especially the occurrence of very cold SST and very high Chl *a* concentration. Although the interpolation methods perform well, they are limited in that they will deteriorate as the gaps between observations become larger and there is no additional information on which to base interpolations. The results are also likely to deteriorate as the class intervals become smaller. For example, the temperature range for a correct prediction, is approximately 1.5°C and temperatures

will not often vary by more than that over a few days, making interpolation much easier. Therefore, the next section investigates methods that can account for the changing environment when filling in the missing data.

4.4.2 Hidden Markov Models

A more informed approach than the standard interpolation methods in Section 4.3.1, is to predict the next most likely value in a time series given observations of previous values. It is intuitive to assume that recent observations have more useful information than older observations for the predictive task. This is the form of *Markov* models, which are a subclass of Bayesian networks (Bishop, 2006). The simplest form is a first-order Markov model that assumes that an observation is dependent on only the immediately preceding observation and not on any prior to that. A second-order Markov model will have an observation dependent on the two most recent observations. The problem with the m-order Markov chain is that the number of possible sequences that must be accounted for grows exponentially with m. This can be resolved by introducing a corresponding *latent* or *hidden* variable (the dimensionality or type of variable may be different from the observations) for each observation in the sequence, and allowing the hidden variable to form a first-order Markov chain. This allows any observation x_m to be connected to the predicted observation x_n via the hidden variables. The hidden variable y_n corresponding to x_n essentially stores the information of all previous observations. The graphical structure of this model is known as a *state space model*, and if the hidden variables are discrete it is known as a *Hidden Markov Model* (HMM). The addition of the input variable extends the HMM framework to the supervised learning domain.

HMMs are commonly used in labelling sequential data where the task is to identify the most likely sequence of labels that are generated by a sequence of observations. This is achieved by assigning a label (hidden variable Y), such as a part-of-speech label to a word that is part of a sequence of words in a sentence (observation variable X). The HMM defines a joint probability distribution of the two random variables $p(X, Y)$. In other words the joint probability of a sequence of hidden states y and an observation sequence x factorizes as (Ghahramani, 1997),

$$p(y_t, x_t) = p(y_1)p(y_1|x_1)\prod_{t=1}^{T} p\,(y_t|y_{t-1})p(x_t|y_t) \qquad (4.7)$$

Determining the joint distribution is made tractable by assuming that the observation x_t depends only on the corresponding state or label y_t and is independent of all the other states and observations at other time indices. Most real-world observation sequences however, are best represented by multiple interacting observations. For example, it is hypothesized here that sequences of surface Chl *a* (represented as discrete labels) depend on observations of SST and wind sequences, and that these observations may interact. A model is then required that supports tractable inference and that does not make unjustified assumptions of independence. This can be achieved by a model that defines a conditional probability $p(Y \mid x)$ of all label sequences given a specified sequence x rather than X (the set of all observation sequences). Conditional models select a label sequence y that maximizes the conditional probability $p(y_i \mid x)$ at each time step. This approach does not waste effort on modelling the probability of the observations and makes no unwarranted assumptions of independence about them. Further, HMMs cannot include observed

information to influence the state transition of the hidden variable (Wallach, 2004). In the situation here it is desirable that the wind data be allowed to influence the change in state of either SST or Chl *a* rather than only the current state. Conditional random fields are flexible enough to handle these issues.

4.4.3 Conditional Random Fields

Where HMMs are a class of model known as *generative* models i.e. they model the joint distribution of the variables x and y and are therefore able to generate samples of sequences, *conditional random fields* (CRFs) fall into the *discriminative* model class (Sutton and McCallum, 2006). They model the dependence of the unobserved variable y on the observed variable x. This is the objective of the task of filling in the missing data; to predict the missing value of either SST or surface Chl *a* based on relevant current and recent environmental information that influences the most likely state and/or the state transition.

CRFs are a framework for labelling sequential data based on the conditional approach described in Equation 4.7. They are built on probability theory and dynamic programming to be able to label statistically sequences or temporal patterns. There may be millions of such patterns and many of them may conflict with each other. By applying the statistical properties of the labels given the observations the most likely label or label with the highest *likelihood* can be determined. They have been shown to outperform HMMs on a number of real-world sequential labelling tasks (Lafferty *et al*., 2001; Pinto *et al*., 2003; Sha and Pereira, 2003).

4.5 **Conditional Random Fields for missing values**

Conditional random fields (CRFs) are represented by an undirected graphical model (Lafferty *et al*., 2001). The structure of the graph may be arbitrary, provided it represents the conditional independencies in the label sequences being modelled. The most commonly used structure, and that used here, is the first-order Markov chain, which states that the label at $t = i$ is only dependent on the label at time $t = i$-1 and no other labels.

A simple example is the prediction of a sequence of surface Chl *a* over a six-day period. Let there be a set of three labels for Chl *a* concentration, Y = {$high$, med, low} and a sequence of six days T = {1, 2,…, 6}. In the first-order HMM it is assumed that the hidden variable y_t (where $y \in Y$ and $t \in T$) is only dependent on the immediately preceding hidden variable y_{t-1}. If we hypothesize that the value of the Chl *a* label at time *t* is influenced by the SST and the transition of labels (y_{t-1}, y_t) is influenced by wind stress at time *t* then *t* can be described by a set of *feature functions* $f_{sst}(t)$ and $f_{wind}(t)$. Each feature function can be a binary output if the set of observation and label conditions are met. For example, f_{85} (where 85 is the arbitrary feature function index) may be *1* if the SST state at t_{-1} is the same as SST state at *t* and the Chl *a* label is med, and *0* otherwise. This will be discussed further below.

4.5.1 Feature functions

The accuracy of labelling is strongly dependent on the set of features selected since they relate the observations or patterns to the labels. In sequence labelling it is common practice to deal with a segment of local data patterns associated to the

label (Sutton and McCallum, 2006). Here a pattern of maximum three consecutive days has been chosen. For example, the SST transition is related to individual one, two and three day sequences of wind stress with no assumption made about the independence among the wind observations. At time t therefore, there will be a set of features relating patterns of various wind duration with a particular label (y_t) or label pair (y_{t-1}, y_t). For example, the features might be as follows; the SST warms, from time t_{-1} to time t, which is represented as a change of state from State 3 to State 4. Two patterns associated with this transition feature may be that there is a pattern of two consecutive days of calm wind or that there is one day of moderate southerly wind followed by two days of weak southerly wind. Each feature has a pattern (g) that returns 1 if it is true or 0 otherwise. A pattern can then be represented as function;

$$g_{200}(x,t) = \begin{cases} 1 & x_{t-1} = 'calm' \; and \; x_t = 'calm' \\ 0 & otherwise \end{cases} \qquad (4.8)$$

where g_{200} is the arbitrary 200th function which is "switched on" if the x vector indicates calm wind at time t_{-1} and t. As the number of patterns can become very large it is useful during training and inference to create a set of only the patterns which are *switched on*[3] for a particular instance. The patterns are extracted and associated with a label or label pair to obtain the model features as one of the following:

[3] A binary value indicates whether to evaluate the feature (*1*) or not (*0*).

- *node association* which takes into account the current label y_t and the pattern $g_m(\boldsymbol{x},t)$ as a node feature $f_k(y_t,\boldsymbol{x},t)$
- *edge association* which takes into account two consecutive labels y_{t-1} and y_t and the pattern $g_m(\boldsymbol{x},t)$ as an edge feature $f_k(y_{t-1},y_t,\boldsymbol{x},t)$

The simplest form of the edge feature is for example;

$$f_{10}(y_{t-1}, y_t, t) = \begin{cases} 1 & y_{t-1} = 3 \; and \; y_t = 4 \\ 0 & otherwise \end{cases} \tag{4.9}$$

which models the co-occurrence pattern of label pairs. However, if it is believed that the transition from y_{t-1}-y_t is influenced by the data then the structure should be for example;

$$f_{100}(y_{t-1}, y_t, x, t) = \begin{cases} g_{200}(x,t) & y_{t-1} = 3 \; and \; y_t = 4 \\ 0 & otherwise \end{cases} \tag{4.10}$$

where the value of the pattern g_{200} is returned for feature function f_{100} if the label pair satisfies the feature.

Typically, as is done here, each pattern $g(\boldsymbol{x},t)$ is associated to each label y_i or label pair y_h, y_i where $y_{h,i} \in \{y_1, y_2, \ldots y_N\}$. So if the pattern of two days of calm wind is observed it is associated with all possible SST state changes to obtain the possible set of feature functions. Then each feature function $f_k(y_{t-1}, y_t, \boldsymbol{x})$ has its own weight w_k that captures the affinity of the pattern and label y_i or label pair y_h, y_i in the training dataset. This approach assumes that there is no prior knowledge of how the SST responds to forcing, and the model is used to *discover* these relationships. A

list of some of the features chosen for the SST and Chl *a* label is presented in Table 4.2.

Table 4.2 Summary of the feature functions used in the SST and Chl*a* Conditional Random Field (CRF) models. Each label or label pair at time *t* from the set of labels *Y* in the Labels column is associated with the pattern in the Pattern column. During training the strength of the association between the label and pattern is encoded in the feature weight.

Label	**Pattern**	
Node Associations		
for all $y_i \in Y$	for each *depth* = {1,…,5} and *month = {1,...,12}*	
for all $y_i \in Y$	For each *sst* = {1,…,7}	For Chl*a* sequence model only
Edge Association		
for all $(y_{i'}, y_i) \in Y$	for each *wind* = {pattern210} and *season* = {1,…,4}	Where *pattern210* is the set of 3 day wind patterns from *t-2* to *t*
for all $(y_{i'}, y_i) \in Y$	for each *wind* = {pattern21} and *season* = {1,…,4}	Where *pattern21* is the set of 2 day wind patterns from *t-2* to *t-1*
for all $(y_{i'}, y_i) \in Y$	for each *wind* = {pattern10} and *season* = {1,…,4}	Where *pattern10* is the set of 2 day wind patterns from *t-1* to *t*
for all $(y_{i'}, y_i) \in Y$	for each *wind* = {pattern0} and *season* = {1,…,4}	Where *pattern0* is the set of 1 day wind patterns at *t*
for all $(y_{i'}, y_i) \in Y$	for each *depth* = {1,…,5} and *month* = {1,…,12}	

4.5.2 Potential functions

The *potential functions* refer to quantities that describe the relationship between the labels and the local data patterns (feature functions). The strength of the relationship of the individual feature functions is encoded in the weight associated with each feature function. For example, if the SST state change from 3 to 4 is often observed in the training data when there are two consecutive days of calm wind, this feature function will have a higher weight than other less frequently associated state transitions. This particular state label pair (and all other potential pairs) can have other features associated with it, such as season and depth that may be "switched on". In other words, each potential label or label pair will have its own set of feature functions given an observation vector. The potential function of a label or label pair is then the exponential sum of weighted feature functions. In the case of the node potential it is given as;

$$\varphi_t(y_t, x) = exp\left(\sum_k w_k f_k(y_t, \boldsymbol{x}, t)\right) \tag{4.11}$$

and the edge potential

$$\omega_t(y_{t-1}, y_t, x) = exp\left(\sum_k w_k f_k(y_{t-1}, y_t, \boldsymbol{x}, t)\right) \tag{4.12}$$

The weights w_k are the model parameters, which are estimated during training. The exponent of the linear function ensures that the solution is always non-negative. The dependence of the label y_t on a global vector $\boldsymbol{x}(t)$ does not necessarily mean that the observations must occur at time t but rather all the observations of any time step that influence the label at time t. Further, no assumptions are made concerning the

relationships of the variables with one another; the observation patterns may overlap so that two or more patterns can refer to common observations. For example, the SST state y_t may be related to the corresponding wind x_t, but also to previous consecutive days [x_{t-1}, x_t] and [x_{t-2}, x_{t-1}, x_t].

4.5.3 Calculating conditional probabilities

Assuming the optimal model parameters have been obtained during the training phase, the model can then be used to estimate the most likely label sequence conditioned on the observation vector. The conditional probability of Y given X is defined as;

$$p(y|x) = \frac{exp \sum_k \psi_k(y, x)}{Z(x)} \tag{4.13}$$

where $\psi_k(y, x)$ are real-valued potential functions and $Z(x) = \sum_{y'} exp \sum_k \psi_k(y', x)$ is the normalization constant known as the *partition function* (Wallach, 2004; y' denotes all possible values of y). The normalization constant sums over all possible state sequences and is thus; it exponential to the number of time steps. Fortunately it can be computed efficiently using dynamic programming, as will be explained in the next section. The form of the CRF model is based on a log-linear model. The exponent of the numerator ensures that the solution is always positive and the denominator ensures that the distribution of y sums to one, hence the results are between 0 and 1 making them valid probabilities. The denominator solves the label bias problem of other discriminative models such as *Maximum Entropy Markov Models* (Lafferty *et al*., 2001). The label bias problem arises when local state transition probabilities are

locally normalized. For example, in the training dataset state a_t might move only to state a_{t+1} (with probability p=0.4) or state b_{t+1} with slightly higher counts for $a_t \rightarrow b_{t+1}$ (p=0.6). State b_t might move to either states a_{t+1} (p=0.25), b_{t+1} (p=0.3), c_{t+1} (p=0.25) or d_{t+1} (p=0.2) with highest probabilities for $b_t \rightarrow b_{t+1}$. The most probable path will be $a_{t1} \rightarrow a_{t2} \ldots \rightarrow a_{tn}$ because the product of probabilities will be higher e.g. $a \rightarrow a \rightarrow a \rightarrow a$ (0.4 x 0.4 x 0.4) as opposed to $a \rightarrow b \rightarrow b \rightarrow b$ (0.6 x 0.3 x 0.3). In other words, even though $a \rightarrow b$ is more likely it is not included in the most likely path. The reason for this is that for a the transition counts are distributed between only two states whereas there are four possibilities for b. By using a global normalization the sequence probabilities are used rather than the local state transition probabilities

4.5.4 Training phase

Developing the conditional random field (CRF) model requires a training phase to obtain the model parameters and a testing phase to analyze the accuracy of the model. In the training phase a subset of the data is used. This subset includes the observation sequences and the corresponding label sequences. The statistical properties of the observation data and associated labels are extracted. Typically for long sequences, only local patterns of the observation data around the label position are collected. These patterns are then weighted. The resulting local patterns with their associated label and parameters, form the CRF model. In the testing phase a subset of the data that excludes the training subset is used to compare the trained model output with the known labels.

The training data D^N are IID (independent and identically distributed) and have the form $\{\boldsymbol{y}^i, \boldsymbol{x}^i\}$ for i=1 to N and where $\boldsymbol{y}^i = \{y_1{}^i, y_2{}^i, \ldots y_T{}^i\}$ and $\boldsymbol{x}^i = \{x_1{}^i, x_2{}^i, \ldots x_T{}^i\}$ are sequences of labels and observations respectively (note, bold text indicates a vector). The objective is to obtain the feature weights w_k or the model parameters $\theta = \{w_k\}$ that maximize the likelihood of the label sequences given the observation sequences. Optimizing θ is done by improving the log-likelihood or objective function using an iterative method. It requires the evaluation of the log-likelihood and its gradient with a particular set of parameters. The log-likelihood is in fact the *conditional log-likelihood* $\ell(\theta)$) (Sutton and McCullum, 2006) and can be written as;

$$\ell(\theta) = \sum_{i=1}^{N} \log p\left(\boldsymbol{y}^i \middle| \boldsymbol{x}^i\right) \tag{4.14}$$

This is a sum of the CRF model over all cases in D. By substituting the model (Eq. 4.13) into the likelihood (Eq. 4.14) the following is obtained;

$$\ell(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} w_k f_k\left(y_t^i, y_{t-1}^i, \boldsymbol{x}_t^i\right) - \sum_{i=1}^{N} \log Z\left(\boldsymbol{x}^i\right) \tag{4.15}$$

To obtain the gradient the partial derivatives are taken;

$$\begin{aligned} \frac{\partial \ell}{\partial w_k} = & \sum_{i=1}^{N} \sum_{t=1}^{T} f_k\left(y_t^i, y_{t-1}^i, \boldsymbol{x}_t^i\right) \\ & - \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{y,y'} f_k\left(y_{t-1}, y_t, \boldsymbol{x}_t^i\right) p\left(y_{t-1}, y_t \middle| \boldsymbol{x}^i\right) \end{aligned} \tag{4.16}$$

The first term is the expected value of feature f_k from the empirical data in D, or more simply, the counts of each feature being switched on over the time steps and over all data cases. The second term is the expectation of feature f_k under the model distribution with θ for all possible label sequences. The optimal solution is then when the two terms summed over all training data are equal and the gradient is zero. Functions of the form $g(x) = log \sum_i exp\, x_i$ are convex and for parameter estimation this means that every local optimum is also a global optimum (Lafferty *et al*., 2001). *Optimization* is commonly achieved using *maximum likelihood* methods, which state that the model chosen should be the one that gives the greatest possible probability (not accuracy) to the training data. A simple and fast approach to optimize ℓ is to compute the gradient using one training example at a time and update θ. *Stochastic gradient* methods are the algorithms used in this approach.

Stochastic Gradient Descent

There are other more sophisticated methods of maximizing the total conditional log-likelihood i.e. summing ℓ over all training examples. These include Newton's method and quasi-Newton methods such as *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) (Bertsekas, 1999) and limited memory BFGS or *L-BFGS* (Sha and Pereira, 2003). However, where these methods require a loop through all the data before updating θ, *stochastic gradient descent* (SGD) does the updating after only one sample;

$$w_k \leftarrow w_k + \lambda G_k^x \qquad (4.17)$$

where G_k^x is the k^{th} element of the gradient associated with the k^{th} feature function and sample data x, and λ is a user-defined learning rate ($\lambda > 0$). It is called

stochastic because the derivative based on a randomly chosen single example is a random approximation to the true derivative based on all the training data. Although this method is fast and easy to implement, its final performance may be less than L-BFGS. Advantages of this method are that for a very large training set the objective function may converge in less than one loop through the data and only those parameters that have their corresponding features switched on need to be updated (Tsuruoka *et al.,* 2009).

Forward-backward Programming

In order to maximize the objective function, an efficient method is required to compute the expectation of each feature function for every possible label sequence given the current observation sequence. This computation needs to be performed for each sample in the training dataset. Fortunately, it is not necessary to compute the expectation over entire sequences as the sequence can be factorized as the summation of marginal probabilities of y_i over time $t \in \{1,\ldots,T\}$. The *forward-backward* algorithm defines the *forward probabilities* (α) and *backward probabilities* (β) of a label y_t which are cached in an Y x T and Y x T*-1* matrix respectively (where Y and T are the set of labels and time-steps in the sequence respectively). The forward-pass allows the computation of the normalization constant while both passes enable the local probabilities to be estimated. These probabilities are defined as (Sutton and McCallum, 2004);

$$\alpha_{ti}(y_t) = p(y_{1:t-1}, y_t = i, \boldsymbol{x}) \tag{4.18}$$

and

$$\beta_{ti}(y_t) = p(y_t = i, y_{t+1:T}, \boldsymbol{x}) \quad (4.19)$$

That is α_{ti} is the probability of y_t having label i at time t over all possible label sequences y_1-y_{t-1}. The probability is cached in column t row i in matrix α. Similarly β_{ti} is the probability of y_t have label i at time t over all possible subsequent label sequences t+1-T. The probability is cached in column t row i in matrix β. Note that the length of this matrix will be T-1. The values of α and β can be computed recursively by (Sutton and McCallum, 2004);

$$\alpha_t(y_t) = \sum_{y_{t-1}} \alpha_{t-1,}(y_{t-1})\ \psi(y_{t-1}, y_t, \boldsymbol{x}) \quad (4.20)$$

$$\beta_t(y_t) = \sum_{y_{t+1}} \beta_{t+1}(y_{t+1})\ \psi(y_{t-1}, y_t, \boldsymbol{x}) \quad (4.21)\ .$$

Here, ψ is the exponent of the weight multiplied by the feature function ($w_k\, f_k$). A special base case is needed for $\alpha_{t=1}$ as there is no label for y_{t-1} at t = 1. In addition for β_t where $t \in$ {1:T-1} as there is no label for y_{T+1}.

Although initially computationally expensive, once constructed, the two matrices enable efficient computations of $Z(x)$ as;

$$Z(x) = \sum_{y_T} \alpha_{y_T} \quad (4.22)$$

that is the sum over all the rows in the last column of α and the feature expectation in the calculation of the gradient where the probability term in Equation 4.12 becomes;

$$p(y_{t-1}, y_t|\boldsymbol{x}) = exp\big(w_k f_k(y_{t-1}, y_t, \boldsymbol{x})\big)\alpha_{t-1}(y_{t-1})\beta_t(y_t) \qquad (4.23)\ .$$

Equation 4.23 is the product of the probability of the label pair, the probability of the label $y_{t\text{-}1}$ and the probability of the label y_t . When the left hand term is summed over T and Y it requires the normalization constant $Z(x)$.

Regularization

Regularization is performed during the maximum likelihood training of the model primarily to prevent the model overfitting the training data (Sutton and McCallum, 2004). This is particularly important when training sets are small. There are two methods commonly applied to such situations namely L1 and L2 norm regularization where the norm has the form;

$$\|w\|_p = \left(\sum_{i=1}^{n} |w_i|^p\right)^{1/p} \qquad (4.24)$$

and where p=1 and p=2 for the L1 and L2 norms respectively. The penalty terms can then be written as;

$$-C\sum_i |w_i| \qquad (4.25)$$

and

$$-C\sum_i w_i^2 \qquad (4.26)$$

In other words, L1 regularization penalizes the weight vector $\overline{w}$ according to the sum of the absolute values of the weights whereas, L2 regularization does so proportional

to the sum of the squares of the weights. In both cases a user defined constant $C \geq 0$ is required to control the degree of regularization or smoothing of the objective function. The value of C is chosen by cross validation. L1 regularization has the advantage that many of the features that are noisy can end up with zero valued weights (Sutton and McCallum, 2004). This is useful for producing compact models as the associated features do not add to the potential functions and can be removed prior to the inference stage. However, during training the gradient at each update is a fairly crude approximation and tends to be noisy and often large in value. The application of the L1 penalty can then easily move weights away from zero. It also penalizes all factors equally, even those features not initialized by the input vector. Further, when the weights are zero the penalized gradient of the objective function is no longer differentiable. Tsuruoka *et al*. (2009) presented a method to solve these problems. Their approach keeps track of the total penalty and the penalty applied to each weight. The penalty is then applied based on the difference between these cumulative values. The accumulated penalty is simply the sum of the learning rate to iteration k multiplied by the regularization constant C. Then for each training sample, w_i is first updated as a function of the unregularized gradient and the learning rate parameter. The penalty is the difference between the accumulated penalty and the total L1 penalty the w_i has actually received up to that point. If the weight value switches sign after the penalty is subtracted it is set to zero. In this way the authors point out that the weight is forced to receive the total L1 penalty that would have been applied if the weight had been updated by the true gradients. The learning rate for all three methods is a simple exponential decay;

$$\lambda_k = \lambda_0 \alpha^{-k} \tag{4.27}$$

where α is a constant and λ_0 is the initial learning rate parameter. Parameter values are selected to allow a gradually decay of the learning rate over the length of the training dataset. The regularization constant C was selected from a range of values that maximized the likelihood of a subset of the testing dataset after 20 runs through the data. In this chapter the cumulative penalty method is compared against the unregularized objective function and the L2 regularization. As there is little difference among the methods, the method of Tsuruoka *et al*. (2009) is selected because it identifies those features that have little to no impact on the model score by setting the weight values to zero.

4.5.5 Testing phase

Inference

The task of inference is to find the most likely sequence y for the given observations $\boldsymbol{x}$. The process follows the Viterbi algorithm (Viterbi, 1967) that infers the most likely sequence by solving the most likely state of the hidden variable at each time step, using the cached solution to move forward iteratively. In this way the message is send along the chain in a forward direction. This is similar to the structure of the forward-backward algorithm. The Viterbi algorithm assumes that it is possible to make an unconstrained optimal decision at each time-step, whereas the hidden variable sequences used here often have partial information. This information is presented as at least one observed state in the sequence. The observed state is important as it indicates a point through which the most likely path must travel and

thus constrains the possible pathway. It provides useful information on the recent state of the system. This path constraint requires that the algorithms be adjusted for the normalization constant and inference. These problems are straight-forward to solve. For the normalization constant, instead of all possible paths, only those paths that pass through the observed state(s) are used. The Viterbi algorithm is not appropriate for the inference step as it cannot move through the partial path. Instead the likelihood of each of the possible paths that move through the observed state(s) is estimated and the most likely path chosen. Although this increases the computational time considerably compared to the Viterbi process, computational time can be reduced by adapting the algorithm to evaluate segments separately. For example, if there are observed states at $y_{t=2}$ and $y_{t=5,}$ then the first segment evaluated will be $y_{t=\text{start}}$ to $y_{t=2}$ followed by $y_{t=2}$ to $y_{t=5}$ and $y_{t=5}$ to y_T . For a six-step sequence with seven possible states this reduces the number of calculations from 7^6 to $7^1+7^2+7^1$. Although the normalization constant needs to be calculated three times instead of once, the number of individual calculations is exponential to the length of the sequence. This method then sets a further constraint on the normalization constant to be the length of each segment and includes all paths between and including the observed state(s). This method can easily be checked by ensuring that the sum of the probabilities over all paths of a segment sums to 1.

4.5.6 Results

To train the model a random number of state observations is removed at random positions from each sequence. The model is then required to predict the missing values. The results are presented in Figure 4.5 as confusion matrices that display

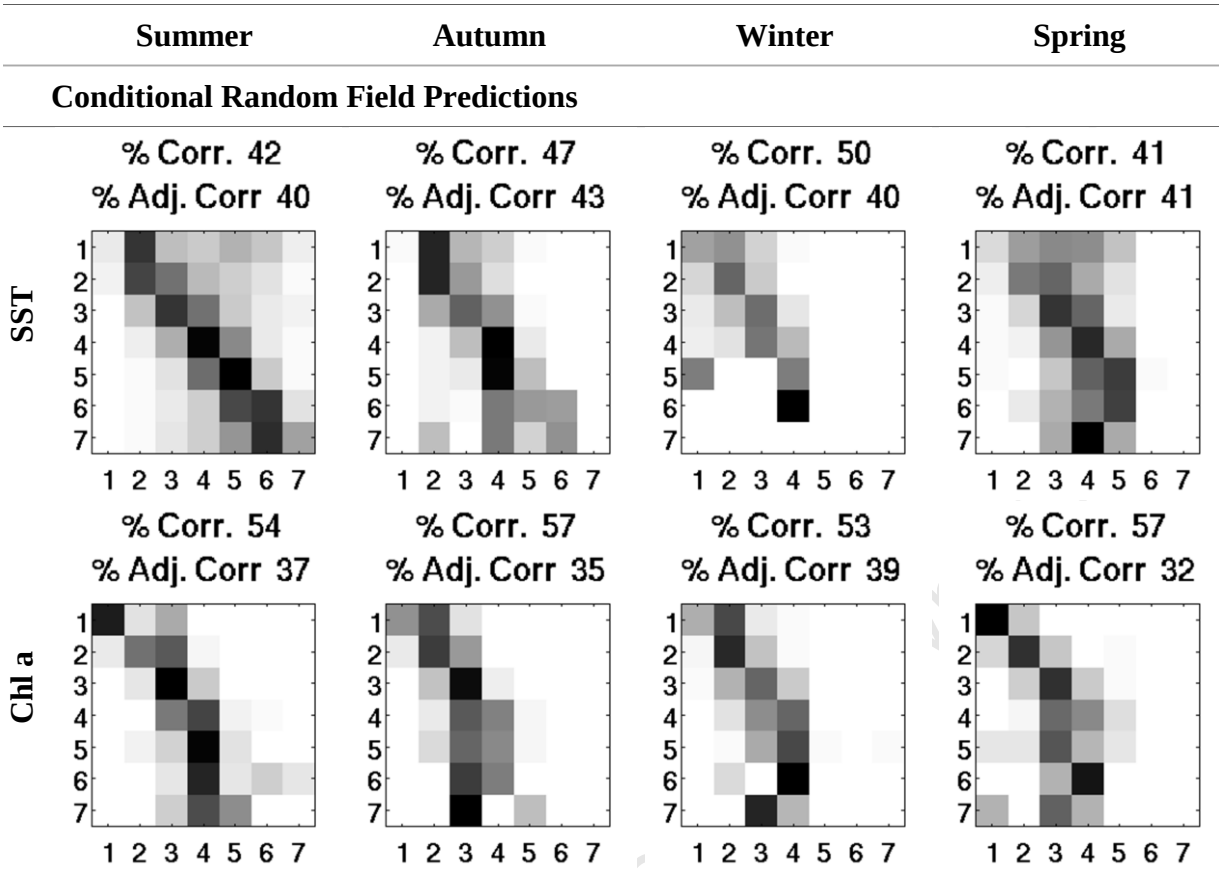


Figure 4.5 Evaluation of the predictive accuracy of the CRF model for missing data in SST and Chl *a* sequences. Grey shading indicates the relative frequency distribution (darker is higher) of the predicted profiles (x-axis) for a particular observed profile (y-axis) i.e. the distribution across the row. The main diagonal indicates the correct predictions.

the accuracy of the model by comparing the predicted labels (x-*a*xis) with the observed labels (y-*a*xis).

For the SST data the CRF performs best in autumn and winter with 90% of the predictions being at most one state above or below the observed state. For the summer and spring seasons this value is 82%. This may be due to more rapid and extreme heating and cooling during summer and spring resulting from a wider range of solar radiation and wind strength. Solar radiation is not accounted for in the model and the combination of variable solar heating and wind strength may complicate the

SST transitions. The prediction accuracy also tends to deteriorate at the lower and higher ends of the SST state. At the lower end the problem may be that the MODIS SST data has not been well sampled in regions of intense upwelling, or the upwelling centres and therefore cold upwelled water is not well represented in the training data (as discussed in Section 2.4.2). At the higher end the cause may be a paucity of observations in the highest temperature range and/or insufficient feature functions that specify the conditions that cause the high temperature values. Additional information such as sea surface height anomalies may help to indicate intrusions of warmer oceanic water, warm cored eddies or filaments.

The Chl *a* results are more accurate than the SST with all seasons obtaining above 50% correct predictions. The autumn and winter predictions confuse State 1 with State 2, incorrectly predicting State 2 instead. This may be the result of the SST sequences which do not differentiate the cold upwelling water from the seasonally cooler surrounding water. There are also errors with higher Chl *a* state predictions, which similar to the SST data, may be the result of the poor performance of MODIS Chl *a* data at high values or the omission of solar photosynthetically active radiation (PAR), which influences the growth of phytoplankton.

4.6 **Revisiting the classification model**

The conditional random field model was trained on a large random dataset from 2002-2009. The training phase calculated the model parameters that produced the highest probability of the observed labels, given the data. The model is used to fill in the missing SST and Chl *a* sequence values so that the classifier can be trained on a larger dataset. The larger dataset is required because the classification rules produced in the dataset are sequences with each element in the sequence having

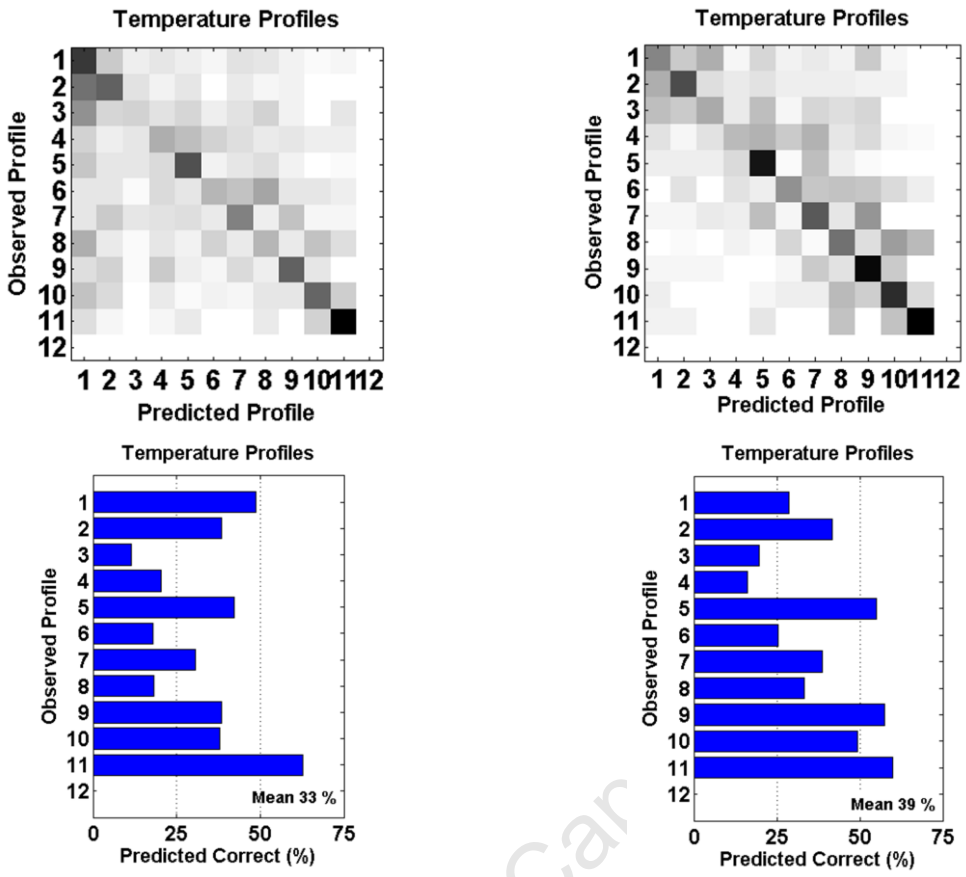


Figure 4.6 Comparison of temperature profile predictions using the model based on the classify-by-sequence model of Tseng and Lee (2009), and the original data with missing data (left panel) and using data with the missing values filled in by the CRF model (right panel). Grey shading indicates the relative frequency distribution (darker is higher) for the predicted profiles.

seven possible states and therefore many combinations are possible. After developing and testing the classification model using the data that included missing values (sequences with missing values were ignored in the training process), the model is retrained on the same dataset but with the missing values filled in by the CRF model. Both the original and the new results of the temperature and Chl *a* profiles are shown for comparison in Figure 4.6 and Figure 4.7 respectively. The new temperature profile predictions (Fig. 4.6, left panel) show an improvement in accuracy when compared to using data with missing values (Fig. 4.6, right panel)

from 33% to 39% correct predictions. The improved score for the temperature profiles can be attributed mainly to an improvement in the predictions of Profile 3 and Profiles 5-10. There is a decrease in the successful predictions of Profile 1, which can be attributed to the CRF model that does not predict the coldest SST state well (Fig. 4.5), therefore, the predicted SST data in the SST sequences associated with Profile 1 are likely to be too warm. There remains some confusion between Profiles 5 and 9 when Profile 7 is observed, and Profiles 6 and 10 when Profile 8 is observed. These groups of profiles have similar structures (see Fig. 2.5) and there will be a continuum of individual profiles among the clusters. In some cases, as a result of the *k*-means clustering process, profiles that share a common mixing layer depth (which is related to wind sequences) will be classed differently if their Euclidean distances (absolute temperature values) are sufficiently large. This will weaken the confidence in the associated wind patterns. In contrast, two profiles with very different mixing layer depths will be classed together if their Euclidean distance is sufficiently small. Thus, a single class may be associated with contrasting wind patterns, which will again weaken each of the patterns. For a given wind observation sequence, the CBS algorithm will then be selecting the class with more consistent mixing layer depths even though the observed mixing layer depth is the same. As an example, suppose a profile has a mixing layer depth of 10 m but because of its surface temperature was classed as Profile 10, which has a mean mixing layer depth of 20 m, rather than Profile 8, where the majority of profiles have mixed layer depths of ~10 m. Then both Clusters 8 and 10 will have the same wind sequences but the observed wind sequence will have higher confidence for Profile 8 because of the consistency of class members. Therefore, the mis-classification of these profiles is

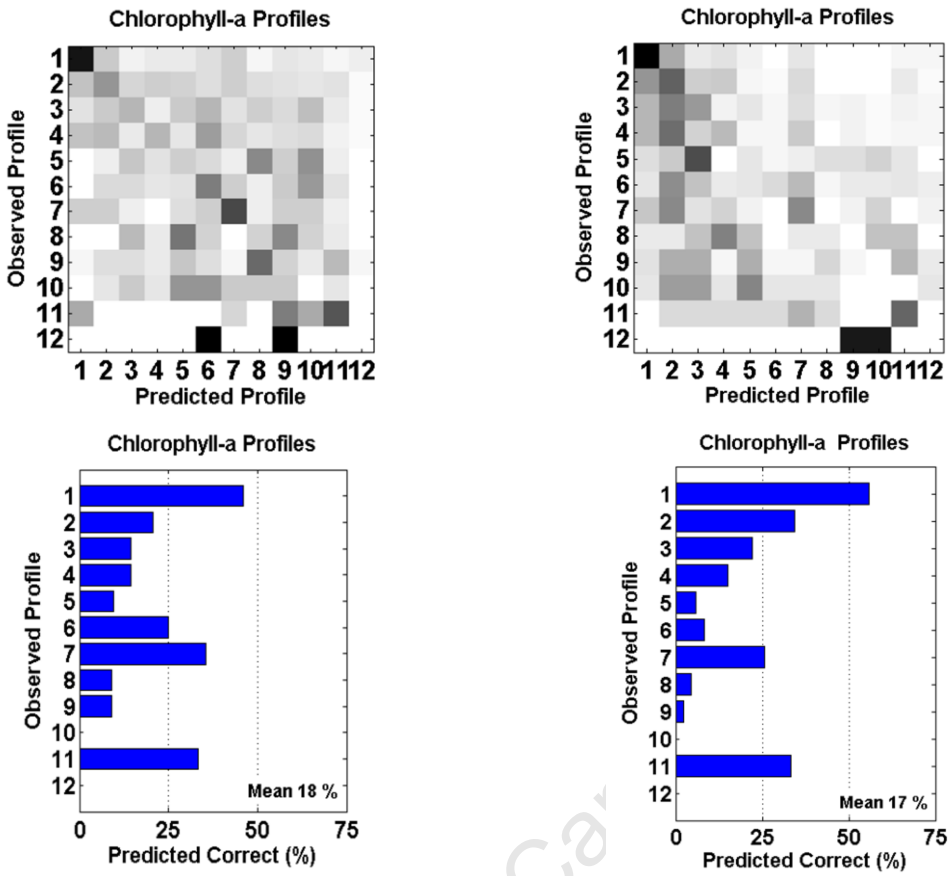


Figure 4.7 Comparison of Chl *a* profile predictions using the model based on the classify-by-sequence model of Tseng and Lee (2009) and the original data with missing data (left panel) and using data with the missing values filled in by the CRF model (right panel). Grey shading indicates the relative frequency distribution (darker is higher) for the predicted profiles.

not considered a weakness of the CBS approach but identifies an area for improvement in the clustering of profiles.

The Chl *a* profiles comparison is shown in Figure 4.7. There is a slight decrease in accuracy when using the CRF model to fill in the missing data to 17% (Fig. 4.7, left panel) from 18% when using the original data (Fig. 4.7, right panel). Prediction accuracy improves for Profiles 1, 2 and 3 (profiles with low integrated biomass; see Fig. 2.6) while deteriorating for Profiles 5-9. For both tests, Profiles 10 and 12 (profiles with high integrated biomass) are never predicted. The improved accuracy

of Profiles 1, 2 and 3 results from the CRF models ability to accurately fill in the missing low Chl *a* values (Fig. 4.5) in the satellite data time-series. Similarly, the weakness in predicting moderate to high biomass profiles is because of poor predictions of high Chl *a* values. Figure 4.7 also shows that as a result of the underestimates of the higher Chl *a* values, there is a shift in predictions from frequently predicting higher profile numbers than the observed profile (area above the matrix diagonal, left panel) to predicting lower profiles numbers (area below the matrix diagonal, right panel). Although the results shown in Figures 4.8 and 4.7 may not appear encouraging at first, they indicate the potential improvement in profile predictions that can be achieved if the missing values can be accurately predicted by the CRF model. These improvements can be achieved by re-evaluating either the pre-processing phase (particularly the clustering of profiles) or the CRF feature functions.

4.7 **Conclusion**

In the Benguela upwelling system the processes responsible for the vertical structures are complex and vary from day to day, forced primarily by fluctuations in wind that result from daily weather systems. In this chapter, algorithms that are capable of extracting patterns of the evolution of these processes have been explored with the purpose of relating these patterns to sub-surface structures. Presently, it is not clear from the literature precisely how these structures develop, with regards to factors such as wind direction, duration or strength, thus it is important to consider algorithms capable of extracting or mining all useful patterns. These algorithms have been tested on their accuracy at predicting profiles using these patterns for new data.

The classify-by-sequence (CBS) approach selected and tested indicates both the simplicity of pattern extractions and the strong potential accuracy in predicting profiles. The patterns in the processes and their relation to a particular water column structure can be based on any hypothesis of the user and can be numerous. The user is then able to make a decision on how much evidence should be available for a pattern to be selected. The patterns, which may represent complex processes, are easy to interpret as their format remains unchanged from the input format.

This is a supervised training approach and so the algorithms require samples of patterns and the related profile. *In situ* profile samples are relatively few compared to the potential wind, SST and Chl *a* sequence scenarios, which increase exponentially with the length of the time sequences. The majority of satellite SST and Chl *a* sequences contain missing data due to cloud cover, which limits the training data. Therefore a similar pattern mining approach is investigated that is capable of estimating the missing data by using information that may or may not influence the state of the variable. Although simple interpolation approaches produce good results, they cannot be substantiated by real forcing processes and are thus limited, particularly where consecutive data gaps are large. The conditional random field model considers both the transition of state probabilities and the degree of influence any available observations may have on the potential state and transition to that state based upon hypotheses concerning the underlying processes. The results show that filling in the missing values can produce significant improvements in the accuracy of the CBS model. This is particularly evident with the temperature profiles that show an improvement in accuracy, which will be larger if including predictions of profiles with similar structure. The improvement is also observed in the predictions of low-biomass Chl *a* profiles, which is attributed to the

accuracy of the conditional random field model in predicting missing low Chl *a* satellite data.

By using all available information, both static and dynamic, that influences the daily physical structures and biological responses in the Benguela upwelling system, a model has been developed that is able to capture the complex non-linear interactions in three dimensions at the event scale. This provides a comprehensive tool capable of providing detailed information on the dynamics of the system.

Chapter 5 - Estimation of Daily Primary Production

5.1 Introduction

In Chapter 4 a model was developed to predict the most likely temperature and chlorophyll *a* (Chl *a*) profiles for a given pixel of a satellite image, based on sequences of satellite-derived surface data; wind, sea surface temperature (SST) and Chl *a*. The model produces the most likely subsurface Chl *a* structure that is required for a more accurate estimation of primary production (Platt and Sathyendranath, 1988). Here, an operational model that utilizes this information and other important processes is assembled and tested.

Bio-optical models that estimate primary production are generally described according to how they resolve the biomass distribution $B(z)$, irradiance $I(z)$ (light transmission through the water column and absorption by phytoplankton), depth and time (Kwewalyanga *et al*., 1992; Behrenfeld and Falkowski, 1997a). It has been demonstrated that models that resolve the biomass profile perform better than those assuming a uniform profile (Platt *et al*., 1991) and models with parameters that are dependent on the wavelength of light, produce better results than those that average over the visible light spectrum (Kwewalyanga *et al*., 1992).

Depth resolved primary production $P(z)$ can be calculated according to ;

$$P(z) = B(z)P^B(z) \qquad (5.1)$$

where P^B is the normalized productivity (the rate of carbon fixation normalized to photosynthetic pigment biomass B) at depth z (Platt and Sathyendranath, 1993). The normalized productivity, P^B depends upon irradiance of the photosynthetically

active radiation (PAR) which is also a function of depth. Irradiance at depth, $I(z)$ is calculated as a decreasing exponential function of the available light at the surface according to Beer-Lambert Law (Gordon, 1989);

$$I(z) = I(0)e^{-Kz} \qquad (5.2)$$

where $I(0)$ is the surface irradiance and K is the attenuation coefficient (Platt *et al*. 1994).

The relationship between P^B and I can be described by two parameters α^B and P_m^B also normalized to biomass in the form (Platt and Sathyendranath, 1993);

$$P^B(z) = p^B(I(z); \alpha^B, P_m^B) \qquad (5.3) \ .$$

The two photosynthetic parameters are the initial light-limited slope α^B, which describes the reaction of the phytoplankton to light (below light saturation) and the maximum photosynthetic rate or assimilation number P_m^B (at light saturation). The explicit function of the relationship of photosynthesis to light or the photosynthesis-irradiance (P-I) curve (Eq. 5.3, right-hand side), has numerous forms for example, Smith (1936), Tamiya (1951), Jassby and Platt (1976), Platt *et al*. (1980). Both the initial slope and available light field at depth are dependent on wavelength as they are based on the absorption by phytoplankton (Bricaud and Stramski, 1990). The broadband models assume that all PAR absorbed by phytoplankton is done with the same efficiency and that the light transmission and the photosynthetic response of phytoplankton to light are wavelength-independent. Kyewalyanga *et al*. (1992) showed that in the oligotrophic North Atlantic Ocean, a model which excludes the wavelength dependencies produced an average coefficient of determination of 60%

whereas a spectral model was able to explain at least 90% of the variation. However, including the dependence of $I(z)$ on the spectral attenuation of surface PAR and the photosynthetic parameters on the spectral distribution of PAR necessitates more complex algorithms and experiments.

To obtain daily water column primary production, the length of the day (DL) and the subsurface irradiance field to the 1% light level throughout the day are required. Primary production is then integrated over depth and time.

5.2 **Methods**

5.2.1 Photosynthesis parameters

The two photosynthesis parameters α^B and P_m^B are obtained from simulated *in situ* P-I experiments usually conducted at sea. Few P-I experiments have been conducted in the southern Benguela upwelling system. Mitchell-Innes *et al*. (2000) and Lamont (2011) conducted non-spectral artificial light experiments in the St Helena Bay region, the former from a biological perspective in February 1996 and the latter from a physical perspective in October 2006 and May 2007. Both Mitchell-Innes *et al*. (2000) and Lamont (2011) used the model of Platt *et al*. (1980), which includes a photoinhibition parameter (β^B) to derive the model parameters from incubations and to estimate daily integrated primary production. The results from their photosynthesis-irradiance experiments are summarized in Table 5.1. Mitchell-Innes *et al*. (2000) observed diurnal variability in the parameters and variability as a result of diatom or dinoflagellate-dominated assemblages. Lamont (2011) found variability that corresponded to physical variability of the environment. Although it is possible to relate the parameters to variables observed via remote sensors, such as

the dependency of rate processes on SST and Chl *a* (Balch and Byrne, 1994; Behrenfeld and Falkowski, 1997b), this is considered beyond the scope of the model developed here. Instead a mean of the available experimental data is used. The means obtained are similar to means observed in the comparable Chilean upwelling system (α^B= 0.021 and P_m^B= 2.84) where the parameters were also highly variable under different environmental conditions (Montecino *et al.,* 2004).

Table 5.1 *In situ* photosynthesis parameters obtained in the southern Benguela upwelling system.

Authors experiment	Date	P_m^B mgC [mgChl]$^{-1}$ h^{-1}	α^B mgC [mgChl]$^{-1}$ h^{-1} [μmol m^{-2} s^{-1}]$^{-1}$	β^B mgC [mgChl]$^{-1}$ h^{-1} [μmol m^{-2} s^{-1}]$^{-1}$
Mitchell-Innes *et al*. (2000) Diatom	**20 Feb '96**	**mean 4.75 range 4.02 – 5.69**	**mean 0.022 range 0.02 – 0.025**	**mean 0.00022 range -0.00016 – 0.00056**
Dinoflagellate	**28 Feb '96**	**mean 2.94 range 1.86 – 4.21**	**mean 0.019 range 0.014-0.025**	**mean 0.00017 range -0.00001 – 0.00064**
Lamont (2011)	**Oct '06**	**mean 2.81**	**mean 0.045**	**-**
	May '07	**mean 4.41**	**mean 0.038**	**-**
Combined		**range 0.18-10.22**	**range 0.01-0.09**	**-**
Overall Mean		**3.73**	**0.031**	**0.00020**

5.2.2 Attenuation of surface PAR

Irradiance at depth *z* is calculated as a decreasing exponential function of the surface irradiance $I(0)$. It is a wavelength integrated measure of the band 400-700 nm, which approximates the PAR. Irradiance as a function of depth *z* therefore

requires the attenuation coefficient K_{PAR}. Satellites however, routinely estimate the attenuation of the 490 nm band (K_{490}) as a function of the ratio of blue to green wavelengths of the water leaving radiances (Austin and Petzold, 1981; Mueller, 2000). Available MODIS data include estimates of average daily PAR, instantaneous PAR (iPAR) at the surface and K_{490}. Previous studies have investigated the relationship between K_{490} and K_{PAR} (Barnard *et al*., 1999; Zaneveld and Kitchen, 1993; Morel *et al.,* 2007; Pierson *et al*., 2008) while others have attempted to estimate K_{PAR} from Chl *a* concentrations (Riley, 1956; Morel, 1988; Platt and Sathyendranath, 1988; Morel and Moritorena, 2001). Both approaches have shown successful results. However, methods investigating relationships between K_{490} and K_{PAR} are based on empirical data used to obtain K_{490} and K_{PAR}. Similarly, models incorporating Chl *a* depend on empirical data from either *in situ* samples or as are used to derive the Chl *a* algorithms from remote sensing reflectance (O'Reilly, 2000), and further empirical data to obtain K_{PAR} from the Chl *a* concentration. As a result of the empirical investigations many of these models have been either regionally specific or specific to clear or turbid waters only. Lee *et al*. (2005) developed a semi-analytical model for determining the attenuation coefficient at any wavelength from remote sensing absorption and backscattering coefficients of water at 490 nm. Lee *et al*. (2007) were able to show that K_{PAR} estimated through this approach and used to calculate Z_{eu}, produced better results than those which derived Z_{eu} from remote-sensing Chl *a*. The error of the latter was attributed to the error of the satellite-derived Chl *a* (78.2%). They found the empirical algorithms (Morel, 1988; Morel and Maritorena, 2001) tested on a broad range of conditions

(measured K_{490} ranged from 0.04-4.0 m^{-1}) significantly underestimated measured values > 0.2 m^{-1} whereas their method produced good results across the range. Their model requires the absorption and backscattering (two optical properties) at 490 nm from satellite data and is thus reliant on accurate estimations of these two parameters. Application of their model to obtain the attenuation coefficient of visible light is not used here as it is not yet a routinely produced satellite product. Further, the dependence of some of the empirical models on *in situ* measurements rather than remotely sensed data and the poor performance of the remote-sensing data when used, points toward a simpler approach.

The algorithm of Riley (1956) is dependent on the *in situ* Chl *a* concentration. It depends on the distribution of Chl *a* and assumes that the attenuation of PAR is not influenced by suspended matter unrelated to phytoplankton. It has been used by Carr (2002) and Demarcq *et al*. (2008) for estimates of primary production in the Benguela. The algorithm is given as;

$$K_{PAR} = K_w + 0.0088 * Chl + 0.054 * Chl^{0.66} \qquad (5.4)$$

where K_w is the attenuation of PAR in pure seawater and is equal to 0.04 and *Chl* is the concentration of Chl *a* in the water column layer. The algorithm was intended to cover a wide range of Chl *a* values to the depth of the thermocline or within the euphotic layer where light extinction is only controlled by phytoplankton, organic matter and the water itself (Riley, 1956). As the average attenuation of light, according to Beer's Law, is governed by the average concentration of the attenuating substance, Equation 5.4 can be applied to any layer of the water column. In the work by Riley (1956) and Demarcq *et al*. (2008) the Chl *a* biomass in a

predetermined layer is used to obtain the average K_{PAR} whereas Carr (2002) used remotely sensed surface Chl *a*. However, these approaches do not consider the problem of self-shading where near-surface peaks may cause a rapid attenuation of light through the peak (Steele and Henderson, 1981). Averaging K_{PAR} over the entire euphotic layer may then lead to an overestimation of irradiance below the peak. Similarly a deep Chl *a* peak may underestimate the irradiance above the peak. To address this, K_{PAR} is calculated at 1 m depth intervals from the biomass profile using the mean Chl *a* concentration from the surface to that depth. This approach is compared against the *in situ* data obtained during three cruises in the Benguela.

5.2.3 Depth of the euphotic zone

The depth of the euphotic zone (Z_{eu}) is usually considered as the depth where the irradiance at the surface is reduced to 1% of its surface value. This is a somewhat arbitrary approach as it indicates a dependence on relative irradiance rather than an absolute value. However, at the euphotic depth the PAR is low and slight variations in the absolute value are likely to have a minimal effect on overall production. Assuming a reliable K_{PAR} for the euphotic layer, the attenuation equation (Eq. 5.2) can be used to obtain Z_{eu}. Given that $z = Z_{eu}$ and $I_{z_{eu}} = 0.01 * I_0$, to estimate the 1% light level the following is obtained:

$$\ln(0.01) = -K_{PAR} z_{eu}$$
$$z_{eu} = {}^{4.6}/_{K_{PAR}} \quad (5.5)$$

However, reliable estimates of K_{PAR} for the Benguela are not readily available. Morel (1988) and Morel and Berthon (1989) derived a model that determines Z_{eu} from the total Chl *a* in the water column. This method computes Z_{eu} for each depth interval using the integrated Chl *a* to that depth, until Z_{eu} becomes deeper than that used for integrating the profile. Their equation for Z_{eu} shallower than 102 m is;

$$Z_{eu} = 568.2 * chl^{-0.746}. \quad (5.6)$$

Lee *et al*. (2007) following the approach of Morel and Berthon (1989) developed a similar model based on the surface Chl *a* derived from the SeaWiFS OC4v4 algorithm (O'Reilly *et al.,* 2000). However, the large uncertainties inherent in using satellite derived Chl *a* in highly productive waters and the different algorithms used between the SeaWiFS and MODIS products detracts from their approach.

Here two methods are compared; K_{PAR} is calculated at 1 m depth intervals using the method of Riley (1956) and Z_{eu} is obtained simply by finding the depth where the calculated irradiance is closest to 1% of the surface irradiance; and the model of Morel (1988), which derives Z_{eu} directly from the Chl *a* profile. Both approaches are compared to *in situ* data obtained during the Benguela Calibration (BENCAL; 2002) and Lamont (2011) cruises.

5.2.4 Daily variability of PAR

Instantaneous surface PAR is available for all satellite pixels regardless of cloud. This is because total atmospheric absorption can be divided into cloud and clear sky characteristics (water vapour above clouds and liquid water within them). As cloud liquid water does not absorb visible light and atmospheric absorption is small

(Gautier, 1995), downward irradiance from the base of the cloud is simply the difference between the downward solar irradiance and upward irradiance from the cloud or albedo (Gautier *et al.,* 1980). Time varying PAR is more useful for primary production calculations as it accounts for the variable irradiance due to the solar zenith angle. For example, Mitchell-Innes *et al*. (2000) observed saturation light intensities between 11h00 and 16h00 on clear sky days. Hourly estimates of PAR can be obtained quite easily from the daily integrated PAR if it is assumed that solar irradiance follows a sine curve of the form;

$$I(0,t) = I_0^m \sin {}^{\pi t}/_{DL} \qquad (5.7)$$

where *t* is the time since sunrise, I_0^m is the maximum surface irradiance at noon and DL is the number of hours from sunrise to sunset. The integral of Equation 5.7 is equal to the daily average PAR provided by MODIS and thus I_0^m can be solved. Average day lengths for each month of the year are available from the Astronomical Applications Dept. US Naval Observatory, Washington DC through their website http://aa.usno.navy.mil/cgi-bin/aa_durtablew.pl. From the curve of daily irradiance hourly estimates of $I(0)$ are obtained and $I(z)$ is computed at 1 m depth intervals for each two-hour interval of daylight using the attenuation coefficients derived from the two methods discussed above for comparison.

5.2.5 Primary production model

Kyewalyanga *et al*. (1992) compared two alternative primary production equations; the first by Smith (1936), which does not account for photoinhibition and the second by Platt *et al*. (1980). The model of Smith (1936) is given as;

$$P^B(z) = \frac{P_m^B[{}^{I(z)}/_{I_k(z)}]}{\sqrt{1 + [{}^{I(z)}/_{I_k(z)}]^2}} \tag{5.8}$$

where $I_k = P_m^B/\alpha^B$. Substituting for I_k and because $P(z) = B(z) * P^B(z)$ Equation 5.9 becomes;

$$P(z) = B(z) * \frac{\alpha^B * I(z)}{\sqrt{1 + \left[{}^{\alpha^B}/_{P_m^B * I(z)}\right]^2}} \tag{5.9}$$

The model of Platt *et al*. (1980) is given as;

$$P(z) = B(z) * P_m^B * \left(1 - \exp\left({}^{-\alpha^B * I(z)}/_{P_m^B}\right)\right) * \exp({}^{-\beta^B * I(z)}/_{P_m^B}) \tag{5.10}$$

As for K_{PAR} the two models are tested against the *in situ* daily primary production data obtained from the BENCAL 2002 and the Lamont (2011) cruises. All the relevant equations were coded in the MATLAB® programming language. To obtain the biomass profile needed to test the the skill of the production model the recorded profiles during the cruise are compared to the mean Chl *a* profiles obtained from the *k*-means clustering (Fig. 2.6) and the closest match is used as the representative profile. Unfortunately, no subsurface biomass distribution was presented by Mitchell-Innes *et al*. (2000) to compare primary production. The calculations of the

photosynthesis parameters during the experiments were based on the non-spectral model of Platt *et al*. (1980) and hence it is expected that the model values should reflect those of the *in situ* observations.

5.3 Results and discussion

5.3.1 Attenuation of PAR

The attenuation of light, using the method adapted from Riley (1956), is tested. Figure 5.1 illustrates the relationship of this method with concurrent *in situ* measurements obtained from the West Coast in October 2007 and May 2008. The solid line is the line of unity with a slope of 1 and a y intercept of zero. The dashed line shows the orthogonal regression. The correlation coefficient (r) shows a fairly strong positive relationship (r = 0.73).

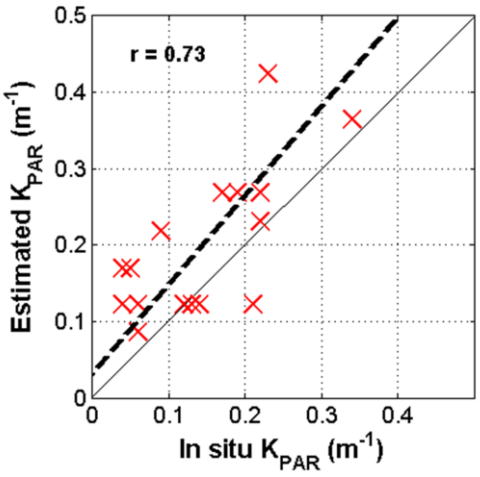


Figure 5.1 Comparison of estimated K_{PAR} using the method of Riley (1956), to *in situ* measurements, where the estimated K_{PAR} is the average K_{PAR} of the euphotic zone. The dashed line shows the orthogonal regression and the solid line indicates the 1:1 relationship. The correlation coefficient (r) is also shown.

The best fit indicates a tendency to overestimate the *in situ* K_{PAR} data across the range of values. The variability of the data in part can be explained by the simplification of the observed profiles to the nearest cluster mean. Many of the observed profiles are mapped to the same *k*-means cluster due to the minimal differences to the cluster mean. This produces the same K_{PAR} estimate as shown in Figure 5.1 where a range of values on the x-axis (K_{PAR} derived from observed profiles) correspond to a single value on the y-axis. For example, *in situ* profiles mapped to Profile 2 from the *k*-means set have an estimated attenuation coefficient of 0.12 m^{-1}, which is close to the mean coefficient of the observed data, 0.11 m^{-1}. However, the observed K_{PAR} for profiles mapped to Profile 2 ranges from 0.21-0.04 m^{-1}, which indicates an optical depth ranging from 4.8-25 m. One optical depth is the inverse of K_{PAR} and is generally considered to be the depth to which satellite ocean colour sensors "see". Further uncertainty lies within the water column constituents. For example, during an anchor station study within St Helena Bay, Mitchell-Innes and Walker (1991) found the vertical attenuation varied in the upper 12 m depth according to phytoplankton species present. They found specific vertical attenuation coefficients for flagellates, small diatoms and large diatoms to be 0.136, 0.017 and 0.009 m^2.mg.Chl a^{-1} respectively. The phytoplankton absorption coefficient is a function of cell size via pigment packaging and the composition of accessory pigments, both of which may change within a cell in response to a changing environment (Kirk 1976; Morel and Bricauid 1981; Sathyendranath *et al.,* 1987; Bricaud *et al.,* 1995; Fujiki and Taguchi, 2002). For example, the content of photosynthetic pigments may decrease as much as one-half with increasing light intensity (Falkowski and LaRoche 1991). Giles-Guzmán and Alvarez-Borrego (2000)

showed that for low Chl *a* waters the subsurface absorption coefficient could increase by as much as 75% of the surface value. Considering the numerous contributions to the uncertainty in the estimated $K_{PAR,}$ the results obtained in Figure 5.1 are reasonable.

5.3.2 Euphotic depth

MODIS data provide an average daily PAR product derived from the instantaneous PAR (iPAR) measured during the satellite overpass. The iPAR value is not useful to primary production models as it cannot account for the diurnal variability of solar irradiance (Platt and Sathyendranath, 1988; Platt *et al.,* 1991). Instead the daily PAR is assumed to be the integral of a simple sinusoidal function of irradiance that allows hourly computations of iPAR using Equation 5.7. Validating PAR and iPAR for the Benguela is not a trivial matter as it requires continuous *in situ* measurements throughout the day and measurements at the exact time the satellite is overhead. Data of this nature is scarce for the Benguela region. It can be expected that cloud cover variability will play a major role in differences between *in situ* and remote-sensing data as daily variability is not explicitly accounted for in the remote-sensing PAR algorithms.

The attenuation of light is not dependent on PAR and therefore errors in estimation of iPAR will only affect the value at the euphotic depth and not the actual depth. Z_{eu} is calculated from the attenuation of the surface iPAR through successive layer of Chl *a* at 1 m depth intervals until it reaches 1% of its surface value, using the method of Riley (1956). Another approach to estimating Z_{eu} is directly from the Chl *a* profile. The results of these two approaches are shown in Figure 5.2. For both approaches

the Z_{eu} for Chl *a* profiles with large near-surface biomass (indicated by a shallow Z_{eu}) is well represented, although using the method of Morel (1988) slightly overestimates the depth. The goodness-of-fit deteriorates with increasing Z_{eu} with greater error associated with Riley's (1956) method. Both methods substantially overestimate Z_{eu} for profiles with low Chl *a* biomass (indicated by a deeper Z_{eu}). There are numerous sources of the uncertainty observed in the estimate Z_{eu}. Of particular importance is the fixed K_{PAR} estimated using the methods of Riley (1956). However, the method of Morel (1988), which does not use K_{PAR} produces similar results. A possible

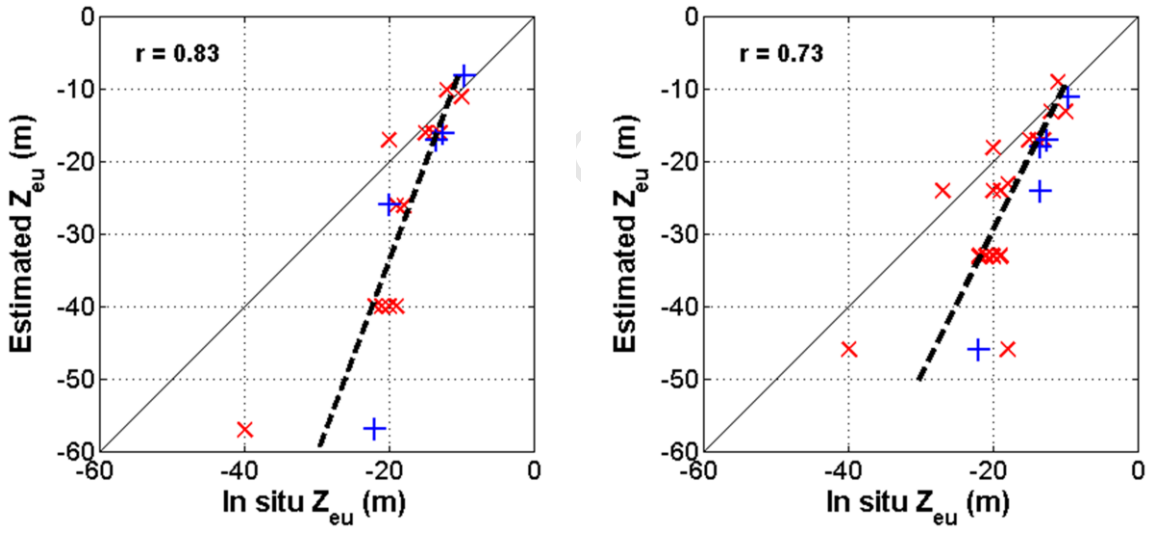


Figure 5.2 Comparison of estimated Z_{eu} (1% euphotic depth) against *in situ* measurements using the method of (a) of Riley (1956) and (b) Morel (1988). The dashed lines indicate the orthogonal regression. The correlation coefficient (r) is also shown. Blue +'s indicates the data from the BENCAL cruise and red x's indicate the data from the Lamont (2011) experiments Fewer samples in (a) result from the missing *PAR* data that are not required by (b).

explanation that affects both approaches is that for a water column with low phytoplankton biomass, the role of other substances in the absorption of light becomes more significant. This might explain why, in those cases of low biomass profiles, the limit of observed Z_{eu} is at ca. 20 m depth. These absorbing constituents in the water column may have been absent when Z_{eu} was observed at ca. 40 m (Fig. 5.2), possibly as a result of the sample location. Further, Gordon (1989) and Liu *et al*. (2002) have reported that for a given homogenous biomass layer the attenuation coefficient will not be constant with depth.

It is worth noting that, as discussed in Section 5.3.1, the *in situ* calculated K_{PAR} for profiles mapped to the same *k*-means category, displayed a broad range of values, yet the observed Z_{eu} for these profiles is quite similar (Fig. 5.2; same category profiles are indicated where estimated Z_{eu} is constant). This suggests that uncertainty may also lie with the *in situ* measurements.

5.3.3 Validation of daily primary production

The two primary production equations that use the light attenuation coefficients obtained from the biomass profiles according to Riley (1956), are compared with the *in situ* measurements. Figure 5.3 shows the results of the primary production models of Smith (1936) and Platt *et al*. (1980) compared to *in situ* experiments. Both methods show good correlation with the *in situ* data (r = 0.74 and r = 0.76 for Smith, 1936 and Platt *et al*., 1980, respectively). However, the model of Platt *et al*. (1980) underestimates the *in situ* values whereas the equation of Smith (1936) shows a closer correspondence with the *in situ* values. Errors produced when estimating the hourly surface irradiance, the computation of Z_{eu} and the averaging of the

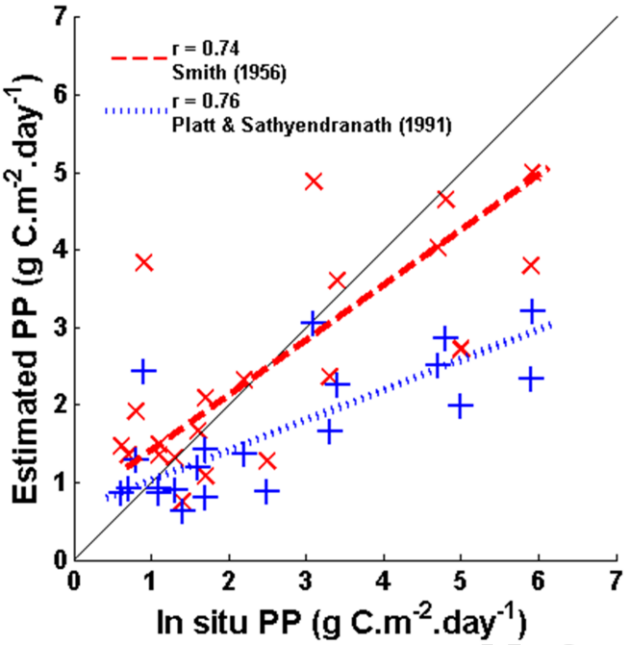


Figure 5.3 Comparison of two bio-optical algorithms for estimating daily depth-integrated primary productivity againt *in situ* measurements. Dashed lines indicate the orthogonal regressions. The correlation coefficient (*r*) is also shown.

photosynthetic parameters are all likely to play a role in the error of estimated primary production. For example, where biomass is large the euphotic depth is well estimated (see Fig. 5.2a) but both primary production models underestimate primary production.Similarly,for low biomass in the water column the deep euphotic depths are over estimated by Riley's (1956) model and the primary production models produce more accurate estimates of primary production. Furthermore, the photosynthetic parameters vary according to species; diatoms generally have higher photosynthetic rates than smaller flagellates and picoplankton (Claustre *et al.,* 2005; Uitz *et al.,* 2008), and a number of physiological and environmental factors (Falkowski *et al.,* 1981; Platt et al., 1992; Côté and Platt 1984; Sathyendranath *et al.,* 1995; Kyewalyanga *et al.,* 1998; Falkowski *et al.,* 1998). Table 5.1 shows a

maximum range across two orders of magnitude for the maximum photosynthetic rate and one order of magnitude for the assimilation number. P_m^B is known to increase from morning to early afternoon as photoadaptation to the increasing irradiance occurs (Cullen *et al.,* 1992; Mitchell-Innes *et al.,* 2000) and both parameters were observed to vary with depth in the Benguela. For example, α^B and P_m^B vary by 9 and 57 (Table 5.1) respectively (Mitchell-Innes *et al.,* 2000; Lamont, 2011). Falkowski (1981) reported α^B and P_m^B varied by factors of ca. 6 and ca. 25 respectively. Others have reported seasonal variability in coastal upwelling regions (Tilstone *et al.,* 2003; Aguirre-Hernández *et al.,* 2004; Montecino *et al.,* 2004). Behrenfeld and Falkowski (1997a) noted the variability of the biomass field with depth as an input to primary production models contributed the most to the variability of the output.

5.4 **Conclusion**

The objective of this chapter was to obtain a simple model for estimating primary production that can account for the vertical variability of production over a photoperiod of one day. To this end the most important variables to resolve are the diurnal incident PAR and its attenuation with depth. It is necessary, due to computational limits, to obtain the input data in a format that will minimize pre-processing but not sacrifice too much on predictive accuracy. Chl *a* derived from satellites in upwelling regions tends to match *in situ* data reasonably well (Moore *et al*., 2009; Dogliotti *et al.,* 2009). Even if there is a good correlation between satellite and *in situ* surface data in the region it still does not resolve the vertical variability and hence the subsurface light field. These issues are currently being addressed by dedicated researchers and experimental products are available (e.g. MODIS K_{PAR}

using the inherent optical approach of Lee *et al.* 2005). By using the model developed in Chapter 4 it is possible to predict the subsurface distribution of Chl *a* from relationships or patterns with environmental variables. Although still relying on remote-sensing data, the output is less sensitive to systematic errors as it relies on statistical model outputs.

The output of the primary production model confirms the potential of this approach. Considering the variability of many of the input variables (primarily solar irradiance, attenuation of PAR and the photosynthesis parameters) and the application of algorithms derived from multiple data sets, the comparison of modelled primary production with that of the *in situ* determinations is encouraging. If the Chl *a* profile can be predicted accurately and assuming that the *in situ* measurements are accurate, the primary production model will provide an accurate description of the vertical distribution of daily primary production integrated over a day.

Chapter 6 - Time-Series of Depth-Integrated Primary Production

6.1 Introduction

In Chapters 4 and 5 separate models were presented that address specific problems that are encountered when attempting to model primary production at the event scale in a very dynamic upwelling system. Each model has been developed separately in using the MATLAB® programming language and the code produced so that the outputs of one model feed into the next. These individual models are now discussed as modules of a single primary production model.

In Chapter 4 a method was developed to classify concurrent sequences of wind, sea surface temperature (SST) and surface chlorophyll *a* (Chl *a*) according to a set of characteristic depth profiles of either temperature or Chl *a*, where Chl *a* is a proxy for phytoplankton biomass. It was assumed that these sequences provide information on the processes that influence the vertical distribution of phytoplankton. For example, wind sequences provide information on the mixing layer, which is dependent on both the strength and duration of the wind. Sequences of SST provide information on the stratification of the upper water column but also on the age of the water mass since it was upwelled. Surface Chl *a* is expected to provide the most influential data for predicting Chl *a* profiles as it indicates the relative state of phytoplankton growth over the time period in question and also the near-surface biomass of the profile, which is predicted at the end of the sequence. In order to fully utilize these sequences it was necessary to develop a method capable of filling in the missing data in the SST and Chl *a* sequences as accurately as possible. To do this

a conditional random field (CRF) module was developed that evaluates all possible label sequences (where a label indicates a discrete range of values) along a time segment of up to six days, which must include any observed labels. For each missing label in the SST or Chl *a* sequence the CRF model determines the probability of each possible label given a series of information; the label of the previous day, the previous one, two and three day wind conditions, the season, month and the depth. The most likely sequence is then the maximum sum of the probabilities of each possible label sequence given the series of information (the probabilities are in the log domain and are therefore summed). Once all the missing labels were filled in by the CRF module the label sequences could be used by the classify-by-sequence (CBS) module. The CBS module scores each profile category according to individual sub-sequences (of three-day and two-day length) in a six-day sequence prior to the day the profile is to be predicted. Each sub-sequence is scored statistically according to each of the twelve profile clusters from a set of training data. These two modules produce a temperature and Chl *a* profile that are the most likely profiles for each pixel of daily satellite images.

The next step is to obtain estimates of daily primary production. In Chapter 5 a production module was developed that explicitly includes information on the available photosynthetically active radiation (PAR) at the ocean surface, the transmission of light through the water column as a function of the phytoplankton biomass distribution (or implicitly the light absorption by the phytoplankton), the maximum photosynthetic rate and the efficiency of photosynthesis at low light levels. This module produces an estimate of primary production for each meter depth for each two-hour interval of the day.

In this chapter the different modules are integrated into a single primary production model that takes satellite data and other known information as input and produces hourly estimates of primary production from the surface to the 1% light level. In Section 6.2 a brief description is given of the processing steps performed by the integrated model. In Section 6.3 the model outputs are presented and they are discussed in Section 6.4 in terms of both depth-resolved and depth-integrated primary production.

6.2 **Methods**

Depth-resolved primary production is estimated for each two-hour interval for a given time period and region and for a spatial resolution of 4 km. The first step is to fill in all satellite SST and surface Chl *a* missing data in the region from the start date of the query to the end date. For each 16 km^2 pixel in the region a vector of wind, SST and Chl *a* is extracted for the query period. Each observation in the sequence is mapped to a particular predefined category or label; seven labels for SST and Chl *a* and 10, 12 and 16 labels for one-, two- and three-day wind periods respectively, five depth labels and four seasons. The CRF module is applied working iteratively in consecutive six-day increments for the SST and Chl *a* sequences. For each six-day sequence queried the missing data locations are identified and the probability of all possible sequences are calculated using available information on concurrent wind data, month, season and depth (and SST for the Chl *a* sequences). The weights for each feature in the CRF are obtained from training on a random selection of 50 daily data points from each day spanning 2002-2009, for the St Helena Bay region,. The number of possible sequences is constrained by the number of observations. Hence, given the possibility of seven labels for each missing data point the number

of queried sequences may be from 7^1-7^5 (or 7-16807) for one to five missing data points. It the CRF module encounters a sequence with no observations the sequence steps back two days and re-extracts a six-day sequence. This allows the last two data points of the previous increment to be used as the first two observations. The most likely sequence is selected and that sequence replaces the queried label sequence before moving on to the next increment.

Once each pixel in the region has an unbroken sequence of surface data labels the CBS module can be employed. This again, is an iterative process that uses six-day sequences. Here the original form of the wind data (meridional and zonal wind stress) is mapped to one of ten labels and only a single day of wind is mapped instead of the one-, two- and three-day wind sequences used by the CRF module. The CBS module scores the extracted sub-sequences for the query date according to each possible profile. The sequences are subdivided into three-day sequences beginning with five days prior to the event and ending with the three most recent days. A two-day sub-sequence is also evaluated that includes the query day and the previous day. The profile-specific scores of each sub-sequence are obtained by prior training on all recorded profiles and their associated surface data where the missing surface data has been filled in by the CRF model. A sub-sequence will receive no score for a particular profile either if there are no observations of that sub-sequence for the given profile in the training data or there are not enough observations to satisfy a user-defined minimum. The profile that scores the highest by summing the profile-specific scores of each observed sub-sequence is selected for the queried day and location. The process continues until each location in the queried region has a predicted profile for each day.

The final step is to calculate the daily primary production at each location using the predicted profile, PAR and a set of productivity parameters obtained from *in situ* experiments. Daily PAR is obtained from the MODIS sensor and hourly PAR is estimated from this using the method described in Section 5.2.4. The transmission of the hourly surface PAR to the 1 % light level is estimated using the phytoplankton biomass at each 1 m light interval, obtained from the profile, which along with pure seawater, are assumed to be the only substances that absorb and scatter the light. With the available light at each 1 m depth interval, the phytoplankton biomass and the set of photosynthetic parameters, primary production is calculated according to the method of Smith (1936). Daily production is simply the sum of the estimated hourly production at each depth, which is then integrated over depth to obtain carbon production per m^2 per day.

Figure 6.1 illustrates the integrated model with a seven-day sequence extracted from a continuous model-run over three months. The wind shows a weak north-westerly wind (probably the result of a cold atmospheric front) followed by a typical south-westerly then south-easterly wind. The south-easterly wind strengthens over three days. The SST sequence indicates cooling after the passage of the cold front and cooling again during the strongest south-easterly wind. The Chl *a* sequence shows an increase or bloom near the surface after the north-westerly wind and a few days of moderate south-easterly wind. The last six days of each sequence are used to predict the last profile of the profile sequence. This profile is used with the hourly PAR at depth to produce hourly estimates of primary production. The figure shows primary production estimated during the first two-hours of sunlight, at midday and during the last two-hours of

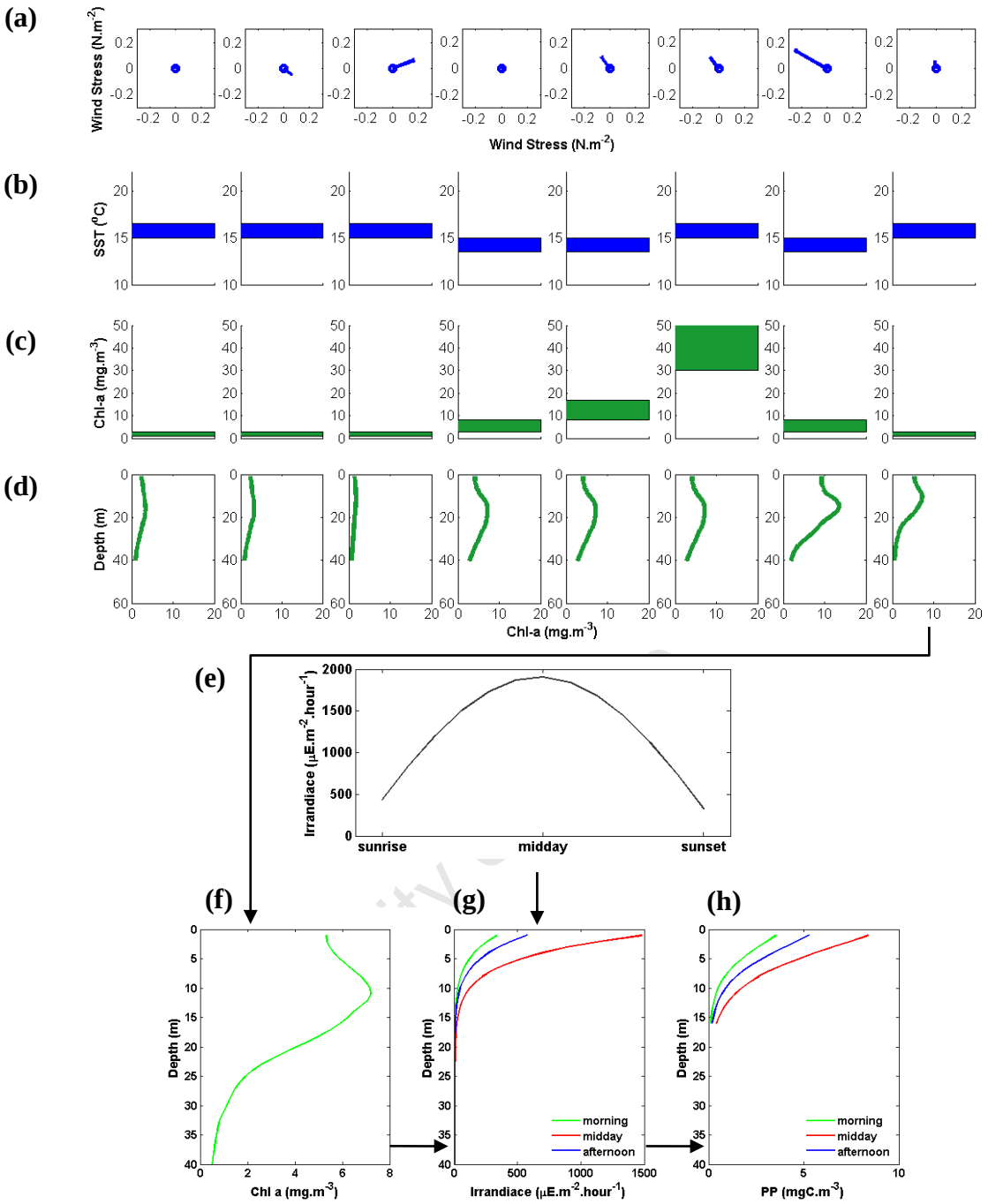


Figure 6.1 An illustration of the model processes: (a) a sequence of wind data (b) SST data and (c) Chl *a* data are used to predict the profiles (d). Hourly (e) PAR is used with the profile (f) to estimate the hourly light field (g) and hourly production (h). The light field and primary production show estimates for the first hour of sunlight, midday and the last hour of sunlight.

sunlight. Initially the model is run over two periods for comparative analysis; firstly for the months of September-October 2006 (austral spring) and secondly from April-June 2007 (austral autumn). The reason for this is that the two periods coincide with the *in situ* experiments conducted by Lamont (2011). The model outputs are further analysed according to depth. Finally, the remaining months between September 2006 and December 2007 are modelled to estimate annual production for the study region.

6.3 **Results**

Figures 6.2 and 6.3 show the *k*-means temperature and Chl *a* profile cluster means and their integrated biomass and near-surface temperature respectively. These profiles are referred to frequently in the rest of this chapter.

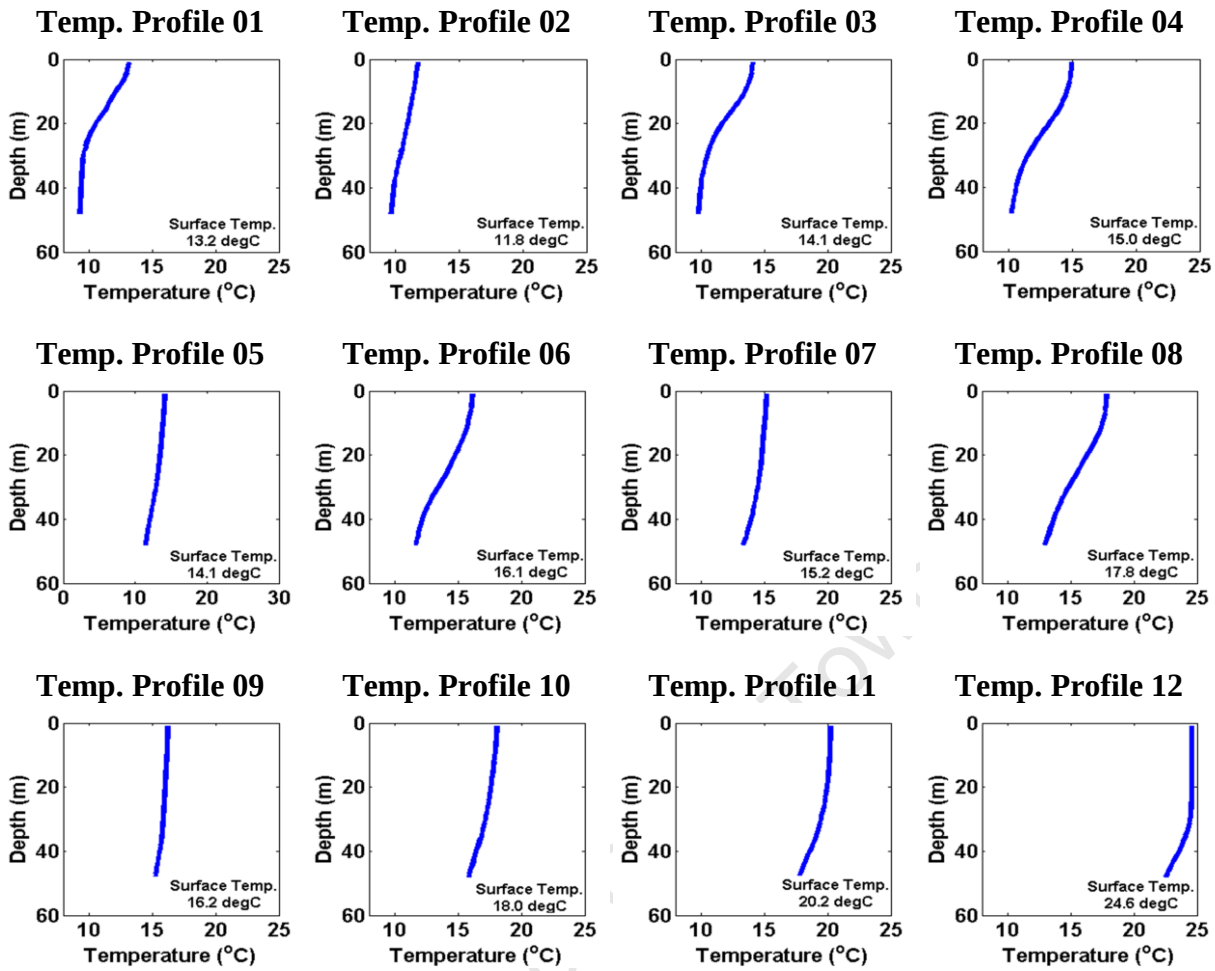


Figure 6.2 Temperature profile cluster means obtained using the *k*-means algorithm from an archive of profiles sampled in the St Helena Bay region between 1988 and 2010. Profiles are arranged according to the average temperature of the profile.

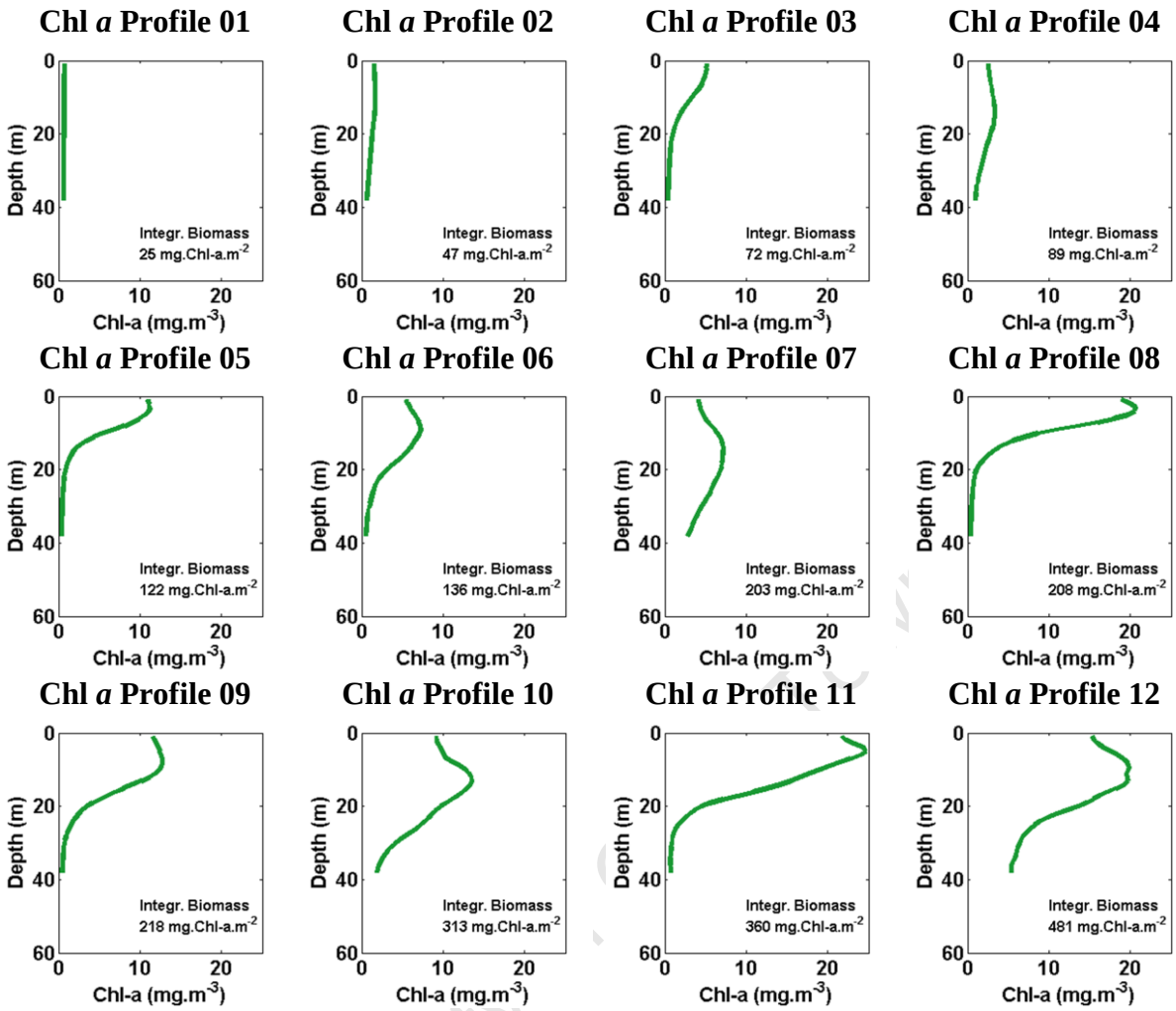


Figure 6.3 Chl *a* profile cluster means obtained using the *k*-means algorithm from an archive of profiles sampled in the St Helena Bay region between 1988 and 2010. Profiles are arranged according to total biomass within the profile.

6.3.1 Frequency distributions of spring profiles

Figures 6.4a-6.4c shows the spring 2006 frequency distributions of the predicted temperature profiles according to months (rows) and depth (columns), where bottom depth is considered a proxy for distance offshore. Bottom depth categories are inshore (< 100 m), inner-shelf (100-199 m) and mid-shelf (200-300 m).

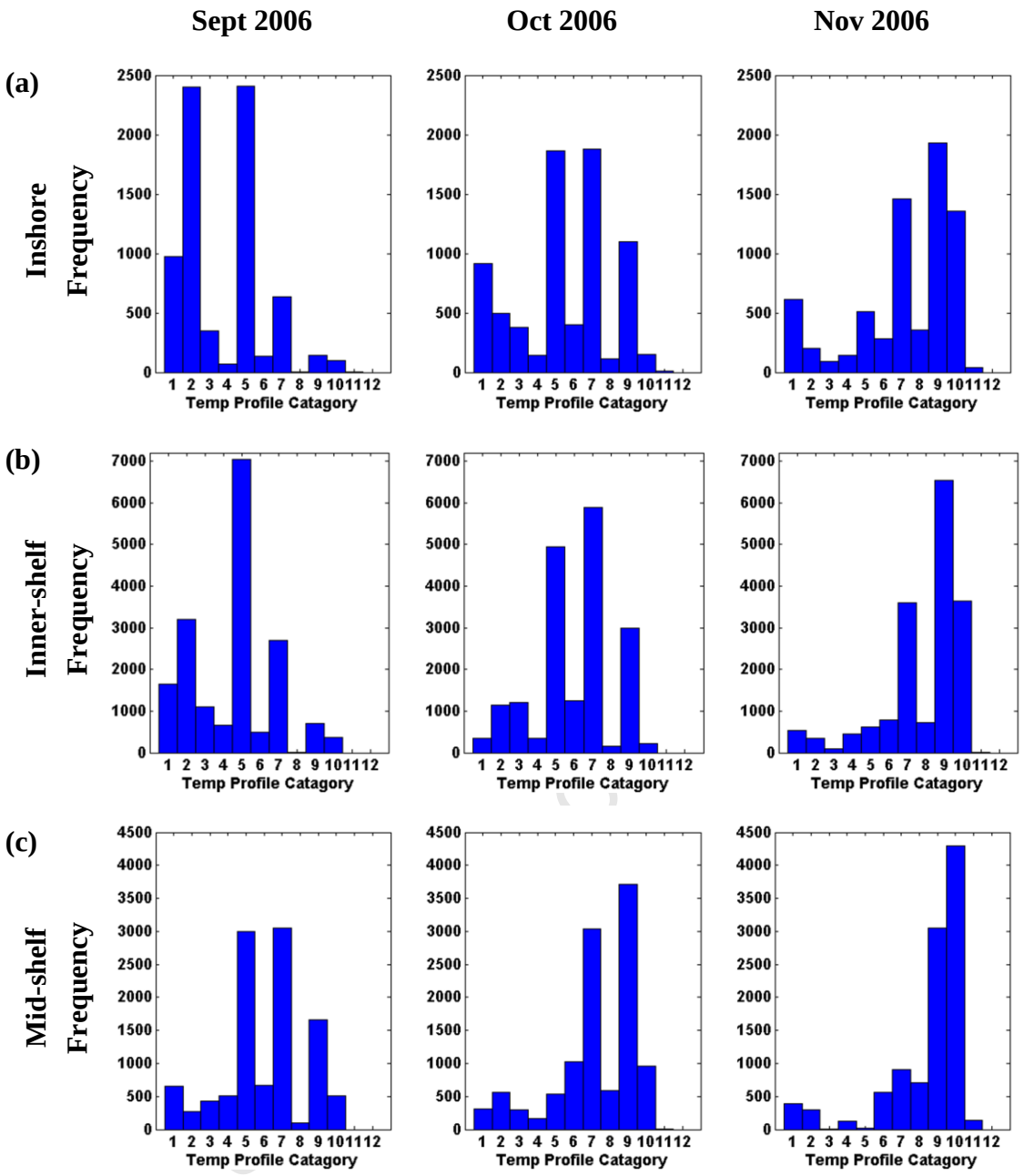


Figure 6.4 Austral spring frequency distributions of the predicted temperature profiles for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay region for September (left column) to November (right column) 2006. Profiles are categorized according to the *k*-means profile clusters (Fig 6.2).

The inshore spring time temperature profiles depicted in Figure 6.4a show a shift in the dominant profile frequencies from Profiles 2 and 5 in September, Profiles 5 and 7

in October and 7, 9 and 10 in November. The mean profiles of these clusters indicate weakly stratified water to depths of around 30-40 m above a more stratified layer in the case of Profiles 5, 7 and 9 (Fig. 6.2). Surface and bottom layer temperatures increase among these profiles from Profile 2-7. Profiles 6 and 8, which have a mixing layer of ca. 10 m (Fig. 6.2) and respective increase in surface temperature, show an increase over this period while Profile 1, which indicates low surface temperature and a mixing layer of ca. 8 m (Fig. 6.2) remains among the most frequent in September and October but decreases slightly in November. Profile 11 is rarely predicted whereas Profile 12 is not predicted at all. These two profiles have the warmest surface temperatures (Fig. 6.2) and would be mostly expected further offshore.

Similar patterns between the inshore and inner-shelf temperature profiles (Fig. 6.4b) are also observed. In September there is a less significant role of Profile 1 and particularly Profile 2 in the frequency distribution and a relative increase in the significance of Profile 7. During October and November the only notable difference to the inshore region is the lower frequencies of Profile 1 and Profile 5 in November. There is a similar succession in the most frequent profiles from September to November from Profiles 2, 5 and 7 to 5, 7 and 9 and finally to 7, 9 and 10, suggesting warming under the influence of increase solar radiation and longer day length according to expectations.

Over the mid-shelf (Fig. 6.4c) there is a change of dominance of Profiles 5, 7 and 9 in September to 7 and 9 in October to a majority of Profiles 9 and 10 in November. Profiles 1 and 2 with the coldest surface temperatures decrease in frequency from the inshore region to relatively insignificant numbers over the mid-shelf, supporting the hypothesis that cool waters indicate upwelling inshore, with surface warming as

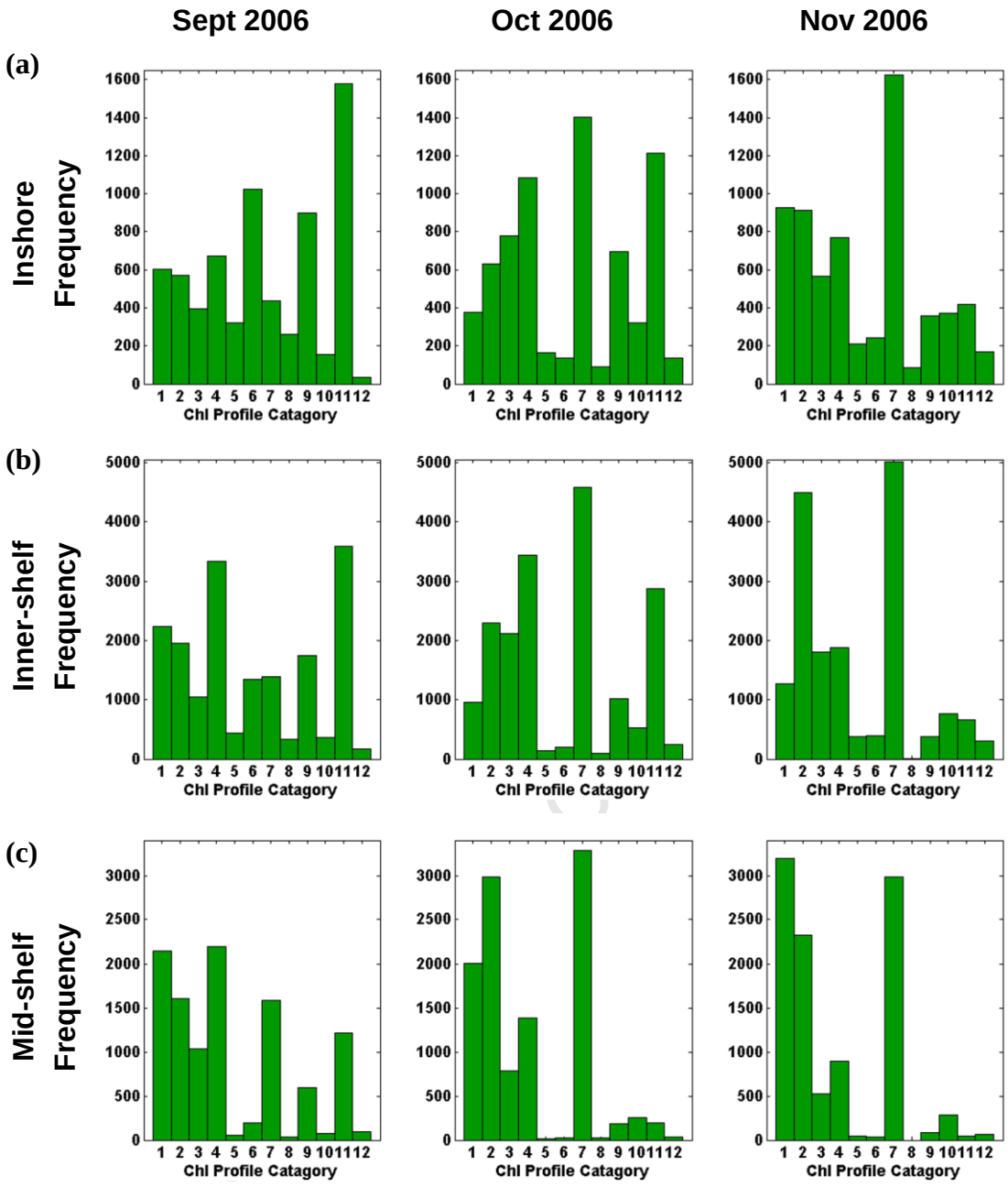


Figure 6.5 Austral spring frequency distributions of the predicted Chl *a* profiles for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay region for September (left column) to November (right column) 2006. Profiles are categorized according to the *k*-means profile clusters (Fig 6.3).

the upwelled water is advected offshore. A surface mixing layer is evident in Profiles 4, 6 and 8 (Fig. 6.2) showing a warming trend as the relative proportions of these three profiles shift.

Figures 6.5a-6.5c shows the spring 2006 Chl *a* profiles. Figure 6.5a indicates that Chl *a* profiles predicted by the model are highly variable over the inshore region from September to November 2006 (column left to right). In September most Chl *a* profiles are well represented with the exception of Profile 12. In particular Profiles 6 and 9, which have peaks at 10 m (Fig. 6.3) and Profile 11 with a shallower peak are frequently predicted. In October Profiles 4 and 7, which have peaks at ca. 20 m (Fig 6.3), and Profile 11 dominate and there is a significant decrease of Profile 6 and similar increase in Profile 7 from September. There is a notable increase in Profile 3 with a small shallow peak (Fig 6.3) and decreases in Profiles 5 and 8 with higher biomass surface peaks (Fig 6.3). By November the predicted frequencies are dominated by the lower biomass Profiles 1-4 with the exception of Profile 7, which is the most frequent. Profile 1 is assumed to indicate upwelling water when occurring inshore (individual profiles in this cluster will have a range of associated surface temperatures). Profiles 9-11 have similar frequencies and Profile 12, which has the highest integrated biomass and a peak at ca. 15 m (Fig 6.3) is most frequently seen in this month.

On the inner-shelf during spring the Chl *a* profile pattern (Fig. 6.5b) is similar to the inshore region but an increase in the frequency of Profile 4 and decrease in Profile 6 indicates a deeper and lower biomass distribution (Fig 6.3). The surface peaks indicated by Profiles 5 and 8 are notably less significant than inshore. In October the pattern is nearly identical to the inshore although Profiles 5, 6, 8, 9 and 10 are less frequent. In November Profile 2, which has low and evenly distributed biomass to ca. 20 m, and Profile 7 dominate the distribution with the remaining profile distribution being similar but lower than inshore, particularly the “upwelling” Profile 1.

Over the mid-shelf region there is a marked change in the frequency distribution of the Chl *a* profiles (Fig. 6.5c). Low biomass Profiles 1-4 collectively dominate the frequency with the exception of Profiles 7 and 11 in September and Profile 7 in October and November. Profile 1 and Profile 2 increasingly outnumber the other profiles from September to November (with the exception of Profile 7). The increase in Profiles 1 and 2 and corresponding decrease in Profiles 3 and 4 suggests lower growth occurred on the mid-shelf as spring progressed. This could occur for example, if the development of strong persistent winds over spring is able to maintain a deep mixing layer.

6.3.2 Frequency distributions of autumn profiles

Figures 6.6a-6.6c shows the temperature profiles for the three bottom depth regions, and for the autumn months of April, May and June (panels left to right). In Figure 6.6a the inshore temperature Profiles 1, 2, 5, 7, 9 and 10 are the most frequent in April. Profiles 1 and 2 indicate cold upwelled water whereas the remaining profiles suggest warmer weakly stratified or well-mixed water (Fig 6.2). In May Profiles 2 and 10 have decreased in number and Profiles 5 and 7 have increased. By June Profile 5 dominates the frequency distribution while Profile 1 has increased and Profiles 7 and 9 have decreased. Profiles 6 and 8, which indicate stronger stratification with a shallow surface mixing layer and Profile 10, which indicates warm very weakly stratified water (Fig. 6.2) are notably infrequent in June.

The inner-shelf temperature profiles (Fig. 6.6b) differ from the inshore pattern in that there are fewer predictions of Profiles 1 and 5 and more of Profile 6 in April, fewer of Profile 1 and more of Profile 10 in May and fewer of Profile 1 in June. Profiles 5, 7 and 9 maintain their relative ratios in May and June.

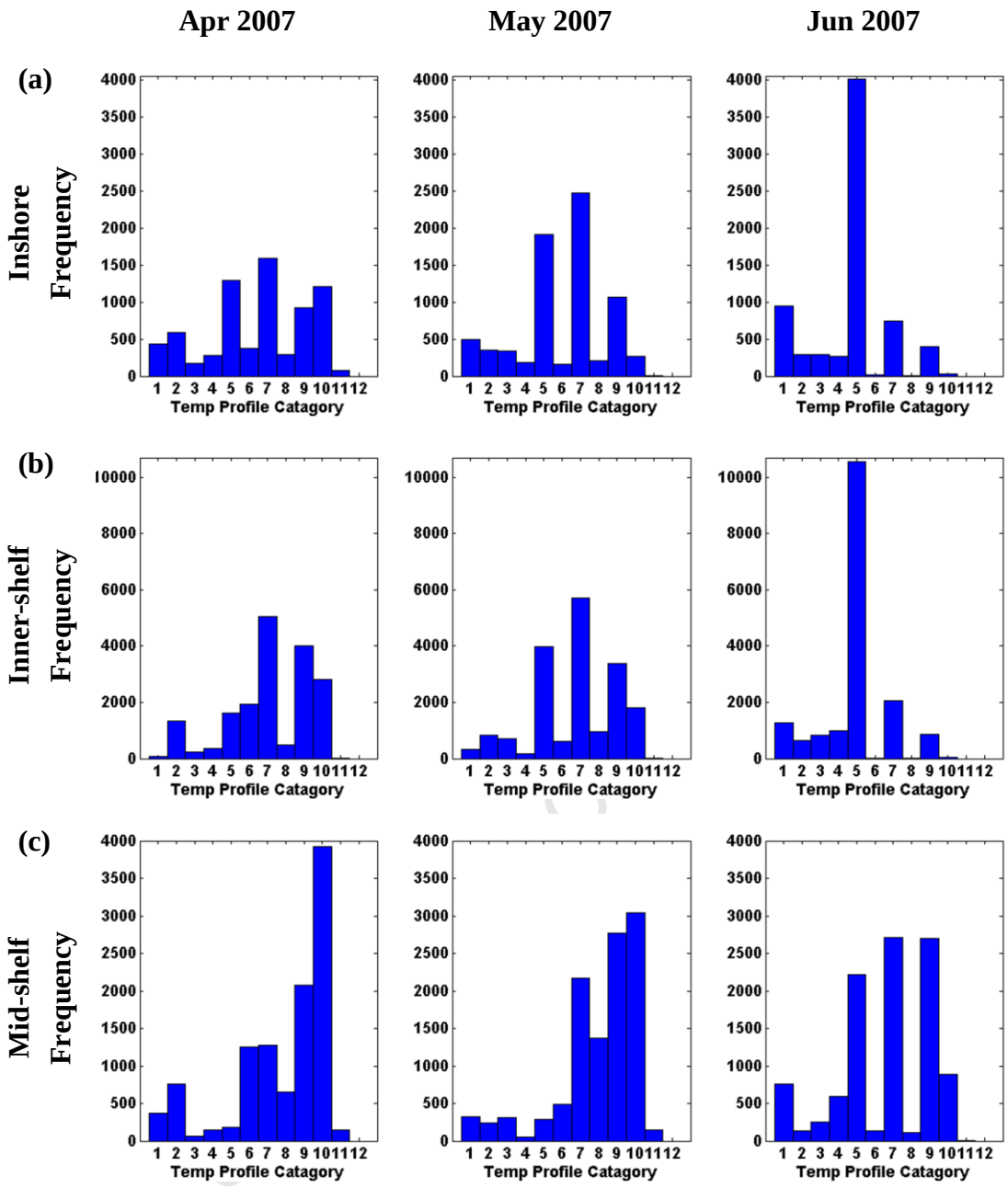


Figure 6.6 Austral autumn frequency distributions of the predicted temperature profiles for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay region for April (left column) to June (right column) 2007. Profiles are categorized according to the *k*-means profile clusters (Fig 6.2).

The temperature profile distribution over the mid-shelf (Fig. 6.6c) shows a decreasing contribution from Profile 10, which is the most frequent profile in April and May. In contrast, from April to May there are increasing contributions from Profiles 7

and 9, and from May to June there are increases in Profiles 5 and 7 from. June also shows higher frequencies of Profiles 1 and 4 that suggest stratification (Fig 6.2).

Figures 6.7a-6.7c shows the autumn Chl *a* profile distributions for the three depth regions. Inshore the Chl *a* profile distribution (Fig. 6.7a) is similar over the three months with regard to the high frequencies of Profiles 2, 4, 7 and 11. They differ in that Profile 3 with its surface peak (Fig 6.3) initially increases from a relatively high frequency in April and then decreases to very infrequent in June. There is also a shift in dominance from Profile 7 in April to no clear dominant profile in May to dominance of Profile 4 in June. The remaining profiles predicted in April tend to decrease towards June.

In Figure 6.7b the inner-shelf is again similar to the inshore region but during April and May Profile 7 dominates more and Profile 11 with a higher integrated biomass and shallower peak than Profile 7 (Fig 6.3) is less significant. In June the relative proportions of the most frequent profiles are practically the same. Profiles 2 and 4, with lower biomass and less defined peaks, increase in May and Profile 4 increases again to dominate in June. Profile 3 also decreases notably on June.

Over the mid-shelf in April (Fig. 6.7c) only Profile 7 and, to a far lesser extent Profile 2 feature in the model outputs. In May and June Profiles 1 and 4 increase as Profile 7 decreases. By June Profiles 1, 2 and 7 all feature with approximately the same frequency with Profiles 4 and 11 contributing to the frequency total.

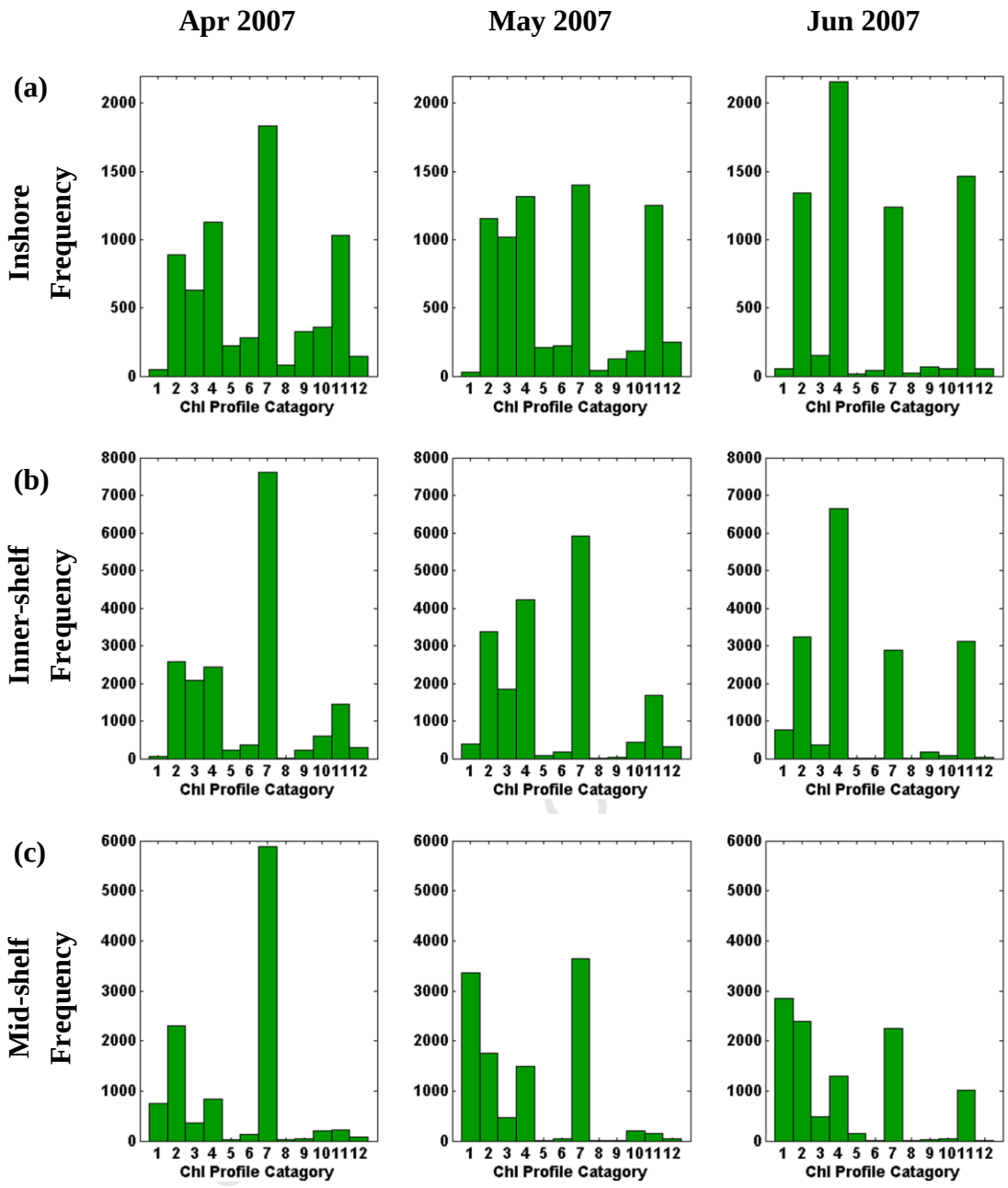


Figure 6.7 Austral autumn frequency distributions of the predicted Chl *a* profiles for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay region for April (left column) to June (right column) 2007. Profiles are categorized according to the *k*-means profile clusters (Fig 6.3).

6.3.3 Depth distribution of primary production

Figures 6.8a-6.8c and 6.9a-6.9c show the daily depth-resolved primary production for three depth regions obtained from the complete primary production model. Note the values are on a log scale and white areas indicate zero values (primary production calculations were not made either because the area is below the 1% light level or there are missing PAR data).

During the spring months of September-October production in the upper 5 m of the inshore (Fig. 6.8a) and inner-shelf (Fig. 6.8b) regions remains relatively constant while deeper in the water column production shows an increases over these months. The upper 5 m also indicates periodic blooms over roughly five days when values of ca. 0.5 $gC.m^{-3}.day^{-1}$ deepen from the surface to 5 m. Over the mid-shelf (Fig. 6.8c) surface layer production is lower although it occasionally does exceed 0.5 $gC.m^{-3}.day^{-1}$. Deeper in the water column production is similar across the shelf but in November the mid-shelf production at depths below ca. 20 m exceeds the inshore and inner-shelf regions.

Over the autumn months of April-June production is shallower than in spring and production values above ca. 0.1 $gC.m^{-3}$ are confined to the upper 10 m whereas in spring these values are often shown below 10 m. Inshore (Fig. 6.9a) and over the inner-shelf (Fig. 6.9b) values above ca. 0.05 gC.m-3.day-1 tend to shallow whereas values lower than ca. 0.05 gC.m-3.day-1 remain relatively constant. Over the mid-shelf (Fig. 6.9c) a period of low production from mid-May marks a change in the depth of production to shallower depths.

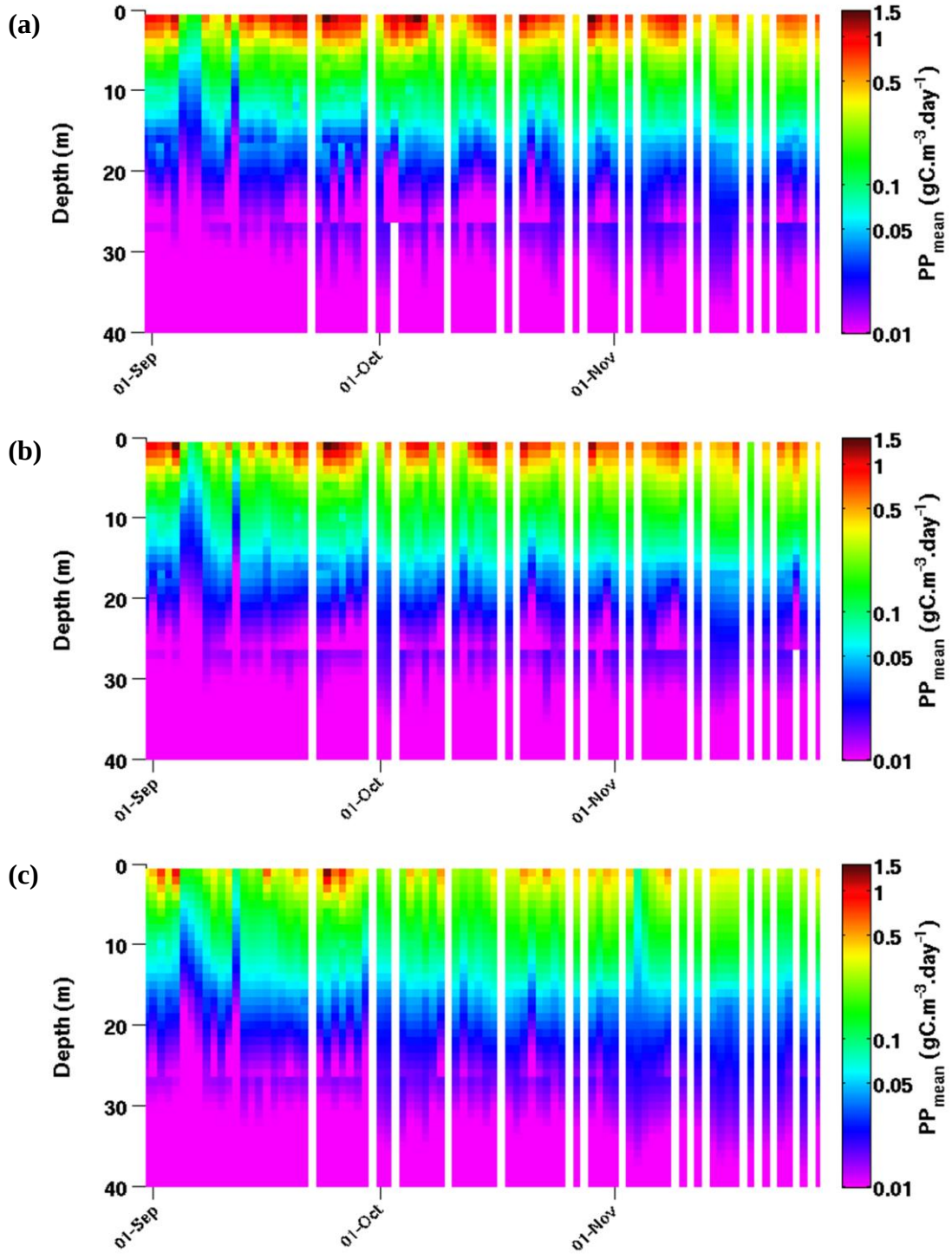


Figure 6.8 Spring time-series of depth-resolved daily primary production for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay region during the period September-November 2006. Note the colour bar is on a log scale.

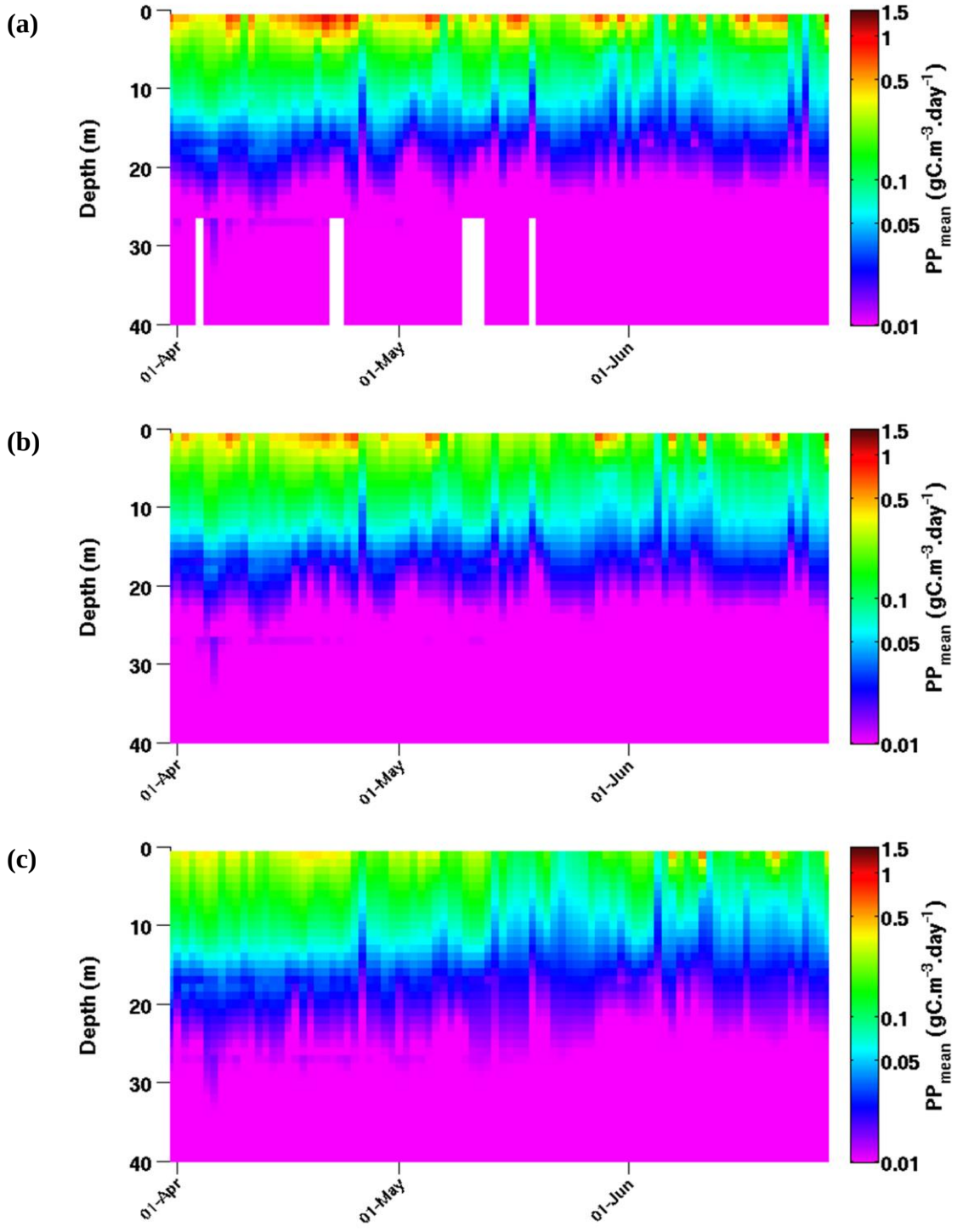


Figure 6.9 Autumn time-series of depth-resolved daily primary production for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay during the period April-June 2007. Note the colour bar is on a log scale.

The variability of the depth-resolved primary production by month, depth and region is shown in Figures 6.10a-6.10c and 6.11a-6.11c for the spring and autumn months, respectively. For spring months there is a decrease in variability from September to November in all regions. Near the surface, for the inshore (Fig. 6.10a) and inner-shelf (Fig. 6.10b) regions, the mean production increases slightly from September to October and decreases in November whereas deeper in the water column there is a small increase. Over the mid-shelf (Fig. 6.10c) production decreases in October and remains constant in November. Inshore in the autumn months (Fig. 6.11a) variability decreases from April to May and is most variable in June whereas the mean production consistently decreases. Over the inner- (Fig. 6.11b) and mid-shelf (Fig. 6.11c) variability remains fairly constant in April and May and increases in June while production decreases most notably from April to May.

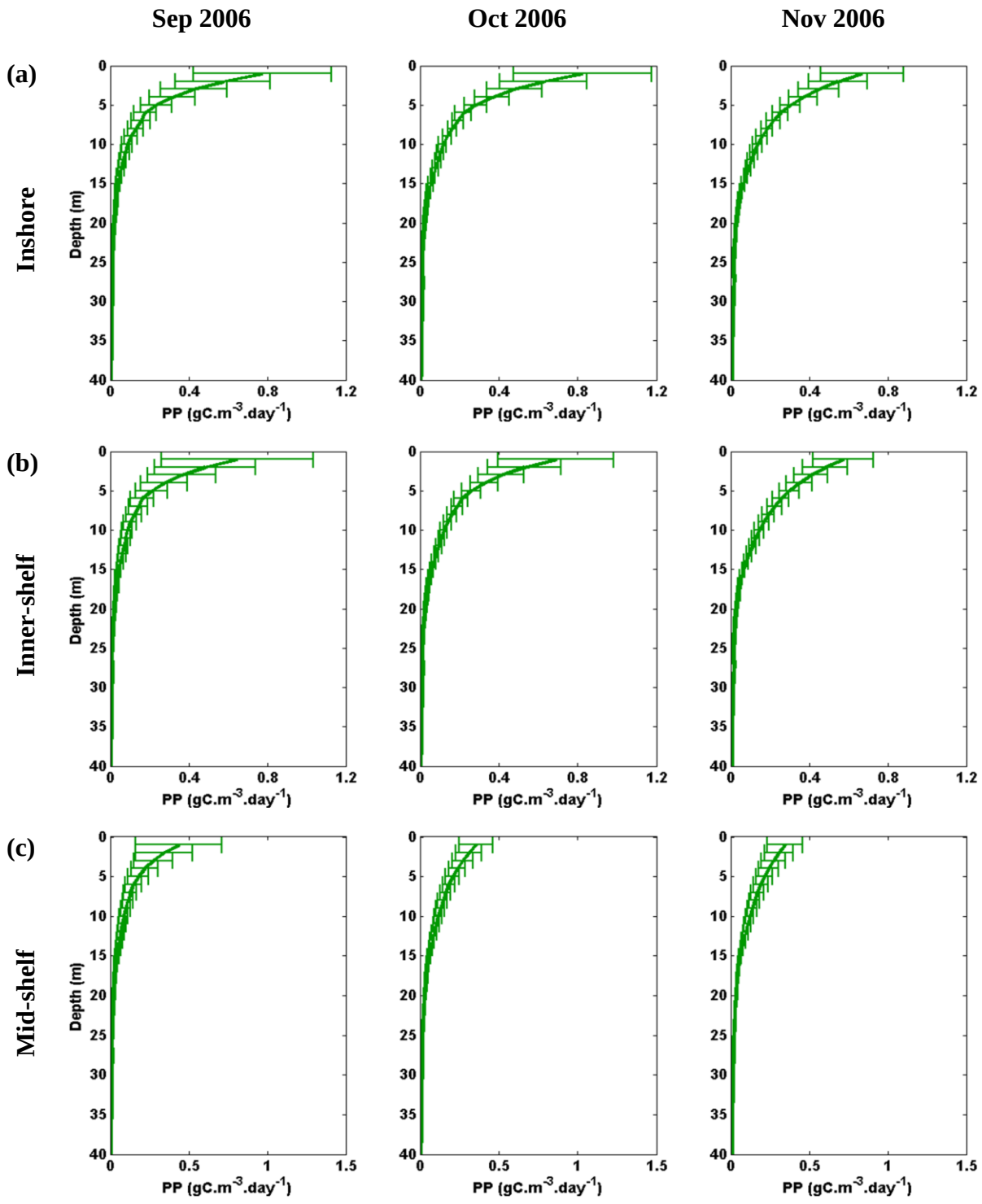


Figure 6.10 Spring production profiles. Mean (solid line) and standard deviation (error bars) of depth resolved daily primary production for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay region for September (left column), October, and November (right column) 2006.

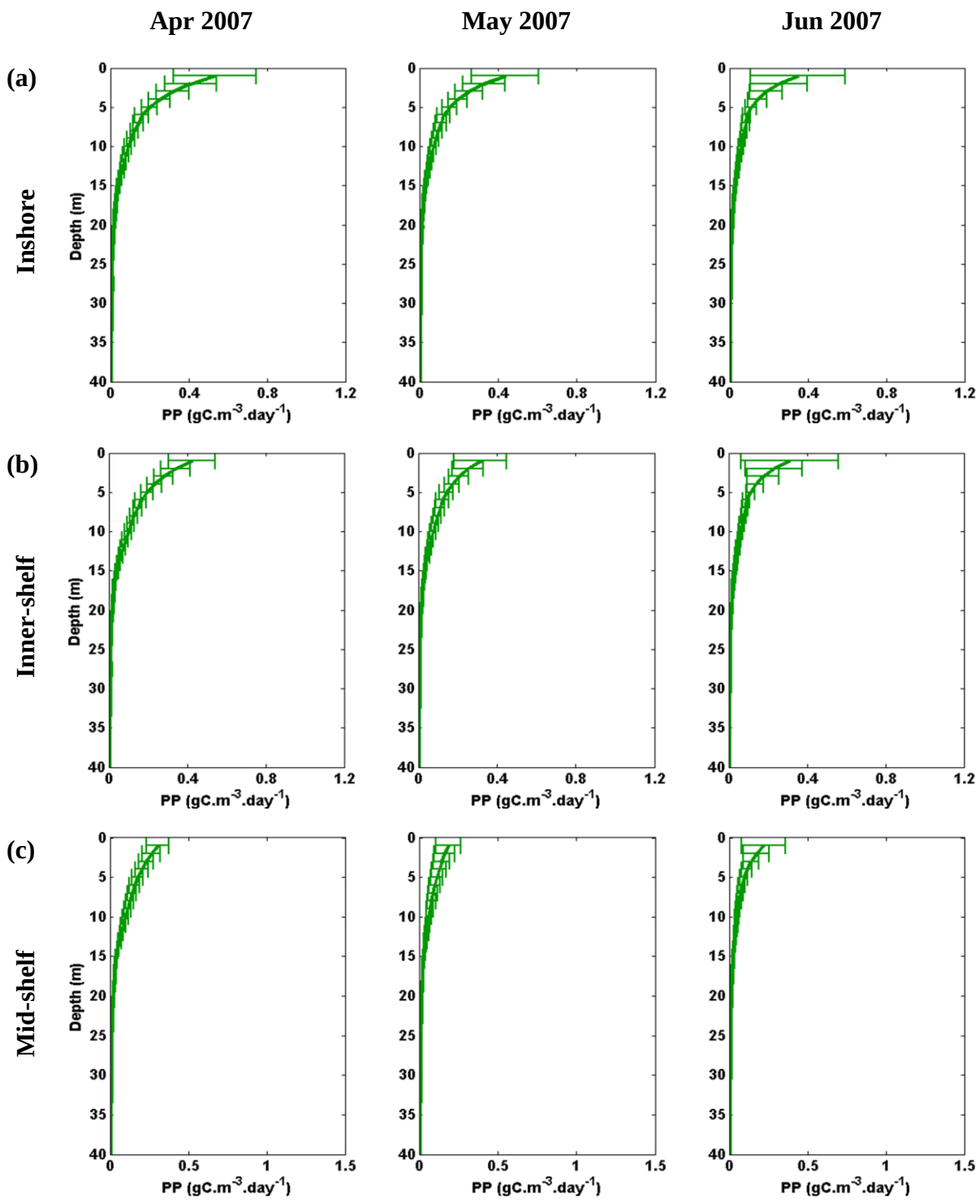


Figure 6.11 Autumn production profiles. Mean (solid line) and standard deviation (error bars) of depth resolved daily primary production for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions of the St Helena Bay region for April (left column), May, and June (right column) 2007.

6.3.4 Depth-integrated production

The total depth-integrated primary production for each of the spring months from September-November 2006 is illustrated in Figure 6.12a-6.12c and the autumn months from April-June 2007 is shown in Figures 6.12d-6.12f. Values > 80 gC.m^{-2}.month^{-1} are predicted over most of the regions in September and October whereas in November values are mostly < 80 gC.m^{-2}.month^{-1} with higher values confined to depths < 100 m and north of ca. 32°S. Total production > 95 gC.m^{-2}.month^{-1} is predicted in St Helena Bay (ca. 32.5°S) and mostly inshore of ca. 100 m bottom depth in September and October. Similar values are restricted to a small area inshore at ca. 31°45′S in November.

Over the autumn months the total depth-integrated primary production (Fig. 6.12d-6.12f) shows a fairly uniform spatial distribution. High production > 70 gC.m^{-2}.month^{-1} is shown across the bottom depth regions in April. In May production decreases to between ca. 50 and 70 gC.m^{-2}.month^{-1} for most of the area and highest values of ca. 65-70 gC.m^{-2}.month^{-1} are predicted inshore. Production decreases further in June to ca. 40-60 gC.m^{-2}.month^{-1} across most of the three regions.

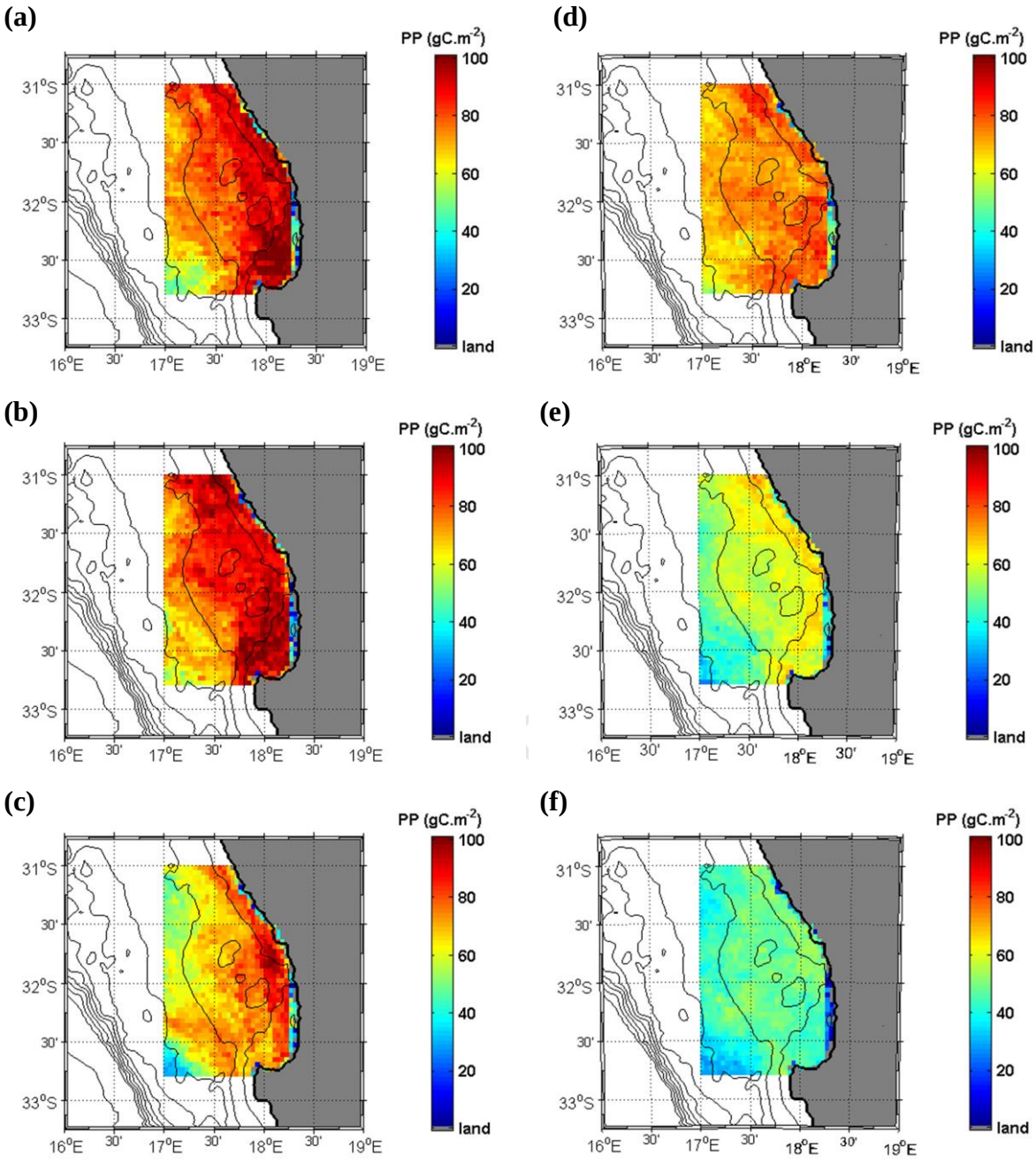


Figure 6.12 Total depth integrated primary production estimated for the spring months (a) September, (b) October, and (c) November 2006 and the autumn months (d) April, (e) May, and (f) June 2007.

6.3.5 Modelled versus *in situ* primary production

In Figure 6.13 the *in situ* primary production data from Lamont (2011) are used for comparison with the model. Whereas in Section 5.2.5 the profile is chosen by finding the best fit *k*-means profile to the measured *in situ* profile, here the model predicts the most likely profile from surface satellite data. Of the ten *in situ* measurements made in October 2006 and the ten measurements made in May 2007, a total of ten model measurements matched up by location and date. The limited matches are due to either the PAR data not being available for the day or the location falling outside the model area. The model results show both under- and over-estimation of the *in situ* data at both low < 2 $gC.m^{-2}.day^{-1}$ and intermediate values 2-5 $gC.m^{-2}.day^{-1}$, with no obvious bias.

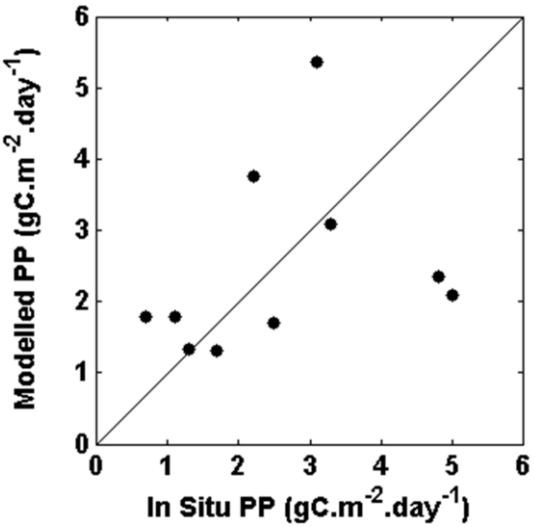


Figure 6.13 A comparison of modelled daily primary production against corresponding *in situ* measurements made in October 2006 and May 2007. The diagonal line indicates the 1:1 relationship.

6.3.6 Annual primary production

Primary production in the water column over a period of 16 months is shown in Figure 6.14. Near the surface, in the upper 5 m depth, production is relatively consistent throughout the period but with a notable decline during the months April-June. High production events appear to last from one-three days during April-August, to three-seven days for the other months and are separated by short periods of low production. The length of high production events in spring and summer is difficult to quantify due to the number of days of no data shown from October-March 2006 and from October-December in 2007. Deeper in the water column between ca. 10 and 30 m depth the estimated production shows a clear seasonal sinusoidal cycle. For example, maximum depth penetration of the 0.05 $mg.m^{-3}$ isopleth occurs in December-January and there is a gradual shoaling until a minimum in June-July. The amplitude of the seasonal fluctuation increases from inshore to the mid-shelf region.

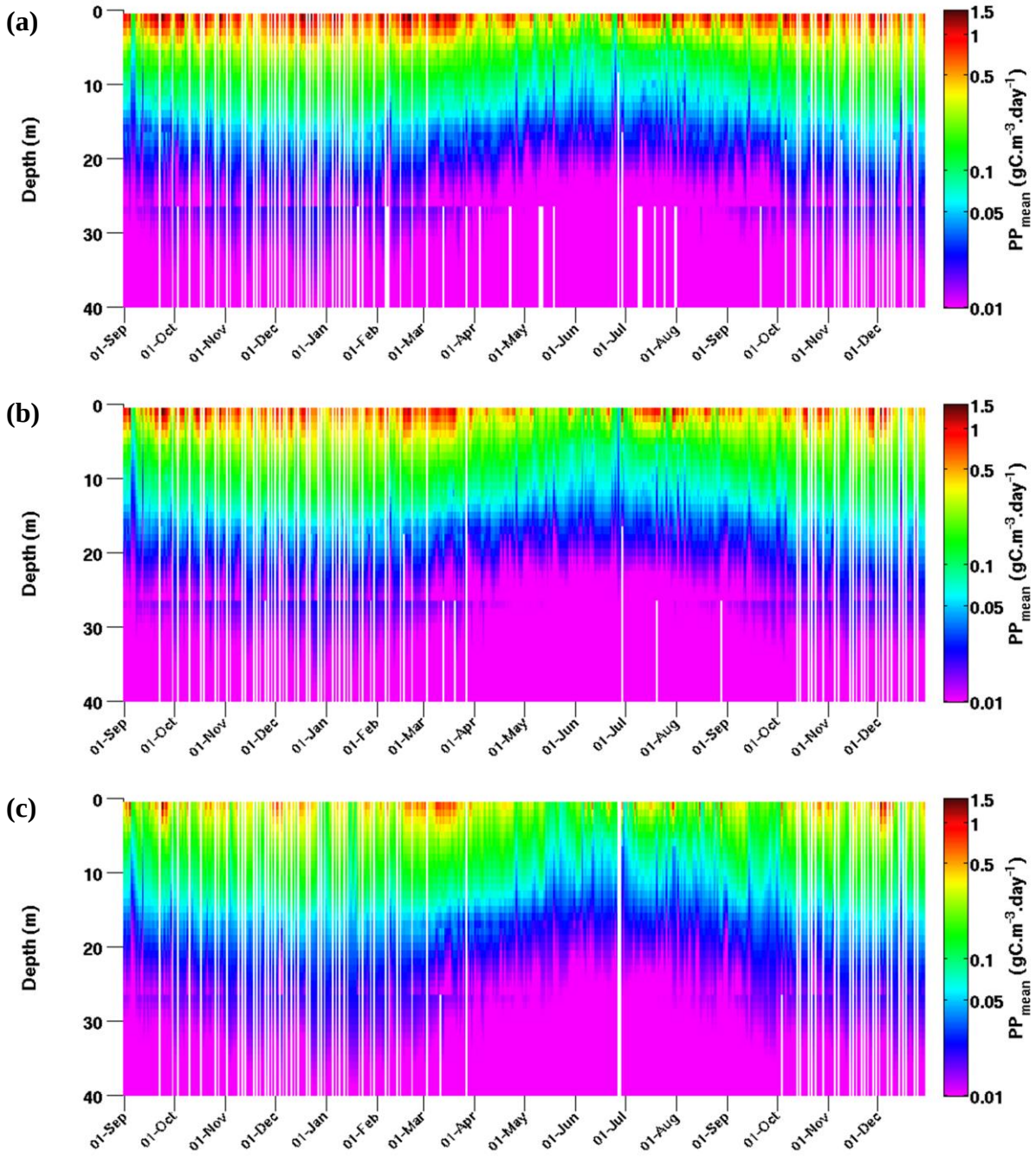


Figure 6.14 Modelled time-series of 16 months of depth-resolved primary production for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions from September 2006 to December 2007. Note the colour bar is on a log scale.

6.4 **Discussion**

6.4.1 Spring processes

The spring months are dominated by a few temperature profiles (Fig. 6.4) and a range of Chl *a* profiles (Fig. 6.5). Temperature profiles show a general warming trend but not much structural change. Weakly stratified water dominates over stronger stratification. This may be the result of gradual spring time warming and few persistent periods of calm wind and clear skies that will maximize surface warming. Temperature Profile 2, which is presumed to indicate newly upwelled or an early stage of maturing upwelled water due to its cold surface temperature, is well represented in all months, suggesting an efficient supply of nutrients to support high biomass development (Cullen *et al.,* 2002) as indicated by Chl *a* Profiles 6-12 (Fig. 6.3). Further, with weak stratification, windy events will easily and rapidly mix the surface layer with subsurface water and replenish nutrients (Brown and Hutchings, 1987; Cullen *et al.,* 2002; Kiørboe, 1993). From September to November there is shift in high biomass profiles from a dominant Profile 11 with a near surface biomass in early spring to a dominant Profile 7 with a deeper peak in late spring. There are also decreases in the frequencies of Profiles 5 and 8, which have near-surface peaks. The dominant vertical distribution and location of the peak can be explained by the stronger winds that occur from September. The stronger winds are caused by the southward shift in the South Atlantic Anti-cyclone and the increased pressure gradient across the southern African coast (Jury, 1980; Nelson and Hutchings, 1983; Kamstra, 1988). Although less frequent, temperature profiles representing stronger stratification and shallow surface mixing (Profiles 1, 3 and 4) do occur during these months (Fig. 6.4). Shallow wind mixing and stratification, indicated by temperature

Profiles 1, 3 and 4 will retain a near-surface biomass but the nutrients will be depleted rapidly, thus limiting phytoplankton growth (Brown and Hutchings, 1987, Pitcher *et al.,* 1996; Cullen *et al.,* 2002). The result will be small near-surface blooms such as Chl *a* Profiles 3 and 5, which are frequently predicted in spring (Fig. 6.5). If moderately calm conditions persist the phytoplankton may move deeper in the water column, either actively or through sinking, to take advantage of the nutricline (Pitcher *et al.,* 1996; Cullen *et al.,* 2002; Kiørboe, 1993), as represented by Chl *a* Profiles 4 and 6. This succession of profiles from near-surface to deeper in the water column is a characteristic of phytoplankton dynamics in the southern Benguela (Pitcher *et al.,* 1989; Mitchell-Innes and Walker 1991; Pitcher *et al.,* 1991; Barlow, 1992; Pitcher *et al.,* 1996; Pitcher and Nelson, 2006). Therefore, frequency pairing of these temporally related Chl *a* Profiles 3 and 4, and Profiles 5 and 6 can be expected. For the majority of months and regions (Figs. 6.5) they are relatively similar in frequency.

6.4.2 Autumn processes

The temperature profiles in autumn (Fig. 6.6) indicate that upwelling (temperature Profile 2; Fig. 6.2) is less frequent than in spring (Fig. 6.4) but is still comparable to late spring. Many studies have shown that upwelling along the West Coast is most frequent from September to April, driven by southerly wind (Andrews and Hutchings, 1980; Hutchings *et al.,* 1984; Barlow *et al.,* 2005). High frequencies of temperature Profiles 6 and 8, which indicate strong stratification in April and May, have declined by June indicating strengthening westerly winds as the low pressure belt to the south of the continent moves northward and low pressure systems begin to interact with the coastline (Tyson, 1986; Taunton-Clark and Kamstra, 1988). This is in agreement

with Horstman (1981) who found that water column stability increased in early autumn in response to a weakening pressure gradient over the region following summer.

Autumn has a more limited set of Chl *a* profiles that occur frequently (Fig. 6.7). The distribution shows that subsurface peaks are most frequent during this season and dominate in June. Mitchell-Innes *et al.* (2001) found variable profiles in June-July 1999, which indicated that typical upwelling and maturing water still occurs in late autumn and early winter. Large biomass profiles are observed near the beginning of autumn but become very infrequent by late autumn. By late autumn low biomass profiles, suggesting deep mixing, have increased whereas near surface peaks have decreased notably. This is probably the result of a combination of the increasing water column stability and decreasing available light (Kiørboe, 1993; Cullen *et al*., 2002). The light is favourable for high phytoplankton production during early autumn and the stability of the water column allows considerable growth (Horstman, 1981). By late autumn the reducing light levels and deep mixing probably inhibit surface bloom development but regular nutrient supply through mixing may sustain production (Brown and Hutchings 1985; Brown and Hutchings, 1987). This is evident from the Chl *a* profile distribution in June, which shows a less variable profile distribution (Fig. 6.7). However, in late autumn, while all other profiles indicating surface blooms decline, Chl *a* Profile 11 persists. This is surprising because there are few occurrences of other profiles (such as Profiles 6 or 10) that are expected to represent earlier or later stages of this profile. This is discussed below.

6.4.3 Modelled and observed profile distributions

Both sets of profiles in the spring months (Figs. 6.4 and 6.5) and autumn months (Fig. 6.6 and 6.7) show month-to-month and cross-shore variability, although the inshore and inner-shelf patterns are quite similar. The changes in profiles seem to follow a progressive change that would be expected as the seasons develop and fit the general hypothesis of primary production in the southern Benguela (Hutchings *et al.,* 1984; Brown and Hutchings, 1987). However, during all months and for all sub-regions one or a few from the group of temperature Profiles 5, 7, 9 and 10 and Chl *a* profiles 7, 9 and 11, tend to dominate. The expected frequencies of these profiles over the shelf are discussed with regards to the sampled profiles below.

Figure 6.15 compares the proportional distributions of observed temperature profiles to those of the predicted profiles for the three shelf sub-regions during spring and autumn. The observed profiles were obtained between 1988 and 2010 and are pooled over September-November (spring) and April-June (autumn). The total number of observed seasonal profiles for the three bottom depth sub-regions ranges from 35 over the mid-shelf in spring to 120 inshore in autumn. For the modelled data, seasonal profiles are pooled over September-November 2006 (spring) and April-June 2007 (autumn).

For most profiles the proportional match is reasonable, especially when considering the few observed profiles. The dominance of temperature Profiles 5, 7, 9 and 10 in the predicted profile distribution is also observed in the sample distribution. For instance, in spring Profile 5 is frequently observed over the inner-shelf and all four

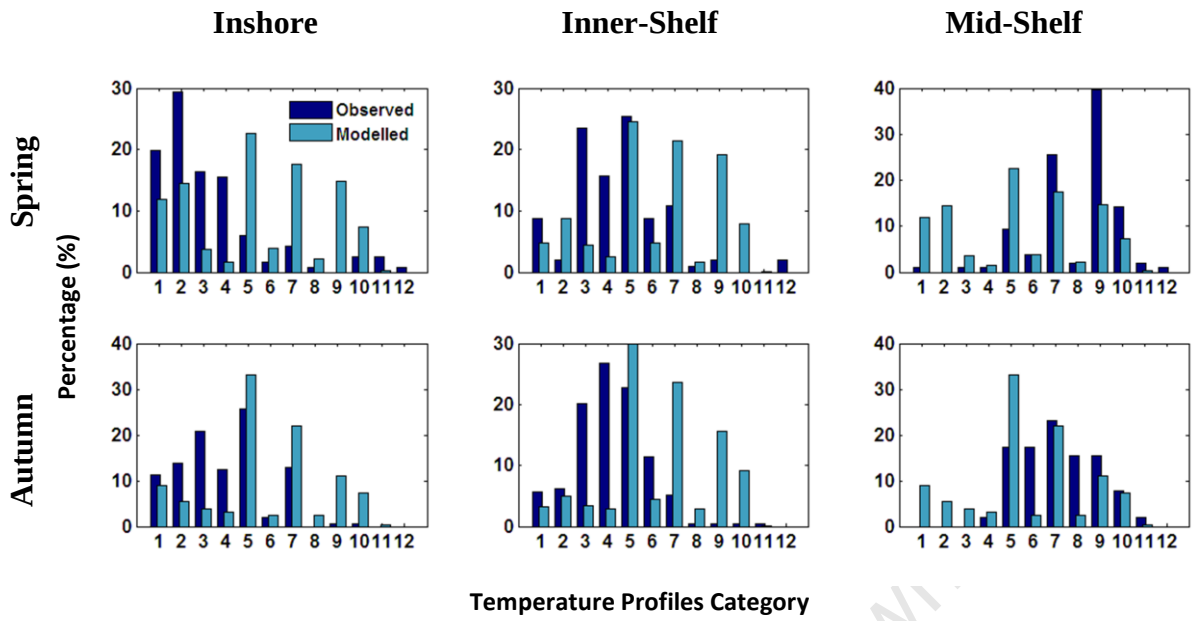


Figure 6.15 Comparison of temperature profile distributions from observed (dark blue) and modelled (light blue) data for spring and autumn. The number of observations ranged from 35 to 120 whereas the number of modelled data is of the order 10^4. The observed data are pooled from September-November (spring) and April-June (autumn) 1988 to 2010 whereas the modelled data are pooled over the months September-November 2006 and April-June 2007.

profiles dominate the mid-shelf. In autumn Profiles 5 and 7 are among the most frequent inshore and Profile 5 is among the most frequent over the inner-shelf. All four profiles are among the most frequent over the mid-shelf in autumn. Although it is difficult to comment on the accuracy of the model with such a small sample of observations and considering the single season for which the model is run, there are predictions that warrant careful interpretation. Specifically, the mismatch between modelled temperature Profiles 1 and 2 over the mid-shelf where the cold water represented by these profiles is not often observed, and the mismatch between Profiles 9 and 10 inshore and over the inner-shelf where such warm water is not often observed (Fig. 6.15). Previous studies have shown that although upwelling is

usually confined close inshore it can extend hundreds of kilometres offshore (Hagen *et al.,* 1981; Lutjeharms and Meeuwis, 1987; Brown, 1991; Boyd and Agenbag, 1996). In the St Helena Bay region this occurs regularly across the shelf, most often in February-April (Weeks *et al.,* 2006). This adds to the plausibility of observing cold water over the mid-shelf as predicted by the model. Similarly, strong onshore flow of warm water associated with relaxation of upwelling (indicated by temperature Profiles 9 and 10 inshore) has been noted by Lutjeharms and Stockton, (1987), Gorzoli and Gordon (1996) and Weeks *et al*. (2006). However, the observations suggest that these extensions and intrusions of water masses should result in smaller proportions of cold water profiles over the mid-shelf and warm water profiles inshore than the model predicts. A possible reason for the difference between modelled and observed profiles is the fine temporal and spatial resolution of the model that increases the probability of these events being predicted by the model. In contrast to the higher proportions of the model, the predicted proportions of Profiles 3 and 4 are much lower inshore and over the inner-shelf compared to the observations. These profiles have a warm surface layer over cold (ca. 10°C) water and will typically follow upwelling if there are periods of calm winds, when solar heating of the surface has a maximum effect (Brown and Hutchings, 1987). They may also occur if warmer surface water is advected inshore. Stratification is a regular occurrence in the St Helena Bay region (Buys, 1957) so it is unclear why these profiles are under-predicted by the model. It may be as a result of how the data are pre-processed before training and running the model or to an unusual study year. This is discussed below.

Figure 6.16 compares the proportional distributions of the observed and predicted Chl *a* profiles in the same way as the temperature profiles in Figure 6.15. The

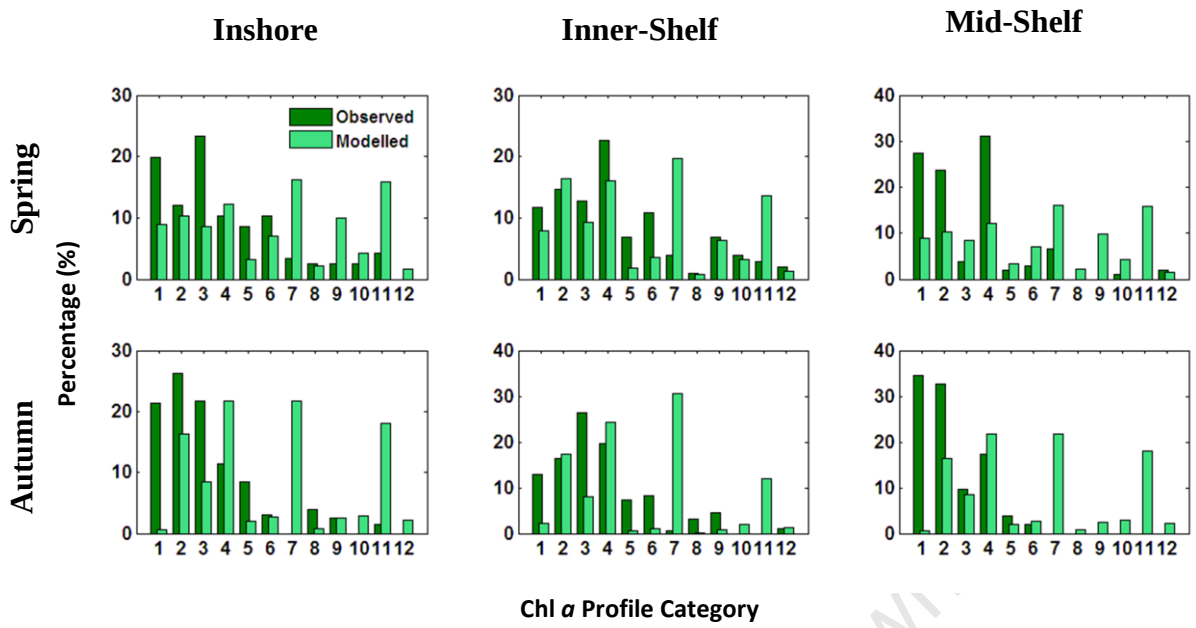


Figure 6.16 Comparison of Chl *a* profile distributions from observed (dark green) and modelled (light green) data for spring and autumn. The number of observations ranged from 35 to 120 whereas the number of modelled data is of the order 10^4. The observed data are pooled from September-November (spring) and April-June (autumn) 1988 to 2010 whereas the modelled data are pooled over the months September-November 2006 and April-June 2007.

predicted Chl *a* profile distribution shows that Profiles 2, 4, 7, and 11 tend to dominate. Profiles 2 and 4 match the observed proportions well but Profiles 7 and 11 show a mismatch with the observation data that is most evident in autumn when very few observations of these two profiles are made. Barlow *et al*. (2005) and Weeks *et al*. (2006) showed that high biomass persists across the shelf in the St Helena Bay region throughout spring, summer and autumn. Shannon *et al*. (2004) reported that Chl *a* concentrations reached a maximum in autumn. Therefore, the mismatch between the high biomass profiles is not clear. As discussed above, certain other profiles are expected to occur with similar proportion to Chl *a* Profiles 7 and 11, which would indicate a succession from phytoplankton growth to decay, although the process can be interrupted at any point in time. The absence of these

profiles suggests a potential problem with predictions of certain profiles. Chl *a* Profile 1, which indicates a homogenously low biomass water column (Fig. 6.3) matches the observations well in spring but appears to be under predicted in autumn. The observations are surprising because although Weeks *et al*. (2006) noted no clear seasonal signal in the surface Chl *a*, they did observe a seasonal signal in the upwelling index of the Cape Columbine upwelling cell that peaked in spring-summer. The model then seems to agree with Weeks *et al*. (2006) although not with the observations. This suggests that data sequences associated with Chl *a* Profile 1 do not occur as often as in spring because the CBS model does not take account of season when predicting the profile.

The over prediction of the worst matched profiles discussed above may originate from the pre-processing of data. Profiles, such as temperature Profiles 7, 9 or 10, or Chl *a* Profiles 7 or 11 probably have the broadest range of associated sub-sequences of wind, SST or Chl *a*. As a result, given a set of sub-sequences that does not clearly identify a single profile, these profiles will probably score highest among the potential profiles simply because they have the broadest range of associated sub-sequences (this will be dependent on the individual confidence and support scores of the sub-sequences). The broad range of associated sub-sequences will occur if the variable's discrete value limits are ineffective in identifying important processes necessary for distinguishing among profiles. This may also occur if the profiles in a particular cluster collectively contain a broad range of sub-sequences, in which case the clustering algorithm has failed to identify sufficiently distinguishable classes.

6.4.4 Modelled primary production

The comparison of daily primary production estimates with those made *in situ* are reasonable when considering the assumptions applied in constructing the light model (Fig. 6.13). For example, the constant photosynthetic parameters or that the satellite PAR represents the daily integral. The slight over- and underestimation but general agreement among the lower and higher values do not indicate any obvious systematic error but probably result from these assumptions and slightly different profiles being predicted than were observed *in situ*. However, the sample size of observations is very small and the comparison should be interpreted with caution.

The spatial distribution of depth-integrated primary production (Fig. 6.12) shows interesting results that warrant careful consideration. In Figures 6.17 and 6.18 the monthly integrated primary production is compared to the monthly mean MODIS surface Chl *a* concentration for the spring and autumn months respectively. The comparison is not and indication of the model skill as the model is dependent on the Chl a data. During the spring months of September and October there is high production across the inshore and inner-shelf with highest production inshore particularly in St Helena Bay (Fig. 6.17a,b). The high inshore production is a result of the retention of water in the bay that allows considerable phytoplankton growth (Holden, 1985; Weeks *et al.,* 2006).

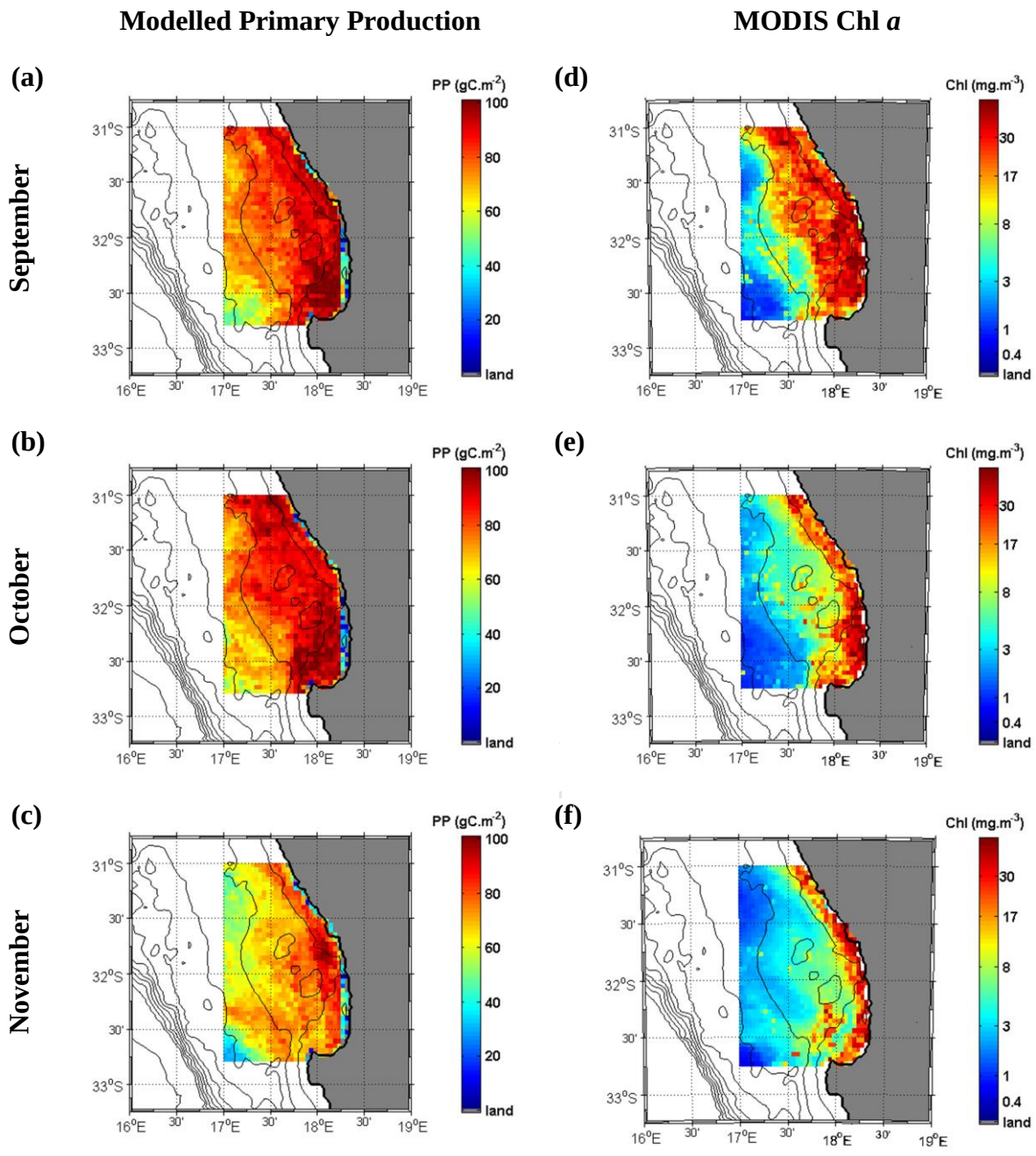


Figure 6.17 Spring depth integrated primary production for the months (a) September, (b) October and (c) November 2006 compared to the mean MODIS surface Chl *a* for the months (d) September, (e) October and (f) November 2006.

In November the modelled production drops significantly over the inner- and mid-shelf (Fig. 6.17c). The mean surface Chl *a* concentration (Fig. 6.17a-6.17c) indicates that highest surface Chl *a* biomass throughout spring is mostly contained inshore, but extends over the inner-shelf in September. The modelled production is in close agreement with the MODIS data in September and November, but in October high modelled production continues to extend over the inner-shelf whereas highest Chl *a* concentration is clearly shown inshore. One possible explanation for this is that missing Chl *a* data in the MODIS daily images have resulted in an underestimation of the true mean concentration. This under estimation is reasonable considering that frontal systems that interrupt the upwelling favourable southerly winds, often bring cloudy conditions (Tyson, 1986) and that growth is maximal during calmer wind conditions (Hutchings *et al.,* 1984, Brown and Hutchings, 1987). Another possible explanation is that the biomass is predominantly deeper in the water column, below the penetration depth of the MODIS sensor. This might explain the poor correlation that Barlow *et al*. (2009) found between the surface Chl *a* and primary production in the region. Deep Chl *a* maxima can occur during periods when upwelling is weak or calm periods persist and thus nutrient replenishment to the surface layer is insufficient to maintain high near-surface production (Pitcher *et al.,* 2006). If this is the case then the Chl *a* profile distribution will reflect this. Figure 6.5 does indeed show an increase in the frequency of profiles with a subsurface maximum (Profiles 4, 7 and 10), particularly Profile 7, from September to November, and a decrease in profiles with a near-surface peak (Profiles 5, 8, 9 and 11). However, high near-surface biomass peaks are still predicted frequently over the inner-shelf in October (Fig, 6.5b). A third possibility is that some of the predicted profiles may be incorrect. This is discussed further below.

From April to June, there is relatively consistent production over the inshore and inner-shelf, which diminishes over autumn (Fig. 6.18a-6.18c). April and May show slightly higher production inshore. Highest Chl *a* concentration is constrained to a narrow inshore band running along the coast (Fig. 6.18d-6.18f). The Chl *a* profile distributions support the MODIS data in that the inshore region shows the highest frequencies of profiles with near-surface biomass (Profile 9 and 11) whereas the inner- and mid-shelf profile distributions are dominated by profiles with deeper biomass maxima (Profiles 2, 4 and 7). The greater dominance of Profile 7, with high integrated biomass and a deep biomass peak (203 mg.Chl *a*.m^{-2}) over smaller biomass profiles (< 89 mg.Chl *a*.m^{-2}), over the inner- and mid-shelf compared to inshore, accounts for the more uniform distribution of production over the shelf.

The significantly higher Chl *a* concentration recorded inshore by MODIS in October compared to the relatively uniform distribution of production indicates a possible problem with the profile predictions. Barlow *et al*. (2009) noted a poor coefficient of variation of 43% for MODIS Chl *a* and primary production for the autumn months of 2002, which can, in part at least explain the difference. If it is assumed that there is a reasonable correspondence between *in situ* surface Chl *a* and remotely sensed MODIS Chl *a* then the inshore region should be dominated by Chl *a* profiles with high near-surface biomass particularly during the upwelling months September-April (Brown and Hutchings, 1987; Barlow *et al.,* 2005; Weeks *et al.,* 2006). However, the model shows high variability among the Chl *a* profiles with a near-surface peak during the spring and autumn months. In Section 4.4.5 the CRF module results are discussed and Figure 4.2 shows that the CRF model does not perform well when predicting high surface Chl *a* states and frequently underestimates the higher states. Although the St Helena Bay area is known for its high surface Chl *a*, it is still the best

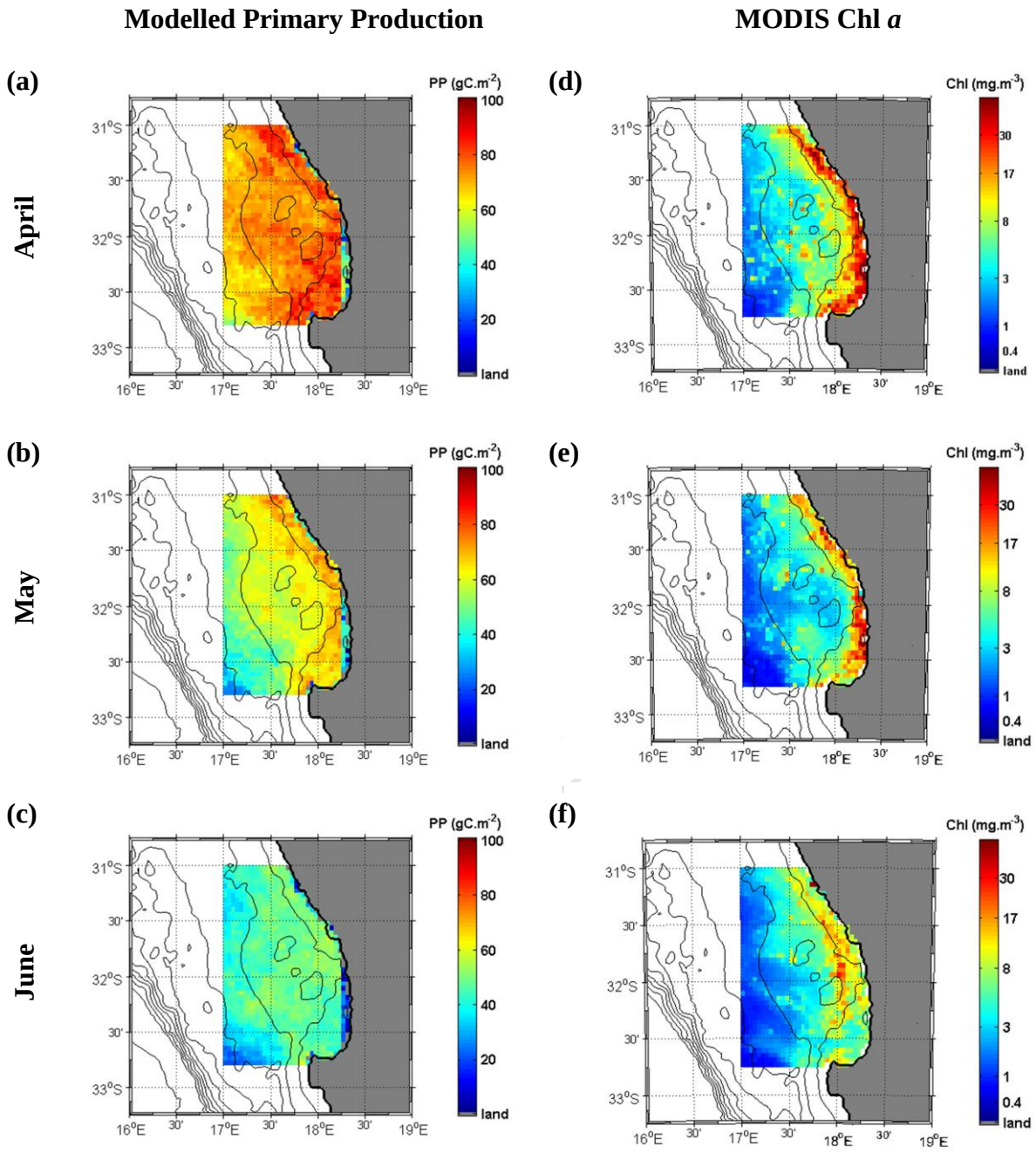


Figure 6.18 Autumn depth integrated primary production for the months (a) April, (b) May and (c) June 2007 compared to the mean MODIS surface Chl *a* for the months (d) April, (e) May and (f) June 2007.

sampled region for *in situ* data that can be used for comparison with the models developed in this study.

Figure 6.19 shows a comparison between the MODIS derived average surface Chl *a* (blue line) and the average model output surface Chl *a* (red line) for the three months September-November 2006. The MODIS data are first mapped to the labels or states used in the model and thus the average is of the ordinal state values. Where there is partial cloud cover over the region (the percentage missing data is indicated by the bar graph) the average MODIS data and average model data should be similar, particularly the relative day-to-day change in state (this assumes that the missing data are randomly distributed over the region and that the uniform distribution of wind over the small area should have a similar affect on the upper layer in terms of mixing and stratification). From Figure 6.19 it is immediately clear that the CRF model is underestimating surface Chl *a* inshore (Fig. 6.19a) and to a lesser degree over the inner-shelf (Fig. 6.19b). The correlation over the mid-shelf is good (Fig. 6.19c). For the most part the day-to-day modelled changes follow the MODIS data well. In autumn (Fig. 6.20) the relationship between the two datasets is different. Inshore (Fig. 6.20a) the model underestimates the higher concentration events and over all regions the model struggles to replicate the more intense short-term variability. The underestimation inshore explains why there is a mismatch between the inshore surface Chl a and the primary production; the missing data are often being underestimated and hence the prediction of higher near-surface biomass profiles is being affected. This is because during training of the CBS module the sequences associated with each profile include both the original MODIS and modelled states.

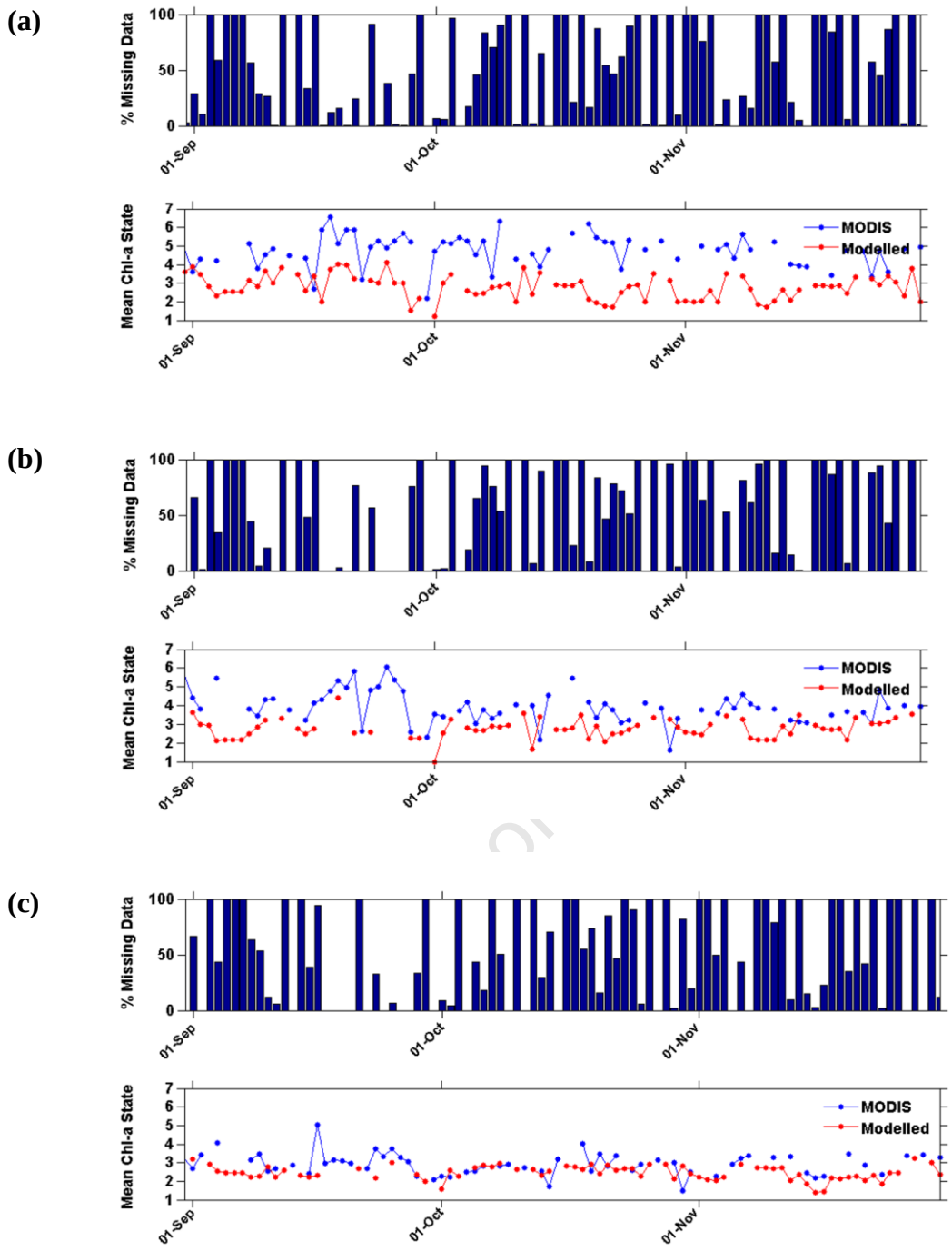
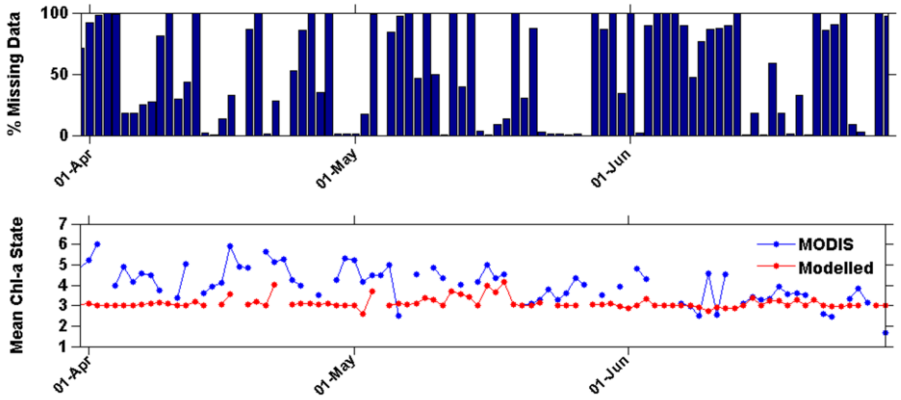
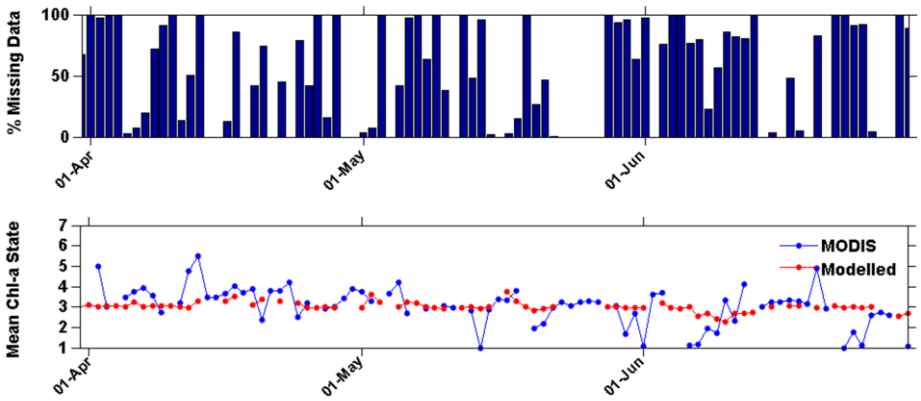


Figure 6.19 Austral spring: percentages of missing data (bar chart) and comparison of the MODIS (blue) and MODIS plus modelled (red) surface Chl a data for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions during September-November 2006.

(a)



(b)



(c)

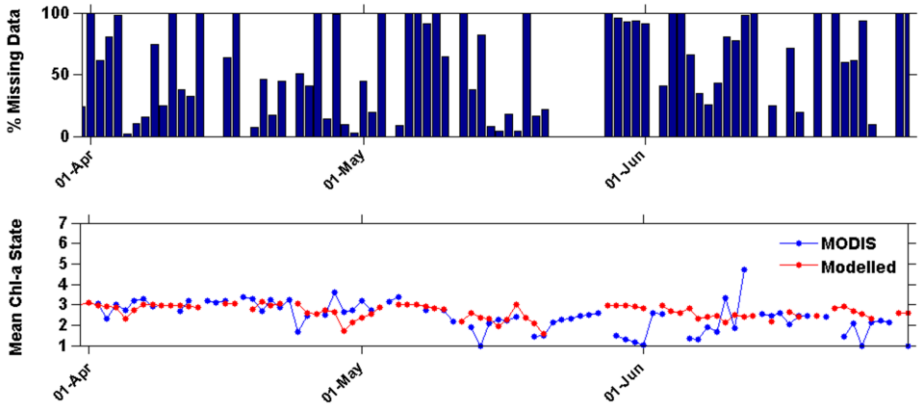


Figure 6.20 Austral autumn: percentages of missing data (bar chart) and comparison of the MODIS (blue) and MODIS plus modelled (red) surface Chl a data for the (a) inshore, (b) inner-shelf and (c) mid-shelf regions during April-June 2007.

There may then be circumstances where certain profiles, such as those with high biomass near-surface peaks, are associated with the higher MODIS Chl *a* surface sequences but also the lower predicted sequences and mixtures of both. The result will be wider confidence limits for the sequences associated with the profiles as discussed in Section 6.4.3. This problem may be resolved by adjusting the interval limits or including PAR data in the CRF module. Consecutive days of high PAR values may be sufficient to discern transitions to high Chl *a* state from transitions to moderate ones, and may also improve the accuracy of predicted SST sequences. Once the CRF module issue of underestimated states is resolved, it is likely that the model will predict more profiles inshore with higher surface biomass (such as Profiles 5, 8, 9 and 11) and fewer with lower biomass and a subsurface peak (such as profiles 4 and 7). This will be seen as an improvement in the predictive accuracy of both the CRF and CBS modules and will result in better correspondence between the MODIS surface data and integrated production inshore.

6.4.5 Long-term primary production

Figure 6.21 shows the mean daily primary production for each month in the inshore, inner-shelf and mid-shelf sub-regions. The dashed lines indicate other reported *in situ* estimates according to the time of year of the experiments. Over the period September 2006-December 2007 the model shows peak production in December 2006 for all three regions and secondary peaks in March 2007 for the mid-shelf. Weeks and Shillington (1994) found maximum Chl *a* levels in summer in the Coastal Zone Colour Scanner (CZCS) data whereas Barlow *et al*. (2005) and Weeks *et al*. (2006) found no clear seasonal signal in the SeaWiFS Chl *a* data and reported high biomass in spring summer and autumn. Shannon *et al*. (1984) reported Chl *a*

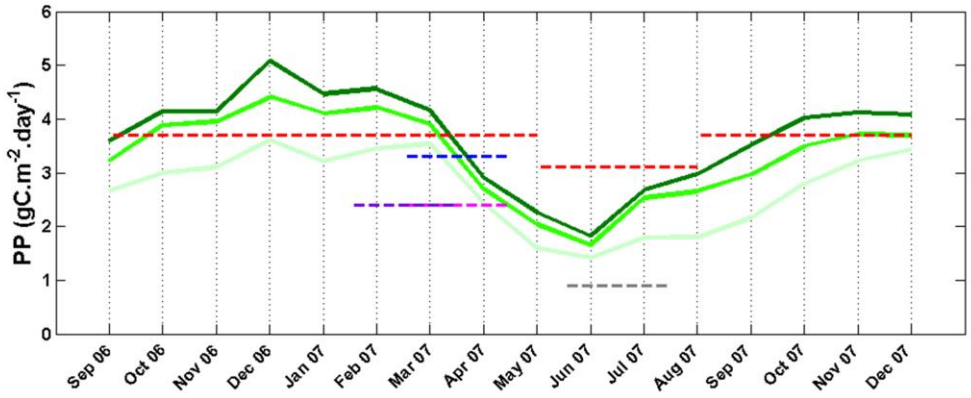


Figure 6.21 Monthly carbon production for the inshore (dark green), inner shelf (green) and mid-shelf (light green) for the study period September 2006 – December 2007. Dashed lines indicate other *in situ* estimates; (red) Shannon and Field (1985), (blue) Mitchell-Innes & Walker (1991), (pink) Walker and Peterson (1991) for St Helena Bay estimates and (purple and gray) Barlow *et al*. (2005) for southern Benguela summer and winter respectively.

concentration reached a maximum in autumn while Barlow *et al*. (2009) showed that production can be twice as high in summer than in winter. These results suggest that the production shown in Figure 6.21 may not be a climatological representation but simply a representation of the specific period. Indeed, there is a clear difference between the December 2006 and December 2007 production. June shows lowest production for all three regions in agreement with expectations as well as confirming the findings of Barlow *et al*. (2005) and Weeks *et al*. (2006).

The proportion of the total production differs among the three regions from month to month, although the inshore region consistently has the highest and the mid-shelf the lowest production. For example, in October and November 2006 the inshore and inner-shelf sub-regions are fairly similar but production increases more inshore than over the inner-shelf in December 2006. In October and November 2007 the

difference between the inshore and inner-shelf is larger than in 2006. From April to June production in all sub-regions is similar. From July the production rapidly increase inshore and over the inner-shelf while only increasing over the mid-shelf in September. These differences can be attributed to the pulsed nature of the inshore sub-region that sees both high and low Chl *a* in accordance with upwelling events, and the relatively more stable Chl *a* inner-shelf with high production (Weeks *et al.,* 2006). Further evidence of this is that some sub-regions of the shelf may increase while others decrease (for example, February-March 2006). This has important consequences on overall production.

Figure 6.22 shows the total monthly production over the shelf. In March 2007 the increase in production over the mid-shelf results in elevated production over the entire region (see Fig. 6.22) and a peak in production only second to that in December 2006. Unlike the Chl *a* studies done in the region (Shannon, 1984; Weeks and Shillington, 1994; Barlow *et al*., 2005; Weeks *et al.,* 2006; Barlow *et al.,* 2009) that do not show a clear seasonal signal, the production estimated by the

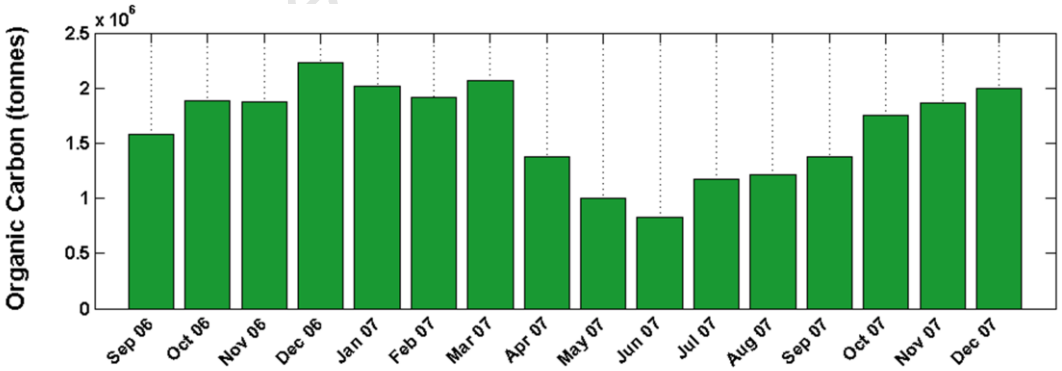


Figure 6.22 Estimations of the total monthly carbon production for the model region (18,000 km^2) for the study period September 2006-December 2007.

model does (Fig. 6.22). Because these studies were based on surface Chl *a* data, it is logical to investigate the production near the surface.

Figure 6.23 compares total production in the upper 4 m of the water column for the shelf to total production from 10-30 m depth. The figure shows that production near the surface is high and variable from September to March with no clear seasonal change in accordance with previous studies. Relatively low production only occurs in May and June. Deeper in the water column the production shows a clear seasonal signal as a sinusoidal pattern. The seasonal signal of production deeper in the water column mirrors that of the seasonal PAR and suggests that light is limiting here. Near the surface nutrients are probably the main limiting factor. These limiting factors interact to produce similar production in the upper and deeper layers in the summer and winter months whereas during spring and autumn the upper layer can produce nearly twice the amount of organic carbon as the deeper layer (Fig. 6.23).

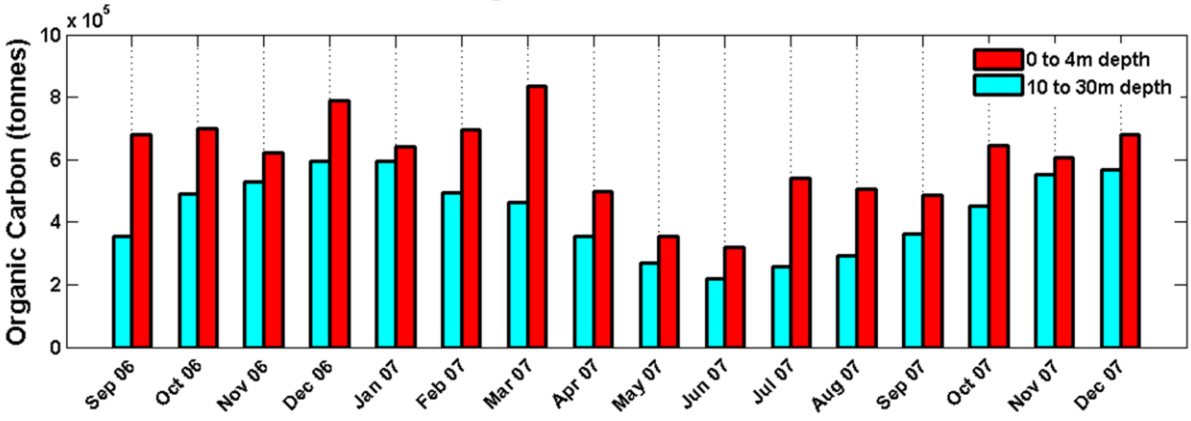


Figure 6.23 A comparison of total monthly production in the upper 4 m of the water column (red bars) and from depth 10 to 30 m (blue bars) for the study period September 2006-December 2007.

The intra-annual differences in production between the surface layer and deeper layers highlight the potential error in using only satellite-derived surface Chl *a* values while ignoring the subsurface biomass distribution and light field in estimates of primary production. This will be even more pronounced in regions with documented subsurface (deep) Chl *a* peaks, such as on the Agulhas Bank.

6.4.6 Annual primary production

Previous *in situ* estimates of primary production in the Benguela are summarized in Table 6.1. Shannon and Field (2005) reported a mean daily production of 2.8 $gC.m^{-2}.day^{-1}$ for the southern Benguela with a mean of 3.1 $gC.m^{-2}.day^{-1}$ in the winter months and 3.7 $gC.m^{-2}.day^{-1}$ for the rest of the year. Mitchell-Innes and Walker (1991) reported a mean *in situ* production of 3.3 $gC.m^{-2}.day^{-1}$ and upper limit of 7.85 $gC.m^{-2}.day^{-1}$ in the St Helena Bay region in March-April 1987. Walker and Peterson (1991) reported an average of 2.4 $gC.m^{-2}.day^{-1}$ in a large cell dominated bloom and 1.5 $gC.m^{-2}.day^{-1}$ in small cell bloom in April the following year. Barlow *et al*. (2009) investigated production throughout the whole Benguela ecosystem in winter 1999 and summer 2002. They obtained a range from 0.14-2.26 $gC.m^{-2}.day^{-1}$ in June-July 1999 and from 0.39-8.83 $gC.m^{-2}.day^{-1}$ in February-March 2002. The results show a generally good agreement amongst the variables especially considering that various photosynthetic parameters were obtained during the *in situ* experiments.

Table 6.1 Summary of productionresults obtained in the Benguela system

Authors	Area (km^2)	Total Production (Giga tonnes)	Primary Production ($gC\ m^{-2}\ day^{-1}$)
Shannon & Field (1985)	**40,000**	**0.04**	**3.1 (win) 3.7 (other) 2.8 (mean)**
Michell-Innes & Walker (1991)	**-**	**-**	**3.3 (mean) 7.85 (max)**
Walker & Peterson (1991)	**-**	**-**	**1.5 (small)- 2.4 (large)**
Barlow *et al.* (2009)	**-**	**-**	**0.9 (win) 2.4 (sum)**
Current Study	**18,000**	**0.019**	**1.4 (win) 5.08 (sum) 3.2 (mean)**

sum = summer, win = winter, other = all other seasons, small = small cells, large = large cells

6.5 **Conclusion**

The southern Benguela region is known to be a dynamic highly variable upwelling system that rapidly responds to changes in local wind (Brown and Hutchings, 1987). Most, if not all, estimates of primary production in the region have been based on averaging over those processes that occur at sub daily to daily scales. In this chapter rapid changes in the system are explicitly modelled as day-to-day variability both in the horizontal and vertical distribution of phytoplankton.

The predicted vertical distribution of phytoplankton is the most probable distribution given information on the current dynamics of the physical and biological environment and is thus directly linked to the forcing mechanisms. The model clearly shows the sensitivity of the vertical distribution of phytoplankton to environmental processes with a broad range of predicted profiles that differ regionally, seasonally and from

month to month. In addition to the variable distribution and biomass of the phytoplankton in the water column, the model also accounts for variability of hourly surface irradiance and the Chl *a*-dependent subsurface light field. The accuracy of these predictions can be improved further with reasonably straight-forward adjustments to the pre-processing of satellite and profile data. For example, the coldest upwelling water, which is expected to be < 13°C is not well represented in the training. Random sampling from individual depth intervals (which includes a < 50 m interval) should capture more data indicating cold upwelled water. Further, the profile clusters do not differentiate optimally among the most informative features such as mixing layer depth for the temperature profiles or deep Chl *a* maximum for the Chl *a* profiles. It is worth investigating clustering the profiles based on specific features using a method such as self-organizing maps that can cluster among multidimensional data.

Nevertheless, the production estimates highlight the variability of the system in both time and space as a result of the complex and non-linear interactions of the input variables, which are captured by the model. For example, the time-series of production at depth shows that the production does not change uniformly with depth over the seasons but shows different responses near the surface and deeper in the water column. This also highlights the potential error when estimating production from surface Chl *a* data only, which will be more pronounced in regions with a persistent subsurface maximum, such as the Agulhas Bank.

The estimates of primary production and total production are derived from depth-resolved, hourly primary production calculations and compare well to others made in the Benguela. Average daily production of 3.2 $gC.m^{-2}.day^{-1}$ is estimated for the

modelled region that shows a range from 1.4 $gC.m^{-2}.day^{-1}$ over the mid-shelf in July to 5.1 $gC.m^{-2}.day^{-1}$ inshore in December. Total annual production of 0.019 Gt $C.year^{-1}$ is estimated for the modelled area of 18,000 km^2. These estimates are particularly close to the results of Shannon and Field (1985), which are based on *in situ* data and slightly higher than those of Carr (2002) who modelled primary production in the southern Benguela using surface Chl *a* and the subsurface light field, over and area of 96,000 km^2 and obtained and average of 2.5 $gC.m^{-2}.day^{-1}$.. However, unlike the earlier research on primary production, the high spatial and temporal detail that is provided by the model allows a closer investigation of processes at event scales.

Chapter 7 - Conclusion

Primary production provides a key process linking marine ecosystems to global climate change. It is of major importance in the global carbon cycle both as a transfer path of carbon to the ocean depths and to other marine organisms. Understanding the timing and scale of production events in response to current and future forcing variability is essential for many fields of ocean science. For example, many zooplankton and fish larvae are dependent on the timing and concentration of Chl *a* blooms (Lasker 1975; Cury and Roy 1989; Platt *et al.*, 2003; Durant *et al.,* 2005). Climate change may cause a mismatch of these predator-prey interactions, which has been referred to as the match-mismatch hypothesis (Cushing, 1990). In the Benguela upwelling system the duration from upwelling to bloom decay typically occurs at the scale of days and is closely linked to changes in the wind field (Hutchings and Nelson, 1985; Brown and Hutchings, 1987; Nelson, 1992; Roy *et al.,* 2001). These local events have important implications for fish recruitment, particularly in St Helena Bay (Peterson and Painting, 1990; Armstrong *et al.,* 1991; Hutchings, 1992; Richardson and Verheye, 1998). Therefore, following the biological processes is important for ecosystem studies.

The present study aimed to produce a three-dimensional estimation of primary production at a similar temporal scale to the events that control production. Two sources of information vital to studying large scale three-dimensional distributions of phytoplankton are readily available; remote-sensing data of the ocean surface, and subsurface temperature structures and biomass distributions obtained from ships. The temporal and spatial scale of satellite data are able to capture the upper-layer forcing (wind) and response processes (sea surface temperature (SST) and

chlorophyll *a* (Chl *a*)) of the system. This thesis hypothesized that daily sequences of forcing and response variables associated with each *in situ* profile, when combined with other information on season and subregion, can be found with regularity in an archive of data spanning nearly a decade. Further, that these patterns, when found in new data sequences can be used to predict the thermal structure and vertical distribution of phytoplankton in the water column for each pixel of a satellite image. The vertical distribution of phytoplankton can then be incorporated into a primary production model to estimate depth-resolved production.

To achieve this, methods were chosen to extract complex and often non-linear statistical relationships among variables, and present these relationships in an easy-to-understand format. A Bayesian framework was selected and first applied in a "static" format (time sequences of variables were ignored) to discrete data that had been pre-processed according to a detailed study of the variability of the system (Chapter 2). During pre-processing continuous variables were discretized according to intuitive temporal and spatial changes in the system with reference to primary production. The temperature and Chl *a* profiles were clustered using the *k*-means algorithm. For both sets of profiles 12 clusters were chosen to represent the variability of the system. The "static" Bayesian approach (Chapter 3) was tested on its ability to predict daily surface Chl *a* from information on other variables (wind, SST, season, region and depth) occurring on the same day. The Bayesian network, which encodes the dependent relationships among the variables, was automatically constructed and parameterized from a subset of the data. The ability of the network to accurately predict discrete states of surface Chl *a* was tested on a separate subset of the data. The network predicted an average of 61% of surface Chl *a* states correctly. The Bayesian network was then tested on its ability to predict the correct

Chl *a* profile clusters from a similar selection of same-day variables. The results showed a range of success from 88% correct for the lowest integrated biomass profile to 0% for the highest. The error shown by these results was attributed, in part, to ignoring the time-sequences of variables that capture the evolution of relevant processes.

To address this, a "dynamic" approach was explored in Chapter 4. Methods were adapted from temporal pattern mining algorithms that search for consecutive events that can be related to a class, or profile cluster. In order to obtain sufficient training data for the algorithms it was necessary to fill in the frequent missing data in the SST and Chl *a* sequences. Whereas in Chapter 3 a static Bayesian approach was used to estimate the Chl *a* state from a set of five states, in Chapter 4 a dynamic version or Conditional Random Field (CRF) was applied to predict one state from a set of seven SST or Chl *a* states. The CRF model is well suited to the MODIS data as the same-day SST data were usually not available when Chl *a* data were missing (as is required with the "static" network). The advantage of this model is that any number of relationships or features believed to be influential on the transition from one state to the next or on the actual state being evaluated, can be included. Once parameterized the weights of each feature indicate the strength of the relationship with the state or state transition. The model was tested on seasonal data and scored on average 45% and 55% correct for predicting missing SST and Chl *a* states respectively. When states adjacent to the correct state were included, the model usually scored above 90% correctly. This is an important result that highlights the problem of using discrete intervals for continuous data. The sequence mining model was trained and tested on the original data, which included missing values, and the same data after using the CRF model to fill in the missing data. The CRF data

improved the accuracy of temperature profile predictions from 33 to 39% correct but had little impact on the accuracy of predicting the Chl *a* profiles (from 17 to 18% correct). After filling in the missing values the sequence mining model tended to predict Chl *a* profiles with lower integrated biomass than the correct profiles. This was attributed to the CRF model, which often underestimated the missing Chl *a* state. The CRF model is dependent on the discretization of the continuous variables as these states are used during training. The predictions probably will be improved by increasing sampling from regions of high surface biomass and upwelling cells where SST tends to be lowest. This will lead to a better representation of the extreme characteristic waters in the sequences used in the training data.

In order to estimate primary production in the water column a model was developed that can incorporate the vertical distribution of Chl *a* and the attenuation of light (Chapter 5). The model uses broad-band light and the Chl *a* biomass distribution to calculate the attenuation of surface PAR. The daily PAR from the MODIS sensor was used to estimate hourly surface PAR from sunrise to sunset. PAR and estimated phytoplankton biomass at 1 m depth intervals was then used to calculate primary production at depth using the equation of Smith (1956) and photosynthesis parameters averaged from previous *in situ* experiments in the southern Benguela. Averaging the parameters greatly simplifies the variability of photosynthesis observed during different seasons (Mitchell-Innes *et al.,* 2001; Lamont, 2011) and among different species (Fujiki and Taguchi, 2002). However, investigations into the variability of the parameters in the Benguela are limited. Nevertheless, comparisons made between production estimated by the model and *in situ* experiments were good and a correlation coefficient of 0.74 was obtained using the equation of Smith (1936).

In Chapter 6 the models developed in Chapters 4 and 5 were combined and applied to a limited area, the St Helena Bay sub-region, because of computer time constraints. The model was run for three spring months (September-November 2006) and three autumn months (April-June 2007) for seasonal comparison with a limited number of actual primary production measurements for validation. The temperature and Chl *a* profile outputs and estimated daily production were discussed according to inshore (< 100 m bottom depth), inner shelf (100-199 m bottom depth) and mid-shelf (200-500 m bottom depth) regions. The predicted profiles show a clear across-shore transition that can be explained by the ageing of inshore upwelled water as it moves offshore, and month-to-month changes in accordance with expected transitions of wind and solar radiation. The model run was then extended over 16 consecutive months to obtain a time-series of daily production and an estimate of annual production. The daily depth-resolved primary production estimates showed pulses of high biomass near the surface in spring and autumn. Deeper in the water column, production increased over the spring months and decreased over autumn. The long-term production estimates showed that the near-surface production was in agreement with published literature; that surface Chl *a* (assumed to be a proxy for near-surface production) remains consistent throughout spring to autumn and declines over winter. However, the model also showed that production has a clear seasonal signal deeper in the water column. The depth-integrated production replicates the seasonal signal (a stronger signal with distance offshore), which highlights the important role of phytoplankton unseen by satellite remote sensors in total production estimates. Estimates of annual production (mean of 0.32 $gC.m^{-2}.day^{-1}$) compared well to other published estimates. The agreement between the model output and published literature supports the model as a fine

resolution (daily 4km^2) three-dimensional estimator of primary production in the St Helena Bay region. Because the estimates are ultimately derived from predictions based on wind patterns and sequences in the surface SST and Chl *a* these results provide good support for the hypothesis tested.

There remains scope for improvement and refinement of the model output. Improved accuracy of the profile predictions can be addressed during the data preprocessing and clustering stages. In particular, alternative methods to *k*-means clustering of the profiles can be tested that are more capable of extracting specific relevant features of the profiles. Further improvement in the prediction of Chl *a* profiles is likely to be achieved by using the predicted thermal structure in conjunction with the sequences of surface data. However, this will be strongly dependent on the accuracy of the temperature profile predictions. Refined estimates of primary production can achieved by incorporating phytoplankton size classes and their Chl *a* specific photophysiological parameters (Brewin *et al*. 2010). For example, Uitz *et al*. (2008) showed that these parameters differ among size classes, and Barlow *et al*. (2005) showed the both large celled diatoms and smaller flagellates play a significant role in shelf production. Since phytoplankton size class and their appropriate photosynthetic rates can be estimated (Morel, 1991; Bricaud *et al*., 1995; Behrenfeld and Falkowski, 1997b; Uitz *et al*., 2006), from the same data already used in the current study it would be interesting to estimate size-class specific production in the water column. Long term variabilty of the relative contribution of different size classes may provide insights into bottom-up effects on ecosystems and hence, changes in ecosystem structure.

This study shows the potential of finding complex relationships hidden among the fine spatial and temporal resolution satellite data and the subsurface temperature

and Chl *a* data. The methods investigated have been applied to the complex upwelling region of St Helena Bay, which represents one of the most productive sub-regions of the southern Benguela upwelling system. These relationships can be used to reconstruct daily depth-resolved primary production from the dynamic biophysical environment. The model can be trained on existing regional data archives to discern local ecosystem variability or future modelled climate scenarios to investigate likely future responses.

References

Agrawal, A. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules. In Bocca, J. B., Matthias, J. and Zaniolo, C. (eds) *VLDB'94, Proceedings of 20th International Conference on Very Large Databases,* Morgan Kaufmann, Santiago de Chile, pp. 487 – 499.

Aguirre-Hernández, E., Gaxiola-Castro, G., Nájera-Martínez, S., Baumgartner, T., Kahru, M. and Mitchell, B. G. (2004) Phytoplankton absorption, photosynthetic parameters, and primary production off Baja California: summer and autumn 1998. *Deep-Sea Res.*, **51**, 799 – 816.

Albert, A., Echevin, V., Lévy, M. and Aumont, O. (2010) Impact of nearshore wind stress curl on coastal circulation and primary production in the Peru upwelling system. *J. Geophys. Res.*, **115**, C12033.

Andrews, W. R. H. and Hutchings, L. (1980) Upwelling in the southern Benguela Current, South Africa. *Prog. Oceanogr*., **9**, 1 – 81.

Armstrong, D. A., Verheye, H. M. and Kemp, A. D. (1991) Short-term variability during an anchor station study in the southern Benguela upwelling system: Fecundity estimates of the dominant copepod *Calanoides carinatus. Prog. Oceanogr*., **28**, 167 – 188.

Atkins, W. R. G. (1928) Seasonal variations in the phosphate and silicate content of seawater during 1926 and 1927 in relation to the phytoplankton crop. *J. Mar. Biol. Assoc. U.K.*, **15**, 191 – 205.

Austin, R. W. and Petzold, T. L. (1981) The determination of the diffuse attenuation coefficient of sea water using the coastal zone colour scanner. In Gower, J. F. R. (ed.), *Oceanography from space*, Plenum Press, New York, pp. 239 – 256.

Bakun, A. (1993) The California Current, Benguela Current and Southwestern Atlantic shelf ecosystems: A comparative approach to identifying factors regulating biomass yields. In Sherman, K., Alexander, L. M. and Gold, B. (eds). *Large marine ecosystems stress mitigation and sustainability*. American Association for the Advancement of Science, Washington DC., pp. 199 – 221.

Bakun, A. and Nelson, C. S. (1991) The seasonal cycle of wind-stress curl in subtropical eastern boundary current regions. *J. Phys. Oceanogr.,* **21**, 1815 – 1834.

Balch, W. M. and Byrne, C. F. (1994) Factors affecting the estimate of primary production from space. *J. Geophys. Res.*, **99**, 7555 – 7570.

Barlow, R. G. (1982) Phytoplankton ecology in the southern Benguela current. Dynamics of a bloom. *J. expl, mar. Biol. Ecol.,* **63** (3), 239 – 248.

Barlow, R. G., Aiken, J., Sessions, H. E., Lavender, S. and Mantel, J. (2001) Phytoplankton pigment, absorption and ocean colour characteristics in the southern Benguela ecosystem. *S. Afr. J. Sci.,* **97**, 230 – 238.

Barlow, R. G., Cummings, D. G. and Gibb, S. W. (1997) Improved resolution of mono and divinyl chlorophylls a and b and zeaxanthin and lutein in phytoplankton extracts using reverse phase C-8 HPLC. *Mar. Ecol. Prog. Ser.,* **161**, 303 – 307.

Barlow, R. G., Lamont, T., Mitchell-Innes, B., Lucas, M. and Thomalla, S. (2009) Primary Production in the Benguela ecosystem, 1999 – 2002. *Afr. J. Marine Sci.,* **31** (1), 97 – 101.

Barlow, R. G., Sessions, H., Balarin, M., Weeks, S., Whittle, C. and Hutchings, L. (2005) Seasonal variation in phytoplankton in the southern Benguela: Pigment indices and ocean colour. *Afr. J. Marine Sci.,* **27** (1), 275 – 287.

Barnard, A. H., Zaneveld, J. R. V., Pegau, W. S., Mueller, J. L., Maske, H., Lara-Lara, R. and Álvarez-Borrego, S. (1999) The determination of PAR levels from absorption coefficient profiles at 490 nm. *Ciencias Marinas*, **25** (4), 487 – 507.

Behrenfeld, M. J. and Falkowski, P. G. (1997a) A consumer's guide to phytoplankton primary productivity models. *Limnol. Oceanogr.,* **42** (7), 1479 – 1491.

Behrenfeld, M. J. and Falkowski, P. G. (1997b) Photosynthetic rates derived from satellite based chlorophyll concentrations. *Limnol. Oceanogr.,* **42** (1), 1 – 20.

Bentamy, A. D., Croize-Fillon, D., Queffeulou, P., Liu, C. and Roquet, H. (2009) Evaluation of high - resolution surface wind products at global and regional scales. *J. Oper. Oceanogr.*, **2** (2), 15 – 27.

Bertsekas, D. P. (1999) *Nonlinear programming,* (2nd edn.). Athena Scientific, Belmont, MA.

Bishop, C. M. (2006) *Pattern recognition and machine learning*. Springer, New York.

Blanke, B., Roy, C., Penven, P., Speich, S., McWilliams, J. and Nelson, G. (2002) Linking wind and interannual upwelling variability in a regional model of the southern Benguela. *Geophys. Res. Lett.*, **29** (24), 41-1 – 41-4.

Borchers, P. and Hutchings, L. (1986) Starvation tolerance, development time and egg production of *Calanoides carinatus* in the southern Benguela Current. *J. Plankton. Res*., **8** (5), 855 – 874.

Botsford, L. W., Lawrence, C. A., Denver, E. P., Hastings, A. and Largier, J. L. (2003) Wind strength and biological productivity in upwelling systems: an idealized study. *Fish. Oceangr.*, **12**, 245 – 259.

Boyd, A. J. and Agenbag, J. J. (1985) Seasonal trends in the longshore distribution of surface temperature off Southwestern Africa 18-34°S, and their relation to subsurface conditions and currents in the area 21-24°S. In Bas, C., Margalef, R. and Rubies, P. (eds), *International Symposium on the Most Important Upwelling Areas off Western Africa (Cape Blanco and Benguela)*. Insituto de Investigaciones Pesqueras, Barcelona, pp. 119-148.

Boyd, A. J. and Shillington, F. A. (1994) Physical forcing and circulation patterns on the Agulhas Bank. *S. Afr. J. Sci.*, **90**, 114 – 122.

Boyd, A. J., Tromp, B. B. S. and Horstman, D. A. (1985) The hydrology off the South African south - western coast between Cape Point and Danger Point in 1975. *S. Afr. J. Marine Sci.,* **3**, 145 – 168.

Brandt, K. (1899) Uber den Stoffwechsels im Meere*. Wiss. Meeresunters, Abt. Keil,* **4**, 213 – 230.

Brewin, R. J., Sathyendranath, S., Hirata, T., Lavender, S. J., Barciela, R. M. and Hardman-Mountford, N. J. (2010). A three-component model of phytoplankton size class for the Atlantic Ocean. *Ecological Modelling*, **221**, 1472-1483.

Bricaud, A., Babin, M., Morel, A. and Claustre, H. (1995) Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: Analysis and parameterization. J. Geophys. Res., 100 (C7), 13,321 – 13,332.

Bricaud, A. and Stramski, D. (1990) Spectral absorption coefficients of living phytoplankton and nonalgal biogenous matter: A comparison between the Peru upwelling area and the Sargasso Sea. *Limnol. Oceanogr.,* **35**, 562 – 582.

Brown, P. C. (1986) The development and decline of phytoplankton blooms in the southern Benguela upwelling region. Ph.D thesis, University of Cape Town: xiii + pp. 151.

Brown, P. C. (1992) Spatial and seasonal variation in chlorophyll distribution in the upper 30 m of the photic zone in the southern Benguela/Agulhas ecosystem. In Payne, A. I. L., Brink, K. H., Mann, K. H. and Hilborn, R. (eds), *Benguela Trophic Functioning*. *S. Afr. J. Marine Sci.,* **12**, pp. 515 – 525.

Brown, P. C. and Hutchings, L. (1987) The development and decline of phytoplankton blooms in the southern Benguela upwelling system. Drogue movements, hydrography and bloom development. In Payne, A. I. L., Gulland, J. A. and Brink, K. H. (eds), *The Benguela and Comparable Ecosystems. S. Afr. J. Marine Sci.*, **5**, 357 – 391.

Brown, P. C., Painting, S. J. and Cochrane, K. L. (1991) Estimates of phytoplankton and bacterial biomass and production in the northern and southern Benguela ecosystems*. S. Afr. J. Marine Sci.,* **11**, 537 – 564.

Buys, M. E. L. (1957) Temperature variations in the upper 50 metres in the St Helena Bay area, September 1950-August, 1954. *Investigational Report of the Division of Sea Fisheries*, **27**, pp. 1 – 114.

Capet, X. J., Marchesiello, P. and McWilliams, J.C. (2004) Upwelling response to coastal wind profiles. *Geophys. Res. Lett.*, **31**, L13311.

Carr, M-E. (2002) Estimation of potential productivity in Eastern Boundary Currents using remote sensing. *Deep-Sea Res. II,* **49**, 59 – 80.

Carter, R. A., McMurray, H. F. and Largier, J. L. (1987) Thermocline characteristics and phytoplankton dynamics in Agulhas Bank waters. In Payne, A. I. L., Gulland, J. A., and Brink, K. H. (eds), *The Benguela and Comparable Ecosystems*. *S. Afr. J. Marine Sci.,* **5**, pp. 327 – 336.

Chapman, P. and Shannon, L. V. (1985) The Benguela ecosystem. 2. Chemistry and related processes. In Barnes, M. (ed.), *Oceanography and Marine Biology. An Annual Review,* **23**, University Press, Aberdeen, pp. 183 – 251.

Claustre, H., Babin, M., Merien, D., Ras, J. and Prieur, L. (2005) Toward a taxon - specific parameterisation of biooptical models of primary production: A case study in the North Atlantic. *J. Geophys. Res*., **110**, C07S12.

Côté, B. and Platt, T. (1984) Utility of the light-saturation curve as an operational model for quantifying the effects of environmental conditions on phytoplankton photosynthesis. *Mar. Ecol. Prog. Ser.,* **18**, 57 – 66.

Crawford, R. J. M., Shelton, P. A., and Hutchings, L. (1980) Implications of availability, distribution and movements of pilchard (*Sardinops ocellata*) and anchovy (*Engraulis capensis*) for assessment and management of the South African purse-seine fishery. *Cons. int. Explor. Mer*, **177**, 355 – 373.

Cullen, J. J., Yang, X. and MacIntyre, H. L. (1992) Nutrient limitation in marine photosynthesis. In Falkowski, P. G. and Woodhead, A. D. (eds), *Primary Productivity and Biogeochemical Cycles in the Sea*. Plenum Press, New York, pp. 69 – 88.

Cushing, D. H. (1971) Upwelling and the production of fish. *Adv. Mar. Biol.,* **9**, 255 – 334.

Demarcq, H., Richardson, A. J. and Field, J. G. (2008) Generalised model of primary production in the southern Benguela upwelling system. *Mar. Ecol. Prog. Ser.,* **354,** 59 – 74.

De Villiers, S. D. (1998) Seasonal and interannual variation in phytoplankton biomass on the southern African continental shelf: Evidence from satellite - derived pigment concentrations. In Pillar, S. C., Moloney, C. L., Payne, A. I. L. and Shillington, F. (eds), *Benguela Dynamics. S. Afr. J. Marine Sci.*, **19**, pp. 169 – 179.

Dogliotti, A. I., Schloss, G. O. and Gagliardini, D. A. (2009) Evaluation of SeaWiFs and MODIS chlorophyll-*a* products in the Argentinean Patagonian Continental Shelf (38°S-55°S). *Int. J. Remote Sens.,* **30** (1), 251 – 273.

Durant, J. I. M., Hjermann, D. Ø., Anker - Nilssen, T., Beaugrand, G., Mysterud, A., Pettorelli, N. and Stenseth N. C. (2005) Timing and abundance as key mechanisms affecting trophic interactions in variable environments. *Ecology Letters*, **8**, 952 – 958.

Edwards, K. A., Rogerson, A. M., Winant, C. D. and Rogers, D. P. (2001) Adjustment of the marine atmospheric boundary layer to a coastal cape. *J. Atmos. Sci.*, **58**, 1511 – 1528.

Eppley, R. W. (1972) Temperature and phytoplankton growth in the sea. Fishery Bulletin, **70** (4), 1063 – 1085.

Esaias, W. E. (1980) Remote sensing of oceanic phytoplankton: present capabilities and future goals. In Falkowski, P. G. (ed) *Primary productivity in the sea*. Plenum Press, New York, pp. 321–337.

Esaias, W. E., Abbott, M. R., Barton, I., Brown, O. B., Campbell, J. W., Carder, K. L., Clark, D, K., Evans, R. H., Hoge, F. E., Gordon, H. R., Balch, W. M., Letelier, R. and Minnett, P. J. (1998) An Overview of MODIS Capabilities for Ocean Science Observations. *Geoscience and Remote Sensing, IEEE Transactions on*, **36** (4), 1250 – 1265.

Falkowski, P. G. (1981) Light-shade adaptation and assimilation numbers. *J. Plankton Res.,* **3**, 203 – 216.

Falkowski, P. G., Barber, R. T. and Smetacek, V. (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science*, **281**, 200 – 206.

Falkowski, P. G. and LaRoche, J. (1991) Acclimation to spectral irradiance in algae. *J. Phycol.*, **27**, 8 – 14.

Falkowski, P. G., Owens, T. G., Ley, A. C. and Mauzerall, D. C. (1981) Effects of growth irradiance levels on the ratio of reaction centers in two species of marine phytoplankton. *Plant Physiol.*, **68**, 969 – 973.

Falkowski, P. G., Wyman, K., Ley, A. C. and Mauzerall, D. (1986) Relationship of steady state photosynthesis to fluorescence in eucaryotic algae. *Biochim. Biophys. Acta,* **849**, 183–192.

Fujiki, T. and Taguchi, S. (2002) Variability in chlorophyll *a* specific absorption coefficient in marine phytoplankton as a function of cell size and irradiance. *J. Plankton Res.,* **24** (9), 859 – 874.

Gautier, C., Diak, G. and Masse, S. (1980) A simple physical model to estimate incident solar radiation at the surface from GOES satellite data. *J. Appl. Meteor*., **19**, 1005 – 1012.

Ghahramani, G. (1998) Learning dynamic Bayesian networks. In Giles, C. L. and Gori, M. (eds), *Adaptive Processing of Temporal Information*. Springer, New York, pp. 168 – 197.

Giles-Guzman, A. D. and Alvarez-Borrego, S. (2000) Vertical attenuation coefficient of photosynthetically active radiation as a function of chlorophyll concentration and depth in Case 1 waters. *Appl. Optics,* **39**, 1351 – 1358.

Gordon, H. R. (1989) Can the Lambert-Beer law be applied to the diffuse attenuation coefficient of ocean water? *Limnol. Oceanogr.,* **34** (8), 1389 – 1409.

Gordon, H. R. and McCluney, W. R. (1975) Estimation of the depth of sunlight penetration in the sea for remote sensing. *Appl. Optics*, **14**, 413 – 416.

Gorzoli, S. L. and Gordon, A. L. (1996) Origins and variability of the Benguela Current. *J. Geophys. Res.*, **101**, 897 – 906.

Gran, H. H. (1912) Pelagic plant life. In Murry, J. and Hjort, J. (eds), *The Depths of the Ocean.* MacMillan and Co., London, pp. 307 – 387.

Gran, H. H. (1931) On the conditions for the production of plankton in the sea. *Rapp. Proc-Verb. Réun. Cons. Int. Explor. Mer*, **7**, 343 – 358.

Hagen, E., Schemainda, R., Michelchen, N., Postel, L., Schulz, S. and Below, M. (1981) Zur küstensenkrechten struktur des kaltwasserauftriebs vor der küste Namibias. *Geod. Geoph. Veröff*, **36**, 1-99.

Hardman-Mountford, N. J., Richardson, A. J., Agenbag, J. J., Hagen, E., Nykjaer, L., Shillington, F. A. and Villacastin, C. (2003) Ocean climate of the South East Atlantic observed from satellite data and wind models. *Prog. Oceanogr.*, **59**, 181 – 221.

Hart, T. J. and Currie, R. J. (1960) The Benguela Current. *Discovery Report*, **31**. 123 – 298.

Heckerman, D. (1997) Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, **1** (1), 79 – 119.

Hewitson, B. C. and Crane, R. G. (2002) Self-organizing maps: Applications to synoptic climatology. *Clim. Res.*, **22**, 13 – 26.

Holden, C. J. (1985) Currents in St Helena Bay inferred from radio-tracked drifters. In L. V. Shannon (ed.), *South Africa Ocean Colour and Upwelling Experiment*, Sea Fisheries Research Institute, Cape Town, pp. 97 – 109.

Hooker, J. D. (1874) On the marine algae of St. Thomas and the Bermudas, and on *Halophila baillonis* Asch. Contributions to the botany of the expedition of H.M.S. Challenger. *J. Linn. Soc. Bot.,* **14**, 311 – 317.

Horstman, D. A. (1981) Reported red-water outbreaks and their effect on fauna of the west coast of South Africa, 1959-1980. *Fish. Bull. of S. Afr.*, **15**, 271 – 278.

Hu, *Y*-H., Chen, *Y*-L. and Lin, E. H. (2007) Classification of time-sequential attributes by using sequential pattern rules. *Fuzzy Systems and Knowledge Discovery, Fourth International Conference on,* **2**, 735 – 739.

Hutchings, L. (1992) Fish harvesting in a variable productive habitat-searching for rules or searching for exceptions. In Payne, A. I. L., Brink, K. H., Mann, K. H. and Hilborn, R. (eds), *Benguela Trophic Functioning*. *S. Afr. J. Marine Sci.,* **12**, pp. 297 – 318.

Hutchings, L., Holden, C. and Mitchell - Innes, B. (1984) Hydrological and biological shipboard monitoring of upwelling off the Cape Peninsula. *S. Afr. J. Marine Sci*., **30**, 83 – 89.

Hutchings, L. and Nelson, G. (1985) The influence of environmental forcing on the Cape pelagic fishery. In Bas, C. Margalef, R. and Rubiés, P. (eds), *International Symposium on Upwelling off Western Africa (Cape Blanco and Benguela)*. *Inst. Inv. Pesq. Barcelona,* **1**, pp. 523 – 540.

Hutchings, L. and Taunton-Clark, J. (1990) The monitoring of gradual change in areas of high mesoscale variability. *S. Afr. J. Marine Sci.,* **86**, 9 – 37.

Jarre, A., Moloney, C. L., Shannon, L. J., Freon, P., van der Lingen, C. D., Verheye, H., Hutchings, L., Roux, J. P. and Cury, P. (2006) Detecting and forecasting long - term ecosystem changes. In Shannon, V., Hempel, G., Malanotte-Rizzoli, P. Moloney, C. L. and Woods, J. (eds), *Benguela: Predicting a Large Marine Ecosystem*, *Large Marine Ecosystems*, **14**, Elsevier, Amsterdam, pp. 239 – 272.

Jassby, A. D. and Platt, T. (1976) Mathematical formulation for the relationship between photosynthesis and light for phytoplankton. *Limnol. Oceanog*r., **21**, 540 – 547.

Jin, X., Dong, C., Kurian, J., McWilliams, J. C., Chelton, D. B. and Li, Z. (2009) SST - wind interaction in coastal upwelling: Oceanic simulation with empirical coupling. *J. Phys. Oceanogr.*, **39** (11), 2957 – 2970.

Jury, M. R. (1980) Characteristics of summer wind fields and air-sea interactions over the Cape Peninsula upwelling region. M.Sc. thesis, University of Cape Town,

Jury, M. R. (1985) Case studies of alongshore variations in wind-driven upwelling in the southern Benguela region. In Shannon, L. V. (ed.), *South African Ocean Colour and Upwelling Experiment.* Sea Fisheries Research Institute, Cape Town, pp. 29 – 46.

Jury, M. R. (1988) A climatological mechanism for wind-driven upwelling near Walker Bay and Danger Point, South Africa*. S. Afr. J. Marine Sci*., **6**, 175 – 181.

Jury, M. R. and Brundrit, G. B. (1992) Temporal organization of upwelling in the southern Benguela ecosystem by resonant coastal trapped waves in the ocean and atmosphere, *S. Afr. J. Marine Sci.,* **12**, 219 – 224.

Kiørboe, T. (1993) Turbulence, phytoplankton cell size and the structure of the pelagic food webs. *Adv. Mar. Biol.*, **29**, 1 – 72.

Kirk, J. Y. O. (1976) A theoretical analysis of the contribution of algal cells to the attenuation of light within natural waters. III. Cylindrical and spheroid cells. *New Phytol*., **77**, 341 – 358.

Kirk, J. T. O. (1994) *Light and photosynthesis in Aquatic Ecosystems*, 2nd edn. Cambridge University Press, Cambridge.

Koblentz-Mishke OJ, Volkovinsky VV, Kabanova JG (1970) Plankton primary production of the world ocean. In Wooster, W. S. (ed) *Scientific exploration of the South Pacific.* U.S. National Academy of Science, Washington, pp. 183–193.

Kohonen, T. (1995) *Self-organizing Maps*. Springer Series in Information Sciences, **30**, Berlin.

Kolber, Z., Wyman, K. D. and Falkowski, P. G. (1990) Natural variability in photosynthetic energy conversion efficiency: a field study in the Gulf of Maine. *Limnol. Oceanogr.,* **35**, 72–79.

Kryszczuk, K. and Hurley, P. (2010). Estimation of the number of clusters using multiple clustering validity indices. In *Multiple Classifier Systems*. Springer Berlin Heidelberg, 114-123.

Kyewalyanga, M. N., Platt, T. and Sathyendranath, S. (1992) Ocean primary production calculated by spectral and broad - band models. *Mar. Ecol. Prog. Ser*., **85**, 171 – 185.

Kyewalyanga, M. N., Platt, T., Sathyendranath, S., Lutz, V. A. and Stuart, V. (1998) Seasonal variations in physiological parameters of phytoplankton across the North Atlantic. *J. Plankton Res.*, **20**, 17 – 42.

Lafferty, J., McCallum, A. and Pereire, F. (2001) Conditional Random Fields; Probabilistic Models for Segmenting and labelling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Morgan Kaufmann, MA, pp. 282 – 289.

Lamont, T. (2011) Bio-optical investigation of phytoplankton production in the southern Benguela ecosystem. Ph.D thesis, University of Cape Town, viii + pp. 129.

Lande, R. and Yentsch, C. S. (1988) Internal waves, primary production and the compensation depth of marine-phytoplankton. *J. Plankton Res.*, **10**, 565 – 571.

Largier, J. L., Chapman, P., Peterson, W. T. and Swart, V. P. (1992) The Western Agulhas Bank: Circulation, stratification and ecology. In Payne, A. I. L., Brink, K. H.,

Mann, K. H. and Hilborn, R. (eds), *Benguela Trophic Functioning*. *S. Afr. J. Marine Sci.,* **12**, pp. 319 – 339.

Largier, J. L. and Swart, V. P. (1987) East-west variation in thermocline breakdown on the Agulhas Bank. In Payne, A. I. L., Gullend, J A. and Brink, K. H. (eds), *The Benguela and Comparable Ecosystems. S. Afr. J. Marine Sci*. **5**, pp. 263 – 272.

Lasker, R. (1975) Field criteria for survival of anchovy larvae: The relation between inshore chlorophyll maximum layers and successful first feeding. *Fish. Bull.*, **73** (3), 453 – 462.

Lee, Z-P., Darecki, M., Carder, K. L., Davis, C. O., Stramski, D. and Rhea, W. J. (2005a) Diffuse attenuation coefficient of downwelling irradiance: An evaluation of remote sensing method. *J. Geophys. Res.,* **110**, CO2017.

Lee, Z-P., Du, K., Arnone, R., Liew, S. C. and Penta, B. (2005b) Penetration of solar radiation in the upper ocean: A numerical model for oceanic and coastal waters. *J. Geophys. Res.,* **110**, C09019.

Lee, Z-P., Weidemann, A., Kindle, J., Arnone, R., Carder, K. L. and Davis C. (2007) Euphotic zone depth: Its deviation and implication to ocean-colour remote sensing. *J. Geophys. Res.,* **112**, C03009.

Lewis, M. R., Cullen, J. J. and Platt, T. (1983) Phytoplankton and thermal structure in the upper ocean: Consequences of nonuniformity in chlorophyll profile. *J. Geophys. Res.,* **88** (C4), 2565 – 2570.

Liu, K-K., Chao, S-Y., Shaw, P-T., Gong, G-C., Chen, C-C. and Tang, T-Y. (2002) Monsoon-forced chlorophyll distribution and primary production in the South China Sea: Observations and numerical study. *Deep-Sea Res. I*, **49**, 1387 – 1412.

Liu, B., Hsu, W. and Ma, Y. (1998) Integrating classification and association rule mining. In *Proceedings of the 4th International conference on Knowledge discovery and Data Mining (KDD-98)*, AAAI Press, pp. 80 – 86.

Liu, B., Ma, Y. and Wong, C. K. (2003) Scoring the data using association rules. *Appl. Intelligence*, **18**, 119 – 135.

Lloyd, S. P. (1982) Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, **28**, 129 – 137.

Longhurst, A. (1998) *Ecological Geography of the Sea.* Academic Press, San Diego.

Longhurst, A., Sathyendranath, S., Platt, T. and Caverhill, C. M. (1995) An estimation of global primary production in the ocean from satellite radiometer data. *J. Plankton Res.,* **17**, 1245 – 1271.

Lutjeharms, J. R. E., Cooper, J. and Roberts, M. (2000) Upwelling at the inshore edge of the Agulhas Current. *Cont. Shelf Res.,* **20**, 737 – 761.

Lutjeharms, J. R. E. and Meeuwis, J. M. (1987) The extent and variability of south-east Atlantic upwelling. In Payne, A. I. L., Gulland, J. A. Brink, K. H. (eds), *The Benguela and Comparable Frontal Ecosystems*. *S. Afr. J. Marine Sci*., **5**, 51 – 62.

Lutjeharms, J. R. E. and Stockton, P. L. (1991) Aspects of the upwelling regime between Cape Point and Cape Agulhas, South Africa. *S. Afr. J. Marine Sci.,* **10**, 91 – 102.

Lutjeharms, J. R. E. and Valentine, H. R. (1987) Water types and volumetric considerations of the South-East Atlantic upwelling regime. *S. Afr. J. Mar. Sci.*, **5**, 63-71.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability,* **1**, pp. 281 – 296.

Mann, K. H. (2000) *Ecology of Coastal Waters, with Implications for Management*, (2nd edn.). Blackwell Science, Boston.

Mann, K. H. and Lazier, J. R. N. (1996) *Dynamics of Marine Ecosystems: Biological-Physical Interactions in the Ocean*. Blackwell Science Inc., Oxford.

Marshall, S. M. and Orr, A. P. (1927) The relation of the plankton to some chemical and physical factors in the Clyde Sea area. *J. Mar. Biol. Ass. U. K.,* **14**, 837 – 868.

Marshall, S. M. and Orr, A. P. (1928) The photosynthesis of diatom cultures in the sea. *J. Mar. Biol. Ass. U. K.,* **15**, 321 – 360.

Mitchell-Innes, B. A. and Pitcher, G. C. (1992) Hydrographic parameters as indicators of the suitability of phytoplankton populations as food for herbivorous copepods. In Payne, A. I. L., Brink, K. H., Mann, K. H. and Hilborn, R. (eds), *Benguela Trophic Functioning*. *S. Afr. J. Marine Sci.,* **12**, pp. 355 – 365.

Mitchell-Innes, B. A., Pitcher, G. C. and Probyn, T. A. (2000) Productivity of dinoflagellate blooms on the west coast of South Africa, as measured by natural fluorescence. *S. Afr. J. Marine Sci*., **22** (1), 273 – 284.

Mitchell-Innes, B. A., Richardson, A. J. and Painting, S. J. (1999) Seasonal changes in phytoplankton biomass on the Western Agulhas Bank, South Africa. *S. Afr. J. Marine Sci.,* **21**, 217 – 233.

Mitchell-Innes, B. A., Silulwane, N. F. and Lucas, M. I. (2001) Variability of chlorophyll profiles on the west coast of southern Africa in June/July 1999*. S. Afr. J. Marine Sci*., **97**, 24 – 250.

Mitchell-Innes, B. A. and Walker, D. R. (1991) Short-term variability during an anchor station study in the southern Benguela upwelling system: Phytoplankton production and biomass in relation to species changes. *Prog. Oceanogr*., **28**, 65 – 89.

Montecino, V., Astoreca, R., Alarcón, G., Retamal, L. and Pizarro, G. (2004) Bio-optical characteristics and primary productivity during upwelling and non-upwelling conditions in a highly productive coastal ecosystem off central Chile (~ 36°S). *Deep-Sea Res*., **51**, 2413 – 2426.

Moore, T. S., Campbell, J. W. and Dowell, M. D. (2009) A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. *Remote Sens. Environ.*, **113**, 2424 – 2430.

Morel, A. (1988) Optical modelling of the upper ocean in relation to its biogenous matter content (Case 1 waters). *J. Geophys. Res.,* **93**, 10749 – 10768.

Morel, A. and Berthon, J-F. (1989) Surface pigments, algal biomass profiles, and potential production of the euphotic layer: Relationships reinvestigated on view of remote - sensing applications. *Limnol. Oceanogr.,* **34**, 1545 – 1562.

Morel, A. and Bricaud, A. (1981) Theoretical results concerning light absorption in a descrete medium and application to specific absorption of phytoplankton. *Deep-Sea Res*., **28**, 1375 – 1393.

Morel, A., Huot, Y., Gentili, B., Werdell, P. J., Hooker, S. B. and Franz, B. A. (2007) Examining the consistency of products derived from various ocean color sensors in

open ocean (Case 1) waters in the perspective of a multi-sensor approach. *Remote Sens. Environ*., **3**, 69 – 88.

Morel, A. and Maritorena, S. (2001) Bio-optical properties of oceanic waters: A reappraisal. *J. Geophys. Res.,* **106**, 7163 – 7180.

Moroshkin, K. V., Bunov, V. A. and Bulatov, R. P. (1970) Water circulation in the eastern South Atlantic Ocean. *Oceanology*, **10**, 27 – 34.

Mueller, J. L. (2000) SeaWiFS algorithm for the diffuse attenuation coefficient, Kd(490) using water-leaving radiance at 490 and 555 nm. SeaWiFS postlaunch technical report series, **11** (3), 24 – 27.

Nathansohn, A. (1906) Über die Bedeutung vertikaler Wasserbewegungen für die Produktion des Planktons im Meere. *Königl. Sächs. Gesellsch. d. Wissensch., Liepzig, Abhandl. d. Math.-Phys. Klasse*, Bd. **29**, no. 5.

Nelson, G. (1992) Equatorward wind and atmospheric pressure spectra as metrics for primary productivity in the Benguela system, *S. Afr. J. Marine Sci.*, **12**, 19 – 28.

Nelson, G. and Hutchings, L. (1983) The Benguela upwelling area. *Prog. Oceanogr.,* **12**, 333 – 356.

O'Reilly, J. J. (2000) SeaWiFs postlaunch calibration and validation analyses: Part 3. *NASA Tech. Memo., 206892*, National Aeronautics and Space Administration, Goddard Flight Center.

Ørsted, A. S. (1844) De regionibus marinus. Elementa topographiae historiconaturalis freti Öresund. Diss. Inaug. Havniae.

Pearl, J. (1988) *Probabilistic Reasoning in Expert Systems.* Wiley, New York.

Peterson, W. T. and Painting, S. J. (1990) Developmental rates of the copepods *Calanus australis* and *Calanoides carinatus* in the laboratory, with discussion of methods used for calculation of development time. *J. Plankton. Res.*, **12** (2), 283 – 293.

Peterson, R. G. and Stramma, L. (1991) Upper-level circulation in the South Atlantic Ocean. *Prog. Oceanogr.*, **26**, 1 – 73.

Pierson, D. C., Kratzer, S., Strömbeck, N. and Håkansson, B. (2008) Relationship between the attenuation of downwelling irradiance at 490 nm with the attenuation of PAR (400 nm-700nm) in the Baltic Sea. *Remote Sens. Environ.*, **112**, 668 – 680.

Pinto, D., McCallum, A., Wei, X. and Croft, W. (2003) Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'03).* ACM, pp. 235 – 242.

Pitcher, G. C. (1986) Sedimentary flux and formation of resting spores of selected *Chaetoceros* species at two sites in the southern Benguela system. *S. Afr. J. Marine Sci.*, **4**, 231 – 244.

Pitcher, G. C. (1988) Mesoscale heterogeneities of the phytoplankton distribution in St Helena Bay, South Africa, following an upwelling event. *S. Afr. J. Marine Sci.*, **7**, 9 – 23.

Pitcher, G. C., Brown, P. C. and Mitchell-Innes, B. A. (1992) Spatio-temporal variability of phytoplankton in the southern Benguela upwelling system. In Payne, A. I. L., Brink, K. H., Mann, K. H. and Hilborn, R. (eds), *Benguela Trophic Functioning*. *S. Afr. J. Marine Sci*. **12**, pp. 439 – 456.

Pitcher, G. C. and Nelson, G. (2006) Characteristics of the surface boundary layer important to the development of red tide on the southern Namaqua shelf of the Benguela upwelling system. *Limnol. Oceanogr.*, **51** (6), 2660 – 2674.

Pitcher, G. C., Richardson, A. J. and Korrubel, J. L. (1996) The use of sea temperature in characterizing the mesoscale heterogeneity of phytoplankton in an embayment of the southern Benguela upwelling system. *J. Plankton Res.,* **18** (5), 643 – 657.

Pitcher, G. C., Walker, D. R. and Mitchell-Innes, B. A. (1989) Phytoplankton sinking rate dynamics in the southern Benguela upwelling system. *Mar. Ecol. Prog. Ser*., **55**, 261 – 269.

Pitcher, G. C., Walker, D. R., Mitchell-Innes B. A. and Moloney, C. (1991) Short-term variability during an anchor station in the southern Benguela upwelling system: Phytoplankton dynamics. *Prog. Oceanogr.,* **28**, 39 – 64.

Platt, T., Caverhill, C. M. and Sathyendranath, S. (1991) Basin-scale estimates of oceanic primary production by remote sensing: The North Atlantic. *J. Geophys. Res.,* **96**, 15147 – 15159.

Platt, T., Fuentes-Yaco, C. and Frank, K. T. (2003) Spring algal bloom and larval fish survival. *Nature*, **423**, 398 – 399.

Platt, T., Denman, K. L. and Jassby, A. D. (1975) The mathematical representation and prediction of phytoplankton productivity. Fisheries and Marine Services Technical Report 523

Platt, T., Gallegos, C. L. and Harrison, W. G. (1980) Photoinhibition of photosynthesis in natural assemblages of marine phytoplankton. *J. Mar. Res.,* **38**, 687 – 701.

Platt, T. and Sathyendranath, S. (1988) Oceanic primary production: Estimation by remote sensing at local and regional scales. *Science*, **241**, 1613 – 1620.

Platt, T. and Sathyendranath, S. (1993) Estimators of primary production for interpretation of remotely sensed data on ocean color. *J. Geophys. Res.* **98**, 14561 – 14576.

Platt, T., Sathyendranath, S., Caverhill, C. M. and Lewis, M. R. (1988) Ocean primary production and available light: Further algorithms for remote sensing. *Deep-Sea Res.,* **35**, 855 – 879.

Qin, Z. (2012) A Method for Determining Optimal Number of Clusters Based on K-means Algorithm. In *Proceedings of the 2012 International Conference on Electronics, Communications and Control*. IEEE Computer Society, 2457-2460.

Platt, T., Sathyendranath, S. and Longhurst, A. (1995) Remote sensing of primary production in the ocean: Promise and fulfilment. *Philos. Trans. R. Soc. London Ser. B,* **348**, 191 – 200.

Platt, T., Sathyendranath, S., Ulloa, O., Harrison, W. G., Hoepffner, N. and Goes J. (1992) Nutrient control of phytoplankton photosynthesis in the western North Atlantic. *Nature,* **356**, 229 – 331.

Preston-Whyte, R. A. and Tyson, P. D. (1988) *The Atmospheric and Weather of Southern Africa*. Oxford University Press, Cape Town.

Probyn T. A., Mitchell-Innes, B. A., Brown, P. C., Hutchings, L. and Carter, R. A. (1994) A review of primary production and related processes on the Agulhas Bank. *S. Afr. J. Marine Sci.*, **90**, 166 – 174.

Renault, L., Dewitte, B., Falvey, M., Garreaud, R., Echevin, V. and Bonjean, F. (2009) Impacts of atmospheric coastal jets in SST off central Chile from satellite observations (2000-2007). *J. Geophys. Res.*, **114**, C08006.

Richardson, A. J., Pfaff, M. C., Field, J. G., Siluwane, N. F. and Shillington, F. A. (2002) Identifying characteristics chlorophyll a profiles in the coastal domain using an artificial neural network. *J. Plank. Res.,* **24** (12), 1289 – 1303.

Richardson, A. J., Silulwane, N. F., Mitchell-Innes, B. A. and Shillington, F. A. (2003) A dynamic quantitative approach for predicting the shape of phytoplankton profiles in the ocean. *Prog. Oceanogr*., **59**, 301 – 319.

Richardson, A. J. and Verheye, H. M. (1998) The relative importance of food and temperature to copepod egg production and somatic growth in the southern Benguela upwelling system. *J. Plankton Res.*, **20** (12), 2379 – 2399.

Riley, G. A. (1956) Oceanography of Long Island Sound, 1952-54. Production and utilization of organic matter. *Bull Bingham Oceanogr. Collect.,* **15**, 324 – 343.

Roy, C., Weeks, S., Rouault, M., Nelson, G., Barlow, R. and van der Lingen, C. (2001) Extreme oceanographic events recorded in the Southern Benguela during the 1999-2000 summer season. *S. Afr. J. Marine Sci.,* **97**, 465 – 471.

Rykaczewski, R. R. and Checkley, D. M. (2008) Influence of ocean winds on the pelagic ecosystem in upwelling regions. *Proc. Nat. Acad. Sci*., **105** (6), 1965-1970.

Ryther, J. H. (1969) Photosynthesis and fish production in the sea. *Science,* **166**, 72 – 76.

Sathyendranath, S., Lazzara, L. and Prieur, L. (1987) Variations in the spectral values of specific absorption of phytoplankton. *Limnol. Oceanogr*., **32**, 403 – 415.

Sathyendranath, S., Longhurst, A. R., Caverhill, C. M. and Platt, T. (1995) Regionally and seasonally differentiated primary production in the North Atlantic. *Deep-Sea Res.,* **42**, 1773 – 1802.

Schumann, E. H. (1998) The coastal ocean off southeast Africa, including Madagascar. In A. R. Robinson, A. R. and K. H. Brink, K. H. (eds), *The Sea, Vol. II, The global coastal ocean, regional studies and synthesis*. Wiley, New York, pp. 557 – 581.

Schwarz, J. N., Raymond, B., Marsland, S. J., Williams, G. D. and Pasquer, B. (2010) Biophysical coupling in remotely-sensed wind stress, sea surface temperature, sea ice and chlorophyll concentrations in the South Indian Ocean. *Deep-Sea Res. II*, **57** (9-10), 701 – 722.

Sha, F. and Pereira, F (2003) Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003).* Association for Computational Lingistics, Morristown, NJ, pp. 134 – 141.

Shannon, L. V. (1966) *Hydrology of the south west coasts of South Africa*. Department of Commerce and Industries, Division of Sea Fisheries, South Africa.

Shannon, L. V. (1985) The Benguela ecosystem, Vol. 1. Evolution of the Benguela, physical features and processes. *Oceanogr. Mar. Biol. Ann. Rev.,* **23**, 105 – 182.

Shannon, L. V. and Field, J. G. (1985) Are fish stocks food-limited in the southern Benguela pelagic ecosystem? *Mar. Ecol. Prog. Ser.*, **22** (1), 7 – 19.

Shannon, L. V. and Nelson, G. (1996) The Benguela: Large scale features and processes and system variability. In Wafer, G., Berger, W. H., Siedler, G. and Web, D. J. (eds), *The South Atlantic: Present and past circulation*. Springer-Verlag, Berlin, pp. 163 – 210.

Shannon, L. V., Nelson, G. and Jury, M. R. (1981) Hydrological and meteorological aspects of upwelling in the southern Benguela Current. In Richards, F. A. (ed.), *Coastal and Estuarine Sciences (1). Coastal Upwelling.* American Geophysical Union, Washington DC, pp. 146 – 159.

Schoenberg, I. J. (1946) Contributions to the problem of approximation of equidistant data by analytic functions, *Quart. Appl. Math*., **4**, 45–99 and 112–141.

Shelton, P. A. and Hutchings, L. (1990) Ocean stability and anchovy spawning in the southern Benguela Current region. *Fish. Bull.*, **88**, 323 – 338.

Silio-Calzada, A., Bricaud, A., Uitz, J. and Gentili, B. (2008) Estimation of new primary production in the Benguela upwelling area, using ENVISAT satellite data and a model dependant on the phytoplankton community size structure. *J. Geophys. Res.,* **113,** C11023.

Silulwane, N. F., Richardson, A. J., Shillington, F. A. and Mitchell-Innes, B. A. (2001). Identification and classification of vertical chlorophyll patterns in the Benguela upwelling system and Angola-Benguela front using an artificial neural network. In Pillar, S. C. and Crawford, R. J. M. (eds), *A Decade of Namibian Fisheries Science. S. Afr. J. Marine Sci.,* **23,** 37–51.

Smith, E. L. (1936) Photosynthesis in relation to light and carbon dioxide. *Proc. Natl. Acad. Sci.*, **22**, 504 – 511.

Steele, J. H. and Henderson, E. W. (1981) A simple plankton model. *Am. Nat.*, **117**, 676 – 691.

Steemann-Nielsen, E. (1952) The use of radio-active carbon (^{14}C) for measuring organic production in the sea. *Journal du Conseil International pour Exploration de la Mer*, **18**, 117–140 .

Strub, P. T., Mesías, J. M., Montecino, V., Rutllant, J. and Salinas, S. (1998) Coastal ocean circulation off western South America. In Robinson, A. R. and Brink, K. H. (eds), *The Sea*. *Vol. II. The global coastal ocean. Regional studies and synthesis.* John Wiley and Sons Ltd, New York, pp. 273 – 313.

Sutton, C. and McCallum, A. (2006) An introduction to conditional random fields for rational learning. In Getoor, L. and Taskar, B. (eds) *Introduction to Statistical Relational Learning*. MIT Press, pp. 142 – 146.

Sverdrup, H. U. (1953) On the condition for vernal blooming of the phytoplankton. *J. Cons. Perm. Int. Explor. Mer.,* **18**, 287 – 295.

Taguchi, S., Kasai, H. and Saito, H. (1994) Estimation of vertical distribution of chlorophyll *a* off east Hokkaido by Gaussian curve fitting. *Proc. NIPR Symp. Polar Biol*., **7**, 17 – 31.

Tamiya, H. (1951) Some theoretical notes on the kinetics of algal growth. *Botanical Magazine*, **6**, 167 - 173.

Taunton-Clark, J. and Kamstra, F. (1988) Aspects of marine environment variability near Cape Town, 1960-1985. *S. Afr. J. Marine Sci*., **6**, 273 - 283.

Tseng, V. S. and Lee, C-H. (2009) Effective temporal data classification by integrating sequential pattern mining and probabilistic induction. *Expert Syst. Appl.*, **36 (5)**, 9524 – 9532.

Tsuruoka, Y., Tsujii, J. and Ananiadou, S. (2009) Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty. In *Proceedings of the 47th Annual Meeting of the ACL, and the 4th IJCNLP of the AFNLP*, pp. 477 – 485.

Tyson, P. D. (1986) *Climate change and variability in Southern Africa. Cape Town*. Oxford University Press, London.

Uitz, J., Claustre, H., Morel, A. and Hooker, S. (2006) Vertical distribution of phytoplankton communities in open the ocean: An assessment based on surface chlorophyll. *J.Geophys. Res*., **III**, C08005.

Uitz, J., Huot, Y., Bruyant, F., Babin, M. and Claustre, H. (2008) Relating phytoplankton photophysiological properties to community structure on large scales. *Limnol. Oceanogr.*, **53**, 614 – 630.

van der Lingen, C. D., Shannon, L. J., Cury, P., Kreiner, A., Moloney, C. L., Roux J-P. and Vaz-Velho, F. (2006) Resource and ecosystem variability, including regime shifts in the Benguela Current System. In Shannon, V., Hempel, G., Malanotte-Rizzoli, P., Moloney, C. L. and Woods, J. (eds), *Benguela: Predicting a Large Marine Ecosystem. Large Marine Ecosystems*, **14**, Elsevier, Amsterdam, pp. 147 – 185.

Viterbi, A. J. (1967) Error bounds for convolutional codes and an asymtotically optimal decoding algorithm. *IEEE Trans. Inf. Theory*, 260 – 269.

Waldron, H. N. and Probyn, T. A. (1992) Nitrate supply and potential production in the Benguela upwelling system. In Payne, A. I. L., Brink, K. H., Mann, K. H. and Hilborn, R. (eds), *Benguela Trophic Functioning. S. Afr. J. Marine Sci*., **12**, 865 – 871.

Walker, N. D. (1986) Satellite observations of the Agulhas Current and episodic upwelling south of Africa. *Deep-Sea Res*., **33**, 1083 – 1106.

Walker, D. R. and Peterson, W. T. (1991) Relationships between hydrography, phytoplankton production, biomass, cell size and species composition, and copepod production in the southern Benguela upwelling system in April 1988. *S. Afr. J. Marine Sci*., **11**, 289 – 305.

Wallach, H. M. (2004) Conditional Random Fields; An Introduction. *University of Pensylvania CIS Techincal Report MS-CIS-04-21*.

Weeks, S. J., Barlow, R., Roy, C. and Shillington, F. A. (2006) Remotely sensed variability of temperature and chlorophyll in the southern Benguela: Upwelling frequency and phytoplankton response. *Afr. J. Marine Sci.,* **28** (3-4), 493 – 509.

Weeks, S. J. and Shillington, F. A. (1994) Interannual scales of variation of pigment concentrations from CZCS data in the Benguela upwelling system and the subtropical convergence zone south of Africa. *J. Geophys. Res*., **99**, 7385 – 7400.

Williamson, R. I., Field, J. G., Shillington, F. A., Jarre, A. and Potgieter, A. (2011) A Bayesian approach for estimating vertical chlorophyll profiles from satellite remote sensing: proof-of-concept. *ICES J. Mar. Sci.*, **68** (4), 792 – 799.

Wooster, W. S. and Reid, J. L. (1963) Eastern boundary currents. In Hill, M. N. (ed.) *The Sea Vol. II.* Interscience, New York, pp. 253 – 280.

Zaneveld, J. R. V. and Kitchen, J. C. (1993) Vertical structure of productivity and its vertical integration as derived from remotely sensed observations. *Limnol. Oceangr.*, **38**, 1384 – 1393.

Zhang, H-M., Bates, J. J. and Reynolds, R. W. (2006) Assessment of composite global sampling: Sea surface wind speed. *Geophys. Res. Lett.,* **33**, L17714.