Modelling Multivariate Spatio-Temporal Structure in Ecological Data and Responses to Climate Change

A thesis presented for the degree of

Doctor of Philosophy of University College London

by

Victoria Harris

Department of Statistics and the Centre for Mathematics and Physics in the Life Sciences and Experimental Biology

University College London

Gower Street, London, WC1E 6BT

FEBRUARY 21, 2013

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed:

Copyright

Acknowledgements

I would like to thank my supervisors Sofia Olhede and David Murrell for their support and guidance over the years. In particular for sharing their wealth of knowledge in Statistics and Ecology respectively to help me overcome the challenges of interdisciplinary research. I also offer my gratitude to the staff at the Sir Alister Hardy Foundation for Ocean Science who have offered advice at various stages of my project during my visits to the lovely city of Plymouth. In particular to Martin Edwards, who has helped me interpret my results and gain new insights in to the world of plankton. In addition I extend my thanks to David Johns, who provided me with the raw dataset, and Pierre Helaout, with whom I have had many an interesting discussion about the applications of statistics to the CPR dataset. The enthusiasm and knowledge of the staff at SAHFOS has given me an appreciation of the both importance and the fascination of plankton research.

I would like to acknowledge my family, who have provided emotional support throughout: my partner Matt and my brother Andrew, with whom I have spent many a lunch in Patisserie Valerie sharing our experiences of life in academia. Finally I would like to thank my parents, whose love and encouragment has helped make this work possible.

Victoria Harris

Table of contents

List of Publications **8**

Abstract **10**

1 Introduction **22**
1.1 Overview of the Aims of this Study and its Importance to Marine Ecology 22
1.2 The Continuous Plankton Recorder Survey 25
1.3 Modelling Spatio-Temporal Multivariate Data Sets 31
1.4 The Biogeography of the North East Atlantic and 'Regime Shifts' in Ecoregions 36
1.5 Modelling Sensitivity to Climate Across Species and Space 40
1.6 Climate Indices and their Influence on Marine Ecology 42
1.6.1 The Atlantic Multidecadal Oscillation 43
1.6.2 The Northern Hemisphere Temperature 44
1.6.3 The North Atlantic Oscillation 45
1.6.4 The East Atlantic Pattern 47
1.6.5 Physical Factors 47

2 Methods **48**
2.1 Smoothing Methods 49
2.2 Principal Component Analysis 54
2.2.1 Decompositions of the CPR Dataset 58
2.2.2 Spatial PCA 61
2.2.3 Species PCA 63
2.3 Sparse Principal Component Analysis 64
2.3.1 Using Mixture Models to Determine the Sparsity Parameter 70
2.3.2 Temporal Misalignment in Biological Data and Fourier PCA 72
2.4 Cluster Analysis 75
2.5 Regression Analysis 78
2.6 Modelling Vulnerability to Climate and Regime Changes 83
2.7 Modelling Trends and Oscillations 85

3 Multidecadal Oscillations **87**
3.1 Overview . . . 87
3.2 Methods Used in this Chapter . . . 90
3.3 Climate Trends Across the Ocean Basin Scale . . . 91
3.3.1 Sea Surface Temperature with Median only Removed . . . 91
3.3.2 Detrended Sea Surface Temperature . . . 92
3.4 Multidecadal Oscillations across the North East Atlantic . . . 97
3.5 Discussion . . . 101

4 The Interpolated Data Modelled using Sparse PCA **103**
4.1 Overview . . . 103
4.2 Methods used in this Chapter . . . 104
4.3 Sparse Principal Component Analysis on the WinCPR Data . . . 105
4.3.1 Verifying the Mixture Model using the WinCPR Data . . . 105
4.3.2 Measures of Diversity . . . 105
4.3.3 Ecoregions Defined by the Loading Vectors . . . 108
4.3.4 Regions Defined by Functional Behaviour . . . 115
4.3.5 Time Delays Across Space . . . 117
4.3.6 Relationship Between the Plankton and Climate . . . 120
4.4 Modelling Seasonal Data using the WinCPR . . . 125
4.5 Discussion . . . 131

5 Modelling the Raw CPR Data **136**
5.1 Overview . . . 136
5.2 Methods Used in this Chapter . . . 137
5.3 Spatial PCA on Indicator Species . . . 138
5.3.1 Spatial PCA for Phytoplankton Colour Index . . . 138
5.3.2 Spatial PCA for *Calanus finmarchicus* . . . 144
5.3.3 Spatial PCA for *Calanus helgolandicus* . . . 145
5.4 Species Principal Component Analysis for Spatially Averaged Data 148
5.4.1 Averaged Zooplankton Data . . . 148
5.4.2 Averaged Phytoplankton Data . . . 150
5.5 Multivariate and Spatio-Temporal Structure Modelled by Sparse PCA 152
5.5.1 Modelling Across all Species . . . 152
5.5.2 Results for the Zooplankton Group . . . 158
5.5.3 Results for the Diatom Group . . . 166
5.5.4 Spatial Structure in Relation to Bathymetry . . . 171
5.5.5 Mixing Regions as Described by Colour Plots . . . 174
5.6 Simulation Studies . . . 177
5.7 Discussion . . . 177

6 Modelling Changes in Biogeographical Regionalisation 180
6.1 Overview . . . 180
6.2 Methods Used in this Chapter . . . 181
6.3 Modelling Changes across all Species . . . 182
6.3.1 Time Courses and Sparsity Parameters Across all Species . 182
6.3.2 Changes in Regionalisation Across all Species . . . 183
6.4 Changes in the Regionalisation for the Zooplankton . . . 185
6.4.1 Non-Stationarity in the Time Courses for the Zooplankton . 185
6.4.2 Changes in Regionalisation for the Zooplankton . . . 193
6.5 Changes in the Regionalisation for the Phytoplankton . . . 196
6.5.1 Time Courses and Sparsity Parameters for the Diatoms . . . 196
6.5.2 Changes in Regionalisation for the Diatom Species . . . 202
6.6 Discussion . . . 203

7 Modelling Vulnerability to Climate Variables 209
7.1 Methods Used in this Chapter . . . 212
7.2 Sensitivity of Different Species to Climate Change . . . 214
7.2.1 The Spatial Sensitivity of *Calanus finmarchicus* . . . 214
7.2.2 The Spatial Sensitivity of *Calanus helgolandicus* . . . 216
7.2.3 The Spatial Sensitivity of *Echinoderm Larvae* . . . 219
7.3 Sensitivity of Joint Responses over Space to Climate Change . . . 221
7.4 Joint Response of the Zooplankton Group . . . 224
7.5 Joint Response of the Diatoms . . . 225
7.6 Multiscale Downscaling . . . 228
7.6.1 Zooplankton Data at Different Scales . . . 233
7.6.2 Diatom Data at Different Scales . . . 237
7.7 Discussion . . . 239

8 Conclusions 242
8.1 Discussion of the Main Results . . . 242
8.1.1 Species Structure . . . 243
8.1.2 Temporal Structure . . . 245
8.1.3 Spatial Structure . . . 247
8.2 Possibilities for Further Study . . . 251
8.3 Conclusions . . . 254

References 276

List of Publications

The following paper is in revision:

[72] Victoria Harris, Martin Edwards and Sofia C. Olhede Multidecadal Atlantic Climate Variability and its Impact on Marine Pelagic Communities. *TBA*, 2012.

Which has the provisional abstract:

A large scale analysis of sea surface temperature and climate variability over the North Atlantic and its interactions with plankton over the North East Atlantic was carried out to better understand what drives both temperature and species abundance. The spatiotemporal pattern of sea surface temperature was found to correspond to known climate indices, namely the Atlantic Multidecadal Oscillation (AMO), the East Atlantic Pattern (EAP) and the North Atlantic Oscillation (NAO). The spatial influence of these indices is heterogeneous. Although the AMO is present across all regions, it is most strongly represented in the sea surface temperature signal in the subpolar gyre region. The NAO instead is strongly weighted in the North Sea and the pattern of its influence is oscillatory in space with a period of approximately 6000 kilometres. It can further be shown that natural oscillations might obscure the influence of climate change effects, making it diffcult to determine how much of the variation is attributable to anthropogenic influences. In order to separate the influences of different climate signals the sea surface temperature signals were decomposed in to spatial and temporal components using principal components analysis (PCA). A similar analysis is carried out on various indicator species of plankton: *Calanus finmarchicus*, phytoplankton colour index and

total copepod abundance, as well as phytoplankton and zooplankton communities. By comparing the two outputs it is apparent that the dominant driver is the average warming trend, which has a negative influence on *C. finmarchicus* and total Copepods, but has a positive one on phytoplankton colour. However natural oscillations also influence the abundance of plankton, in particular the AMO is a driver of diatom abundance. Fourier principal component analysis, an approach which is novel in terms of the ecological data, is used to analyse the behavior of various communities averaged over space. The zooplankton community is found to be primarily influenced by climate warming trends. The analysis provides compelling evidence for the hypothesis that cold water species are gradually being replaced by more temperate species in the North Atlantic. This may have detrimental effects for the entire marine ecosystem, by impacting on organisms such as fish larva for example. The second group, a phytoplankton subset consisting primarily of diatom species, is primarily influenced by the AMO rather than the average temperature trend. This result highlights the importance of natural oscillations to certain functional groups, in particular those subgroups which are less directly metabolically affected by changes in temperature.

Abstract

In this study the behaviour of multivariate plankton communities and their relationships with climate is explored. Existing statistical methodology is adapted to analyse both the plankton communities and sea surface temperature. In the first part of this study a large scale exploratory analysis is applied using principal component analysis. Dominant temporal trends and spatial patterns for a number of indicator species and the joint responses of functional groups of species are found. The community analysis focuses on on the zooplankton and the phytoplankton, the latter respresented by diatoms. This research is novel because the full multivariate structure of the plankton data has not been studied across communities before. The common trends are regressed against different climate signals to determine dominant drivers and cluster analysis identifies regions based on species.

In the second part 'regime shifts' described by changes in ecoregions are explored. Whilst changes in spatial patterns over time have been studied over indicator species, this study describes the shift across communities, providing an overview of how the 'regime shift' is differently expressed for the two species groups. To explore changes in biogeographical patterns, the data is then divided in to a pre-1985 and post-1985 regimes. The results show a northwards movement of zooplankton species and increased spatial structure across the diatom group, following the bathymetry. In the final part the model is used to predict vulnerability of different indicator species and the community as a whole to changes in climate drivers across space, which is used to find climate change 'hotspots'. Vulnerability is defined as a significant change in abundance in response to a relatively small change in the climate signal. Vulnerability is also explored at different scales. These results

highlight the spatial inhomogeneity of species responses and are of great interest to environmental policy makers.

List of Tables

1 Notation used throughout this report. 21

3.1 Table summarising the Pearson's correlation coefficients between the linearly detrended sea surface temperature data over the ocean basin scale and different climate indices. 92
3.2 Table summarising the Pearson's correlation coefficients between the principal components of the sea surface temperature with the median only removed over the CPR region and the climate indices. . 97

5.1 Pearson's correlation coefficients between the first three principal components of *Calanus finmarchicus* and the climate indices. 144
5.2 Pearson's correlation coefficients between the first three principal components of *Calanus helgolandicus* and the climate indices. . . . 145
5.3 Pearson's correlation coefficients between the first three principal components of the zooplankton data averaged over space and the climate indices. 148

7.1 Table of the Pearson's correlation coefficients between the principal components of the zooplankton data over each cluster and different climate indices. 233
7.2 Table of the Pearson's correlation coefficients between the principal components of the diatom data over each cluster and different climate indices. 237

8.1 Species numbers part 1: zooplankton. 277
8.2 Species numbers part 2: phytoplankton. 278

List of Figures

1.1 Map showing the shipping routes towing the CPR device across the North Atlantic and the North Sea. 28

2.1 Plot of the raw abundances of *Calanus finmarchicus* across months for a single one degree by one degree region. 51
2.2 The CPR dataset can be thought of as a 3-dimensional space with abundance dependent on time, space and species. The different forms of PCA fix one of these dimensions. The dimensions of the resulting 2-dimensional matrix determine what the weights and signals found using PCA depend on. (a) Representation of the CPR dataset. (b) Spatial PCA. Species has been fixed and so PCA is performed on a time by location matrix. (c) Temporal PCA. Species has been fixed but now PCA is performed on a location by time matrix to give weights dependent on time. (d) Species PCA, where space has been fixed resulting in a time by species matrix. The weights found by PCA depend on species. 59

3.1 Plot of the change in sea surface temperature measured as the sea surface temperature in 2009 minus the average for 1890 till 1960 at each grid point. 90
3.2 Loading vectors plotted in space and principal components for the linearly detrended sea surface temperature. 93
3.3 Plots of the amount of variation in sea surface temperature left unexplained by the principal components modelled as responses to climate indices culminatively. 98
3.4 Plots of the first three principal components on the sea surface temperature over the CPR region. 100

4.1 Plot of the density of the absolute values of the loadings along with the densities estimate from the mixture model. The kernel density estimates of the true signal are shown in blue, the estimated probability density for the noise in green and the estimated probability density for the signal in red. 106

4.2 Plot of the sparsity parameter for the first four Fourier transformed principal components across space. Here a red pixel depicts a large value of π (so a very non-sparse solution) and a blue pixel depicts a small value. The limits are set between 0 and 0.5. The number of principal components across space is also shown, with red indicating a higher number and blue a lower value. 109
4.3 Plot of the clusters on the absolute value of the Fourier transformed loading vector. On the spatial plots of the clusters: cluster 1 is shown in light blue, cluster 2 in orange and cluster 3 in dark red. . . 114
4.4 Plot of the clusters on the first principal component for the Fourier transformed data. In the plots of the averages the centre of cluster one is shown in blue, cluster two in green and cluster three in red. . 118
4.5 Plot of the clusters on the second principal component for the Fourier transformed data. In the plots of the averages the centre of cluster one is shown in blue, cluster two in green and cluster three in red. . 119
4.6 Time delays in for the first principal component for two frequently non-zero species 121
4.7 Centres of the clusters for time delays as a function in space clustered on species and the mean squared error for the distance from the centre of the cluster plotted with species on the x-axis for principal component1. 122
4.8 Plot of the correlation coefficients between the principal components for the Fourier transformed data and the Atlantic Multidecadal Oscillation. 126
4.9 Plot of the correlation coefficients between the principal components for the Fourier transformed data and the NAO. 127
4.10 Plot of the correlation coefficients between the principal components for the Fourier transformed data and the Northern Hemisphere Temperature. 128
4.11 Plot showing the number of locations where the correlation between the first principal component and the physical variables is significant when controlling the false discovery rate for both the Fourier and non-Fourier PCA. The x-axis is the locations ordered by increasing p-value and the p-values are plotted on the y-axis. 129
4.12 Values of the sparsity parameter and number of principal components for the monthly data. 132
4.13 Clusters on the first PC and the first loading vector for the monthly data. For the centres of the clusters, cluster one is shown in blue, cluster two in green and cluster three in red. 133
4.14 Plot showing the instantaneous frequencies in blue and the median of those frequencies in red for the first and second principal components at two locations. 134

4.15 Plot showing the principal components in the Fourier domain for the first and second principal components at two locations. 135

5.1 Plot of the pixel locations for the gridded data across space. 137
5.2 Plot of the loading vector and the first principal component for phytoplankton colour. 139
5.3 Plot of the loading vector and the second principal component for phytoplankton colour. 140
5.4 Plot of the loading vector and the first principal component for *Calanus finmarchius*. 142
5.5 Plot of the loading vector and the second principal component for *Calanus finmarchius*. 143
5.6 Plot of the loading vector and the first principal component for *Calanus helgolandicus*. 146
5.7 Plot of the loading vector and the second principal component for *Calanus helgolandicus*. 147
5.8 Plots of the first three species principal components on the zooplankton subgroup averaged over the North East Atlantic. 151
5.9 Plots of the first three species principal components on the phytoplankton subgroup averaged over the North East Atlantic. 153
5.10 Plots of the Pearson correlation coefficients between the principal components for the plankton data and the principal components on the sea surface temperature data, see figure 3.4. 154
5.11 Values of the sparsity parameter and the number of components plotted for the whole of the North East Atlantic. 155
5.12 Clusters on the first principal component for all species. 156
5.13 Clusters on the real values of the first loading vector for all species. . 157
5.14 Plots of Pearson's correlation coefficient between the first principal component at each location and the climate indices. 159
5.15 Sparsity parameter and number of principal components across zooplankton species for the entire time course from 1958 till 2009. . . . 161
5.16 Plots of regions over all time based on clusters on the first loading vector. Cluster one is blue, cluster two is orange and cluster three is dark red. 162
5.17 Plots of regions based on the time courses for the first component over all time. 164
5.18 Plots of regions based on the second component over all time. Cluster one is blue, cluster two is orange and cluster three is dark red. . . 165
5.19 Plots of Pearson's correlation coefficient between the first principal component on the zooplankton species at each location and the climate indices. 167

5.20 Plots of the real parts of the weights on two indicator species for the first principal component. 168
5.21 Sparsity parameter and number of principal components across Phytoplankton species for the entire time course from 1958 till 2009. . . 169
5.22 Plots of regions based on the first component on the phytoplankton over all time. Cluster one is blue, cluster two is orange and cluster three is dark red. 170
5.23 Plots of regions based on the time courses for the first component over all time. 172
5.24 Plots of Pearson's correlation coefficient between the first principal component on the phytoplankton species at each location and the climate indices. 173
5.25 Spatial plots of the level below the sea surface, where a red pixel represents shallow waters and a blue pixel represents deeper waters. 174
5.26 RGB plots of the clusters on the real values of the first loading vector. 176
5.27 Spatial patterns on randomised data. 178
5.28 Plot of the footprint of the Kernel smoothing function. 179

6.1 Principal components all species for the entire time course from 1958 till 2009. Time course for the full dataset is shown in blue, before 1985 in green and after 1985 in red. a) Averaged first principal component across all time and for each half of the data before and after 1985 for all species. b) Averaged second principal component across all time and for each half of the data before and after 1985 for all species. c) Averaged third principal component across all time and for each half of the data before and after 1985 for all species. d) Averaged forth principal component across all time and for each half of the data before and after 1985 for all species. 186
6.2 Sparsity parameter across all species. There is a decrease in the sparsity parameter in the North West and a slight decrease in the sparsity parameter in the North Sea. 187
6.3 Number of PCs on across all species. 188
6.4 Plots of regions based on the first loading vector before and after 1985. 189
6.5 Plots of regions based on the second loading vector before and after 1985. 190

6.6 Principal components zooplankton species for the entire time course from 1958 till 2009. The time course for the entire period is shown in blue, before 1985 in green and after in red. a) Averaged first principal component across all time and for each half of the data before and after 1985 for zooplankton species. b) Averaged second principal component across all time and for each half of the data before and after 1985 for zooplankton species. c) Averaged third principal component across all time and for each half of the data before and after 1985 for zooplankton species. d) Averaged forth principal component across all time and for each half of the data before and after 1985 for zooplankton species. 197
6.7 Sparsity parameter across zooplankton species. 198
6.8 Number of PCs on across zooplankton species. 199
6.9 Plots of regions before and after 1985. 200
6.10 Plots of regions based on the time courses for the first component before and after 1985. 201
6.11 Principal components Phytoplankton species for the entire time course from 1958 till 2009. The time courses for the full dataset are shown in blue, before 1985 in green and after 1985 in red. a) Averaged first principal component across all time and for each half of the data before and after 1985 for Phytoplankton species. b) Averaged second principal component across all time and for each half of the data before and after 1985 for Phytoplankton species. c) Averaged third principal component across all time and for each half of the data before and after 1985 for Phytoplankton species. d) Averaged forth principal component across all time and for each half of the data before and after 1985 for Phytoplankton species. 204
6.12 Sparsity parameter across Phytoplankton species. 205
6.13 Number of PCs on across Phytoplankton species. 206
6.14 Plots of regions before and after 1985. 207

7.1 Plots of the climate variables and the fitted climate variables. The true climate variables are plotted in blue, the fitted variables in red and for the NHT the trend line is plotted in green. The units of the NHT and the AMO are in degrees Celsius. The NAO is taken from a dataset with standardised units. 211

7.2 Plots of the true log-abundances of Calanus finmarchicus and the log-abundances estimated from the modelled climate signals. a) Abundance of Calanus finmarchicus in 1958. b) Abundance of Calanus finmarchicus estimated from the regression model and using the modelled climate signals in 1958. c) Abundance of Calanus finmarchicus in 2008. d) Abundance of Calanus finmarchicus estimated from the regression model and using the modelled climate signals in 2008. 217
7.3 a) Plot of the standard errors of the regression model given by equation 7.3 for Calanus finmarchicus. b) A histogram of the residuals from the regression model across all locations, with different locations being denoted by different coloured bars. 217
7.4 Plots of the predicted change in logged Calanus finmarchicus abundance over 10 years under the model. a) Estimated abundance of Calanus finmarchicus under a 1 degree increase in NHT over 10 years. b) Abundance of Calanus finmarchicus estimated from the regression model in 2008 minus the estimated abundance under a 1 degree increase in NHT over 10 years. 218
7.5 Plots of the true log-abundances of Calanus helgolandicus and the log-abundances estimated from the modelled climate signals. a) Abundance of Calanus helgolandicus in 1958. b) Abundance of Calanus helgolandicus estimated from the regression model and using the modelled climate signals in 1958. c) Abundance of Calanus helgolandicus in 2008. d) Abundance of Calanus helgolandicus estimated from the regression model and using the modelled climate signals in 2008. 219
7.6 a) Plot of the standard errors of the regression model given by equation 7.3 for Calanus helgolandicus. b) A histogram of the residuals from the regression model across all locations, with different locations being denoted by different coloured bars. 220
7.7 Plots of the predicted change in logged Calanus helgolandicus abundance over 10 years under the model. a) Estimated abundance of Calanus helgolandicus under a 1 degree increase in NHT over 10 years. b) Abundance of Calanus helgolandicus estimated from the regression model in 2008 minus the estimated abundance under a 1 degree increase in NHT over 10 years. 220

7.8 Plots of the true log-abundances of Echinoderm larvae and the log-abundances estimated from the modelled climate signals. a) Abundance of Echinoderm larvae in 1958. b) Abundance of Echinoderm larvae estimated from the regression model and using the modelled climate signals in 1958. c) Abundance of Echinoderm larvae in 2008. d) Abundance of Echinoderm larvae estimated from the regression model and using the modelled climate signals in 2008. . . . 222
7.9 a) Plot of the standard errors of the regression model given by equation 7.3 for Echinoderm larvae. b) A histogram of the residuals from the regression model across all locations, with different locations being denoted by different coloured bars. 222
7.10 Plots of the predicted change in logged Echinoderm larvae abundance over 10 years under the model. a) Estimated abundance of Echinoderm larvae under a 1 degree increase in NHT over 10 years. b) Estimated abundance of *Echinoderm larvae* from the regression model in 2008 minus abundance of *Echinoderm larvae* under a 1 degree increase in NHT over 10 years. 223
7.11 Plots of the true first principal component on the zooplankton species in 2008 and the modelled values. 227
7.12 Plots of the standard errors of the regression model for the first principal component of the zooplankton communities. 228
7.13 Plots in space showing the difference between the first principal component modelled under varying the covariates and the real first component. 229
7.14 Plots of the true first principal component on the phytoplankton species in 2008 and the modelled values. 230
7.15 Plots of the standard errors of the regression model for the first principal component of the diatom communities. 231
7.16 Plots in space showing the difference between the first principal component modelled under varying the covariates and the real first component. 232
7.17 Plots of the clusters on the first loading vector for the zooplankton data with the spatial average for the whole North East Atlantic removed. 235
7.18 Plots of the principal components for each region for the zooplankton with the spatial average for the whole North East Atlantic removed across the each of the clusters. The first component is shown in dark blue, the second in green, the third in red and the fourth in light blue. 236
7.19 Plots of the clusters on the first loading vector for the Phytoplankton data with the spatial average for the whole North East Atlantic removed. 239

7.20 Plots of the principal components for the Phytoplankton with the spatial average for the whole North East Atlantic removed across the each of the clusters. The first component is shown in dark blue, the second in green, the third in red, the fourth in light blue, the fifth in purple and the sixth in yellow. 240

Notation

Notation	Use
n	The number of measurements eg. time points or spatial points.
p	The number of variables eg. species.
k	The number of variables with a non-zero loading.
π	The proportion of variables that are true non-zeros, a sparsity parameter, $\frac{k}{p}$
$\mathbf{Y}$	The data matrix (eg. of species abundances), an $n \times p$ matrix.
l_j	jth position vector, where $l = (\text{longitude}_j, \text{latitude}_j)^T$
t	Time point.
Y_j	The jth column of $\mathbf{Y}$, or the jth variable (species).
$\tilde{Y}_j$	Fourier transform of Y_j.
$\hat{Y}_j$	Observed abundances.
$\mathbf{A}$	Matrix of principal component loadings.
$\mathbf{Z}$	Matrix of principal components.
ϵ	Matrix of random noise.
Σ	The data covariance matrix, given by $\mathbf{Y}^T\mathbf{Y}$.
$\mathbf{S}$	A diagonal matrix of the variances associated with each principal component.
a_i	The ith column of $\mathbf{A}$, ie. the ith loading vector.
z_i	The ith principal component.
L_p or $\|\|.\|\|_p$	The p-norm of a vector, given by $\|\|\mathbf{u}\|\|_p = (\sum_{i=1}^{n} \|u_i\|^p)^{1/p}$
λ	A sparsity parameter using in conjunction with a norm based penalisation.
Card$(.)$	The cardinality (number of non-zeros) in a vector.
ρ	A sparsity parameter used in conjunction with a cardinality penalisation.
$\delta(.)$	The delta function.
μ	A mean signal.
σ_k^2	Variance of the noise in our observed data.
σ_γ^2	Variance of the true noiseless signal.
$K(.)$	A kernel function, symmetric about zero.
h	The bandwidth, a smoothing parameter.
$\mathbf{H}$	A multidimensional bandwidth matrix.
$K_h(.)$	A kernel function weighted by a smoothing parameter, $K_h(u) = h^{-1}K(u/h)$
K	The number of partitions in a dataset.

Table 1: Notation used throughout this report.

Chapter 1

Introduction

1.1 Overview of the Aims of this Study and its Importance to Marine Ecology

Over recent decades the apparent increase in global temperatures and its impact on the environment has become a growing concern for many different sectors and policy makers [136, 19]. Changes in the climate have begun to have an impact on many different habitats and the ecosystems that rely on them [127, 166]. Interpreting ecological data is an important part of understanding the interplay between changes in physical variables and the impact this has on the joint behaviour of different species. The oceans in particular are known to play an important role in the global climate [130, 10]. In order to better understand variability in the marine ecosystem related to climate this study will explore the relationship between climate drivers and marine pelagic communities. The first chapter outlines the background of the ecological problem and the dataset in question in order to provide a framework for understanding the multidisciplinary nature of this research. Since this work covers both ecology and statistical methods it is necessary to provide this overview before describing the analysis. In the second chapter the statistical techniques are described in more detail and the interpretation of the output in terms of the biological problem is described. Chapters three through to seven contain original work

and the results of the analysis.

In marine ecology organisms living on the sea floor are said to be benthic, whilst those occupying regions of the ocean not including the sea floor are pelagic [163, 103]. These two systems are often interrelated [88], as changes in pelagic communities can propagate down to other organisms. This research focuses on the plankton, which plays an important role in the marine ecosystem [68, 15]. Plankton are small, ranging from microscopic to a few millimetres in size, organisms that form a community near the sea surface [68] and are subject to seasonal cycles [79]. Plankton forms a rich and diverse ecosystem, comprising of a varied mixture of single celled and multicellular organisms[68]. Species of plankton are generally grouped in to phytoplankton and zooplankton. The phytoplankton consist of a multitude of plant-like organisms, from diatoms, which have a rigid cell wall, to the dinoflagellates, which are characterised by their tendril-like flagella [68]. In the open sea there is relatively little free-floating multicellular plant life, so the phytoplankton exist as the primary producers in the marine environment [68] and as such are very important. Phytoplankton are also responsible for large amounts of the earth's oxygen production during photosynthesis [68]. As a result they perform a crucial role in ecology. Zooplankton comprise of anything from single celled organisms, small crustaceans, small jellyfish and fish and shellfish larvae [68]. Many marine organisms, such as fish and shellfish, will spend part of their lifecycle in the plankton [68]. Organisms that only spend part of their lifecycle as plankton are known as transient plankton during this part of their lifecycle [68] and during this phase are dependent on other plankton as a food source. For those species of fish that spend part of their lifecycle amongst plankton changes in the behaviour of the species they predate upon whilst in their planktonic stage will have a knock on effect for the abundance and distribution of adult fish [21, 27, 28]. This means that there is a direct interaction between the abundance of plankton and the abundance of larger marine organisms. The spatial distribution of the plankton can also be tied

to the spatial distribution of other organisms [29] and so understanding the distribution of plankton is important to understanding the marine ecosystem as a whole. Furthermore the plankton are known to be sensitive to changes in climate [3] and can therefore amplify changes observed in physical variables [96].

Various different climate indices have an influence on the abundance of plankton [73, 65]. Some of these effects are believed to be anthropogenic, i.e. driven by human behaviours, and others thought to be natural oscillations in pressure or currents [43, 150]. Understanding the role of both natural and anthropogenic effects is important to marine and environmental policy makers, as it will allow them to better plan for the future [53]. Natural climate oscillations are typically measured as pressure differences between spatial locations and switch between high and low phases at regular intervals [121, 139]. Previous studies [43] have shown that the detrended sea surface temperature can be decomposed in to various climate signals, such as the Atlantic Multidecadal Oscillation (AMO), the East Atlantic Pattern (EAP) and the North Atlantic Oscillation (NAO), the influence of which can vary in space. The spatial pattern of these climate indices influence on sea surface temperature might also relate to the spatial distribution of plankton [98], with the region a particular species is most well adapted to being referred to as its ecological niche [23]. Understanding spatial heterogeneity is an important part of ecology across all types of system and spatial scales [63]. For example not all species of plankton respond to climate in the same way [65, 73] across space. In addition an understanding of the sensitivity of the plankton to climate effects and the ways in which this varies might help influence the decisions of policy makers looking to safeguard the marine ecosystem against potentially detrimental effects from climate change [53].

1.2 The Continuous Plankton Recorder Survey

This study will focus on data drawn from the Continuous Plankton Recorder (CPR) survey. The CPR survey is an ecological dataset of particular interest because of the abundance of data and long term records available [14, 11, 160]. The CPR survey contains monthly abundance data for 450 different species of plankton [160], of which approximately 110 are studied in this report, across the North Atlantic [11]. The CPR survey is unique in the temporal range of data available [138], which in some cases dates from the 1930's, as for many datasets only a short period of data is available or there are missing data entries due to financial constraints, the spatial coverage and the number of taxa investigated [138]. This is highly problematic because long term records are absolutely essential for understanding the effects of climate change [16, 54, 4]. Such extensive ecological data is useful for understanding species structure in both time and space but the multivariate nature of the data presents a number of challenges to modelling. Previous studies of the CPR data have studied the influence of climate variables [25, 31], as well as its structure in time and space, but often for only one or a handful of species. Most studies tend to concentrate on particular *indicator species*, which are said to represent something of the structure of the entire dataset, but few have looked at the joint behaviour of different species groups [14, 25]. Those that do look at joint behaviour, for example Beaugrand, Ibanez and Reid [24] who study 11 species of Copepod, concentrate on a small subset of species [25], as the computations required to analyse the data across multiple species can be very time consuming. The limitations of focusing on only a few indicator species is that it is impossible to identify the joint behaviour of functional groups or differential species responses to climate for a large number of species. This study therefore provides entirely new insights in to the complex structure of the CPR dataset. The community analysis presented in this report is novel in that similar studies have not been carried out before. We propose statistical

techniques that enable one to analyse the data across space, time and species simultaneously, which consequently provides new insights in to the behaviour of these marine pelagic communities.

The CPR survey is a useful tool for a wide variety of reasons. Recently the impact of climate change on phytoplankton has been widely reported in the media, with some studies showing a decline in the global biomass [41], although other studies show that phytoplankton biomass is increasing [54, 112]. The CPR survey records both total biomass of phytoplankton, represented by a colour index, and counts of individuals [54, 13, 99]. Abundances are useful for studying interactions between species [99] but do not account for the difference in size of individuals of different species [99]. Biomass is a measure of the total volume, meaning larger species contribute more per individual, which is useful for investigating overall effects [99]. Changes in the phenology (seasonal blooming patterns) can be detrimental to the development of fish larvae [27]. Whilst the blooming periods of certain plankton are dependent on light, the spawning patterns of most fish depend on temperature [96]. Consequently the blooms of young fish no longer correspond to the times when there will be an abundance of food in the plankton, since climate warming is known to bring these seasonal cycles forward, and this will have consequences for adult fish stocks.

The Continuous Plankton Recorder device was originally envisaged by Sir Alister Hardy as a tool to aid the fishing industry [69, 70, 71]. At the time fishers used discs of gauze to collect plankton samples, which were known to be good indicators of where fish would be found [68]. Alister Hardy came up with the unique idea for a device which would collect continuous records of plankton abundances. The CPR device was originally designed to collect samples on a continuous band of silk, which would then be stored with a covering sheet in a well of formaldehyde [69, 70]. Samples on each section of the band of silk will correspond to a specific area of the sea, which is determined by the speed of the ship. Save a few improve-

ments to the efficiency of the device the design remains largely unchanged today [138]. Samples are still routinely returned to the laboratory in the Sir Alister Hardy Foundation for Ocean Science (SAHFOS), where they are identified and counted by skilled analysts. The resulting database is a vast collection of monthly samples from the North Atlantic [11, 14]. Of course the original purpose of the CPR survey has now become redundant with the efficiency of the fishing industry increasing to such an extent that there are now grave concerns about over-fishing [159]. What Alister Hardy could not have foreseen was that the CPR survey would come to have a new importance when researchers and policy makers began to observe the potential impact of climate change on the oceanic environment [96].

One of the challenges in modelling the CPR data that has been previously discussed is the irregularity of the sampling. The CPR survey is collected by so-called 'ships of opportunity', which refers to the voluntary towing by merchant vessels [160, 25]. The fact that shipping routes have occasionally changed in time gives slightly better spatial coverage than might otherwise be the case [160] but the routes have not been designed for optimal interpolation. Figure 1.1 shows the shipping routes across the North Sea and North East Atlantic, which shows the irregularity of the sampling. The data is also thought to be subject to a certain amount of noise, due to human error, the time of day the data was collected (plankton are known to rise nearer the surface during daylight) and other environmental factors such as wind and currents [160]. One approach to dealing with these confounding factors is to average over large spatial regions but this sacrifices the spatial resolution that is necessary for finding regional behaviours. The approach taken in this study is to use smoothing methods, which can also be used to reduce the effect of noise, and will be discussed in more detail later. Irregular sampling is often a problem in ecological datasets [51, 120, 34]. Some surveys use 'preferential' sampling, where regions known to be particularly rich in certain taxa are sampled [113, 39]. Other studies use 'non-preferential' sampling, where the sampling is either done across

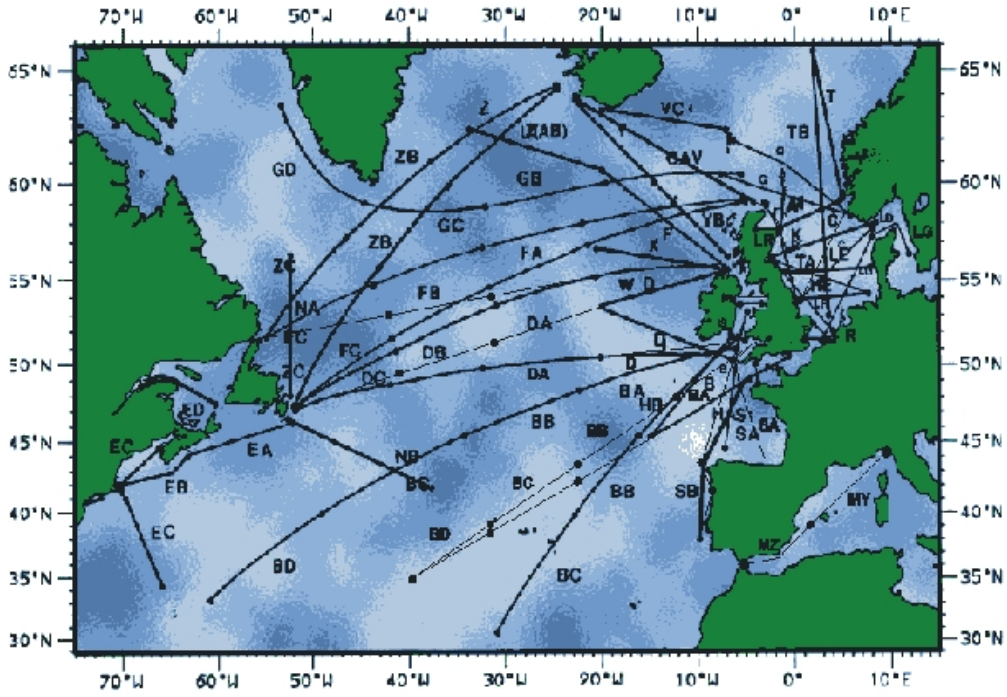


Figure 1.1: Map showing the shipping routes towing the CPR device across the North Atlantic and the North Sea.

a regular grid, randomly or is limited by physical constraints on where it is possible to sample. This means that missing data must often be estimated from these samples or otherwise accounted for [66, 143], for example by using interpolation methods [167], regression based methods [95] or Bayesian methods based around priors on the probability of observing a particular species [143]. In the case of the CPR dataset the sampling methods are not preferential but rather constrained by shipping routes.

Analysis is initially carried out on the pre-processed WinCPR dataset and then we look at raw abundances over a larger spatial region over which the analysis is repeated. The WinCPR dataset is a gridded database containing abundance of approximately one hundred species recorded in the CPR survey for the North Sea since 1958 [137]. This data has been interpolated using inverse distance methods, which are methods for estimating the abundance of plankton at unsampled locations by taking weighted sums of nearby locations. The WinCPR data can be used to verify different models fitted to the data, in order to test the validity of our methodology. In order to study a larger spatial region than available in the WinCPR, extending to the entire North East Atlantic the raw abundances are used. This data must also be interpolated in order to model it on a regular spaced grid.

The WinCPR data has been produced by first transforming the raw abundance data by taking the logarithm of the data plus a constant and then interpolating the transformed data in order to estimate values at unsampled locations [160]. The constant is added in order to avoid taking logarithms of zero, which can not be calculated. Count data is typically modelled using a Poisson distribution. A Poisson distribution is positively skewed, in that most of the observations will with high probability lie around the median but there will be a small number of very high observations leading to large observations at the extremes. This distribution is typical of plankton abundance data. In this type of data it is expected that the variance will be dependent on the expected values of the observations. The logarithmic trans-

form removes some of this dependence and so is often used so that data will satisfy distributional assumptions, in particular it is expected that the transformed data will be closer to a Gaussian distribution. The logarithm also deals with the positivity of the data, as the abundances can only take positive values but the log-abundances can take both positive and negative values. The abundances might also be thought of as the product of a deterministic process and a stochastic variability, where a deterministic process is non-random and a stochastic process is governed by an underlying probability distribution and can be used to model the random variability. A logarithmic transform turns multiplicative relationships in to additive relationships, which allows it to be used as a variance stabilising transformation. When the raw data is analysed later the same transformation will be used in order for the results to be comparable to those found using the WinCPR dataset.

For plankton count data the abundance of species p at time t and location l can be said to be distributed $\text{Po}(\mu^{(p)}(t,l))$, where $\mu^{(p)}(l,t)$ is the mean intensity at time t and location l. The expected value and variance are equal in a Poisson distribution, so the variance is also $\mu^{(p)}(l,t)$. Various statistical methods rely on the assumption that data are approximately Gaussian and have constant variance, for example linear regression which shall be used later relies on the assumption that the dependent variable is normally distributed [168], or at the least that any error terms will be normally distributed. This means that the functional dependence between the expected value and the variance might be undesirable for the statistical analysis. Once the data has been transformed it is then interpolated using geostatistical methods [51] to produce the WinCPR dataset. The data is gridded on a 1 degree by 0.5 degree grid and recorded for each year and month. In the later analysis where the dataset over the whole North East Atlantic is analysed a slightly different interpolation method will be used. Since the monthly data is strongly dominated by seasonality yearly averages are often taken in studies wishing to focus upon long term variability. An alternative approach might be to model and analyse each month separately.

The raw CPR data is recorded by longitude, latitude, year and month as counts of each species. Counts are estimated from cross sections of samples [160]. Also recorded is the phytoplankton colour index [13], a measure of the Chlorophyll in the sample, which serves as an estimate of phytoplankton biomass [54]. This study shall investigate over a hundred species across the whole of the North East Atlantic in order to determine how different ecoregions respond to the effects of climate change. In order to do this the data must first be interpolated and then gridded. The raw data is irregularly sampled but the results in space from the analysis are more interpretable when presented across a regular grid in space. Various interpolation methods exist for the purposes of estimating missing values. Before any smoothing is carried out, however, a stabilising transform is performed. As with the WinCPR data a logarithm is used, after first shifting the data by one to avoid any zero entries, as was done with the WinCPR data. The data is interpolated in space, treating each time point separately. For regions where there is insufficient spatial data by necessity some smoothing in time is also carried out. Some smoothing is also carried out during the interpolation in order to reduce the noise. There is a trade-off in the smoothing between smoothing sufficiently in order to remove the noise but not over-smoothing in order to remove small scale effects, and this will be discussed further later.

1.3 Modelling Spatio-Temporal Multivariate Data Sets

The first sets of results using the raw data are used to analyse responses of both indicator species and species communities to climate. However the multi-dimensional nature of the data provides many statistical challenges, even once it has been transformed to a regular spatial grid. Long term trends across multiple species and locations are difficult to isolate by eye and so variable reduction methods are need in order to make it simpler to identify the ecological patterns [25]. In theory one

could model all species across all time in space but any resulting model would be complex with at least nominally a large number of parameters. The method used to find the dominant trends in both sea surface temperature and plankton abundance is principal component analysis (PCA).

PCA is a method for finding maximal directions of variation through a dataset by taking linear combinations of variables [85] and is often used to analyse ecological data [95]. Although it is often thought of as being 'model-free', it could be seen as modelling each variable as a linear response to common signals. Spatial PCA, which takes weights as functions of space and signals as functions of time, has been used in other studies to separate individual indicator species of plankton in to their spatial and temporal representations [25, 31, 14]. The advantage of this is that it reduces the number of dimensions and produces summaries of the main modes of variability [86], which are easier to compare with climate signals [31]. Similar analysis can be carried out on the sea surface temperature data. The study by Cannaby *et al* [43], for example, breaks the sea surface temperature down in to its dominant spatial and temporal components using what they term as empirical orthogonal function (EOF) analysis, which is functionally the same as spatial principal component analysis (PCA)[86]. In this study they find the dominant mode of variability to be given by the AMO, followed by the EAP and the NAO, accounting in total for 48% of the variation across the first three principal components in the sea surface temperature over their region of interest. In this thesis a new approach to analysing the CPR data using existing statistical techniques is proposed for visualising the behaviour across species as well as space and time. Species principal component analysis is used to find species groups and their joint behaviour at different locations. It is shown that accounting for time lags [149, 102] between species is important, in particular when comparing these group responses with climate trends. The resulting joint behaviours can be compared with climate indices using linear regression analysis, which assumes that each component is proportional to the cli-

mate variable plus a constant and can be used to find the strength and nature of this linear relationship [168]. It has been hypothesised that climate warming is driving changes in species distribution for species of zooplankton, with cold water species declining and temperate species increasing in abundance [12, 17, 22, 74, 127] and so this analysis can be used to investigate this hypothesis. Meanwhile other species groups, such as the Diatoms, are thought to be less directly driven by climate warming trends as much as they are driven by the changes in mixed layer depth caused by fluctuations in currents [56].

One potential difficulty is that because principal component analysis assigns a non-zero weight to every variable this can impede the interpretation of the results [172], which is especially pertinent in the case of the CPR data where there are a large number of species. When investigating the functional behaviour of species assemblages small non-zero weights may be difficult to interpret, since they are typically assigned to rarer species. Since the data is counted by eye there may be noise resulting from human error, which can affect the resulting weight vectors. Furthermore it is often difficult to identify those species that contribute most to the joint behaviour of the ecosystem by eye. Simple thresholding, that is setting all the observations that fall below a specific value to zero, can be used on the species abundances, to remove those with little variability, but choosing the correct threshold can be highly non-intuitive [172] because it is difficult to determine which weights are drawn from the white noise by eye. Furthermore the number of true non-zero species can vary across space and functional groups, meaning the thresholds are not necessarily constant. By constructing principal components from those variables with the greatest variability one can identify functional groups of species that behave together, which may be linked to biological groupings [47, 68]. One method for dealing with the case where the number of variables is large in comparison to the number of observations is sparse principal component analysis [172, 50]. This method is used for datasets where a number of variables might take small non-zero

values but not contribute in a meaningful way to the total variation. An example of this is gene expression data [44], where only a few genes that are expressed are related to the observed response but other genes will show some levels of expression due to noise. It is assumed that at any given location only a few species contribute to each common signal and that the rest is noise. Since species are optimised to survive in certain climate and environmental conditions [75] it is unlikely that all groups of species will be present across all locations, hence modelling some of them as having zero weight is appropriate. Sparsity in statistics is where only a few of the variables in a dataset are non-zero [152]. Of course in most experimental data there is some noise and so true zero observations are rarely observed [152]. In most cases sparsity is assumed to be represented by a few non-zero observations and the rest as small error terms [152]. Modelling the data as sparse introduces a trade-off between explained variance and the ease of interpretation [49]. The appropriateness of a sparse model can be assessed by investigating this trade-off in a similar way to traditional model selection methods. It is generally the case, however, that if the model of sparsity is appropriate to the data the loss of explained variance will be relatively small [172, 49], whilst the improvement in understanding the structure will be significant. Some methods, such as the method proposed by Zou, Hastie and Tibshirani [172], for incorporating sparsity in to the principal component analysis also result in the loss of orthogonality of the principal components. Again this is thought to be a relatively small decrease and so is more than compensated for by the improvement in interpretation [172]. Sparse principal component analysis methods incorporate a penalty in to the normal PCA algorithm, which forces small loadings to be zero.

One remaining question is how to determine how many species should contribute to each common trend. In most formulations of sparse PCA there is typically a parameter which controls the number of variables that have non-zero weights [172, 50]. If it is assumed that some of the variables mostly consist of noise then one

might seek to estimate the number of variables that have true non-zero weights and those for which the weight is a function of noise only in order to gain a better representation of the dataset. In this case the signal is drawn from one distribution whilst the noise is drawn from another, naturally leading to by placing a mixture model on the loading vector [154, 82, 83]. A mixture model is used in situations where data is thought to have been drawn from two or more probability densities and is dependent on a parameter which determines what proportion of the data is drawn from each distribution [154]. Johnstone and Silverman [82, 83] use a similar approach for estimating sparsity on data where there are a few true non-zero values and the rest are decaying values. Expectation maximisation, which is an iterative procedure for finding the maximum likelihood, or another maximum likelihood method can then be used to find the proportion of the variables that should belong to the latter distribution [82, 83]. The value of this parameter is proportional to the group size for each trend across space. The number of different groups for each location is determined by thresholding on the culmative explained variance for the principal components [86]. Together this will give a measure of the diversity of each location. Changes in the ecology and the vulnerability of the ecosystem can be assessed in terms of measures of diversity. In general, diversity is thought to be a measure of the health of an ecosystem, with more diverse ecosystems being considered to be more successful [20, 63]. In the case of this model there are two measures of diversity: the number of functional groups and the proportion of the species belonging to each group, denoted by the sparsity parameter. An ecosystem might also be considered in terms of its vulnerability, how likely a change in the environment is to give rise to the collapse of the ecosystem. A system with a large number of functional groups might be considered less vulnerable than one with a single group with many members [125, 20]. Paine [125] discusses the concept of keystone species, which are species that when removed from the ecosystem can cause the collapse of multiple other species, which is likely to be the case if a species interacts with

many other species (i.e. is part of a large functional group). This is because in the latter case the collapse of a single species is more likely to have a knock-on effect on other species in the ecosystem. If it is assumed that the common signals found by PCA are joint responses to climate, then either the species are interacting with one another are they are all responding to the same driver. In either case a large group size indicates that a change in one species will be accompanied by a change in many other species, either because it is a result of a climate shift or because of the relationship between different species.

1.4 The Biogeography of the North East Atlantic and 'Regime Shifts' in Ecoregions

In this study the application of this model to exploring non-stationarity in the ecoregions of the North East Atlantic is explored. There are a number of ways that the ecoregions of the North East Atlantic might be defined. Regions can be defined on the species assemblages [1, 7, 103] or on climate variables and current patterns. McGinty *et al* [108] for example find regions likely to have similar plankton dynamics by clustering on chlorophyll patterns. The marine ecosystems of the world (MEOW) [145] instead define provinces and ecoregions on coastal and shelf areas using biogeographic classifications. Provinces are classified as large areas defined on more general species groupings, showing similar behaviour over evolutionary time. They might also be described by hydrographic features, such as currents and upwellings, and geochemical influences, which include the broadest scale nutrient dynamics such as salinity. All the coastal regions of the world are classified in to 62 provinces under this system [145]. Areas around the British Isles can be divided in to the cold-temperate boreal province, containing the North Sea; the warm temperate lusitanian province, including the Bay of Biscay, and the overlap between those two provinces, including the seas around Ireland [161]. North of Iceland one enters

in to the Arctic Province. The areas around the British Isles are very much a mixing zone, which makes determining the ecoregions a complex task [161]. Ecoregions in the MEOW [145] are the smallest units and are defined as having a relatively homogeneous species composition determined by a small number of ecosystems. Whilst species composition on large scales is determined by climate variables, the main factors at an ecoregion scale are isolation, upwelling, nutrient inputs, freshwater influx, temperature regimes, ice regimes, exposure, sediments, currents and bathymetric or coastal complexity [145, 103].

There are various ways of defining the ecoregions of the North Atlantic from the CPR data. Rangel et al [129] comment on not only the statistical challenges in modelling spatial structure but also its importance in understanding underlying biological processes. The ecoregions can be regressed against spatial variables, which drive the biological behaviour. In the North East Atlantic, for example, sea surface temperature and salinity are said to be major drivers in which species are found in different areas [9], with the North Sea behaving very differently to the rest of the North East Atlantic [108]. Not all regions are equally sensitive to climate change and in some regions only a small change in temperature is required to cause a significant shift in the ecosystem [22, 75], which can also be considered in terms of ecological niches [75]. Beaugrand et al [22] explore how to address the problem of spatial variability by focusing on chlorophyll concentration and *Calanus Copepods*, as well as a number of physical variables such as sea surface temperature.

One of the areas of exploration in this thesis is how to understand species interactions in particular in relation to the 'regime shift' [30]. Part of this has been previously described in both the CPR data and in other ecological datasets. Some previous work has been done in attempting to quantify the 'regime shift' * but the community analysis in this study adds new insights. It is believed that rising av-

*A 'regime shift' in this context is a stepwise change in the behaviour of species assemblages and can be considered either spatially (i.e. the northwards movement of species [12, 30]) or temporally as a change in the average abundance [17].

erage temperatures have caused taxa to be found increasingly far north over the past decades [22, 75, 42]. Long term regime shifts around the 1980s are well documented in the CPR dataset [30, 25, 31, 33, 92] but these regime shifts are difficult to quantify [146] because it is unclear whether they take the form of step-wise changes or smooth trends over time. It is also possible that different species will respond in different ways to a change in regime and that this shift will not be constant over space [108]. The complexity of the data makes any potential 'regime shift' difficult to identify. Beaugrand [30] approaches the problem of quantifying the regime shift by taking a sliding window across a number of species and comparing the Euclidean distance between the data in the first half of the window with the second to determine whether there has been a shift or at least identify where the distance is greatest. An approach to quantifying 'regime shifts' is to use change point analysis. This assumes that the data are stochastic but that the distribution will different before and after a fixed point in time, in particular it may have a different mean or variance. If the 'regime shift' is instead viewed to be a trend, i.e. a more gradual change over time, a deterministic model might be more appropriate. Change point analysis in statistics has been explored by various authors and there are many methods for identifying changes in mean or variance. With growing concern over the effects of climate change ways of modelling change points have been explored in many areas of ecology. Instead of using traditional change point analysis a more applied mathematical approach might consider looking at instability and bifurcations [107], which are sudden transitions in the behaviour of a system resulting from only a small change in the values of the parameters, in the population model. Ecological models may be non-linear, where a non-linear model is used to account for small changes in the independent variables potentially having a large impact on the outcome. For example once recent study of extinction in deteriorating environments [52] explores using bifurcation points as a predictor of impending extinction of a particular species. The limitation of all of these approaches is they tend to lend

themselves to investigating the time course of one species at a time.

Regime shifts can be investigated over different temporal scales. Typically the CPR data is dominated by monthly seasonal cycles, which can make long term trends difficult to identify. Most studies of the CPR dataset deal with this issue by taking yearly averages. There is, however, interest in both looking at yearly data *and* looking at monthly data. The yearly data can inform about long term changes in the average biomass or in the general trend compared to climate variables. The monthly data contains information about how the seasonal cycles have changed over time. It is believed that seasonal cycles in the plankton have changed as a response to climate, leading to misalignment between species [53, 96, 31], i.e. species that previously had concurrent blooms now have their blooms at different times of the year.

The methodology used in this thesis allows ecoregions of the North Atlantic to be found either based upon species or upon temporal trends. The ecoregions of the North Atlantic based on species is governed by physical features such as climate, salinity and bathymetry. This is because different species are adapted to survive better in different conditions [75], with some responding to temperatures [12, 5] and others responding to currents [56] or being reliant on certain nutrients. Marked changes in the biogeographical regions of the North Atlantic can be seen before and after 1985, adding weight to the conclusion in this report that species distribution in the North Atlantic is changing [108, 132, 16, 18]. K-means clustering is used to analyse spatial structure by clustering on both the weight vector, to find spatial regions defined on the species, and the temporal trend, to find spatial regions defined by common functional behaviour. K-means clustering is a method which finds groupings of variables [168]. For a pre-defined number, K, of clusters the algorithm first randomly selects K variables as centres [81]. The rest of the variables are then assigned to a cluster by finding the minimal Euclidean distance from the centres. The centres are then recalculated by taking the average of each cluster

and the process is repeated until convergence is reached. Some care must be taken as the resulting clustering can be dependent on the initial choice of centres and so multiple runs of the algorithm are necessary [81]. In the case of clustering on the results of the principal component analysis for the CPR data, clusters on the weight vectors will determine spatial structure in the species groupings, whilst clusters on the temporal signals will describe spatial structure in the joint behaviour of the species, which is believed to be related to climate trends. If a species has a zero weight for all principal components then it is regarded as not being important at that location, whilst if only one species has a large weight then a location is thought to be dominated by a single taxa and therefore will be vulnerable to changes that affect that taxa. For the purposes of studying ecosystem shifts clustering on the loading vectors is likely to be most informative. PCA has fixed weight vectors over the entire time course but changes in the ecoregions are found by carrying out the analysis before and after 1985, which is thought to be the approximate time the 'regime shift' may have occurred [30]. The ecoregions are found on both halves of the data and then compared to show that there is non-stationarity in the spatial distribution of species, which can be explained by ongoing ecosystem shifts.

1.5 Modelling Sensitivity to Climate Across Species and Space

For policy makers the issue of vulnerability is of great importance [53]. In this study vulnerability to climate is viewed in terms of a large change in the abundance proportional to the previous abundance of a species or functional group of species in response to a comparatively small change in a given climate index. This is of interest to environmental policy makers because it can determine the tendency of a species or group of species to either disappear from a region or to increase dramatically under different climate scenarios. The issue of vulnerability to climate variation can be explored by investigating sensitivity of *different regions* or *differ-*

ent species to climate effects. The vulnerability of different species may well vary across space. The spatial pattern of the vulnerability of a single species can also be compared with the vulnerability of functional groups across space. Both the vulnerability of single species and the vulnerability of the joint behaviour across space are informative. The first might be of interest to those looking at a particular part of the ecosystem, for example species upon which fish predate, and the latter is of interest to those looking at wider scale changes in the habitat. As well as looking at diversity a way of determining vulnerability is to model changes in abundance over hypothetical changes in the climate drivers. The vulnerability of a species or region to climate change is then viewed as its sensitivity to particular climate variables, i.e. how drastically changes in the physical variable influence a species or the joint behaviour of different species groups at a particular location. The results of such analysis is therefore useful not only to ecologists but also to environmental policy makers [53] because it allows one to investigate changes in community structure across the region, which will in turn have consequences for the entire ecosystem. A particular region can be said to be sensitive to a particular climate variable if the model shows the magnitude of the weight relating to the appropriate covariate is large at that location. In terms of policy making this will allow sensitivity 'hotspots' to be identified, since vulnerability of the plankton to climate change may well induce instability in the rest of the marine ecosystem which are dependent on the plankton. Drastic changes in the local behaviour of the plankton can lead to instability in the ecosystem as a whole, which can be damaging to many different organisms [21, 27]. A particular species can be said to be vulnerable to a particular climate index at a specific location if the correlation between a component and the climate is strong and the weight on that species is large at that location, allowing the identification of which species will benefit under changes in climate and for which it will have a detrimental effect.

Sensitivity to climate effects might also be investigated at different scales by

taking a 'macro-scale downscaling' approach. In this study 'macro-scale downscaling' refers to finding dominant drivers of average responses over large regions and then finding drivers of the residuals once the average has been removed at smaller scales. The first part finds large scale influences, whilst the second part investigates more localised effects. It is expected that long term climate trends will dominate over larger scales because they impact the environment over a large region, whilst at smaller scales more regional effects will be important. The abundances of all species can be averaged over a large spatial region and the analysis performed across the whole region. The larger region can then be divided in to smaller regions and the average abundances across the large region removed before the analysis is carried out again on the smaller spatial region. This process can be repeated multiple times, dependent on the spatial resolution of the data. On large scales it is expected that the trends will be most strongly influenced by climate trends, whilst at a smaller scale more local phenomena, such as currents, nutrients or short term oscillations, will be important. Bringing together all the different parts of this research allows one to gain a better understanding of the complex systems and changes that impact the marine environment.

1.6 Climate Indices and their Influence on Marine Ecology

When studying the effects of climate on the marine system what is difficult to determine is how much of the changes observed can be attributed to natural climate variability, and how much is due to anthropogenic influences. Since the increase in carbon dioxide is a result of increased industrialisation, emissions and the use of aerosols [38], the average warming trend is believed to be anthropogenic [142]. The issue of climate change can also be controversial [136]. Therefore the question of whether apparent changes in climate are due to anthropogenic effects or natural fluctuations is of vital importance. Natural phenomena that might influence climate

include a number of oscillations in pressure and currents. What are believed to be natural oscillations can often be seen in the temperature signal, where a temporary increase might be seen during a high period and a decrease during the low period. Since these different indices represent different fluctuations in the physical conditions, eg. changes in currents, pressure centres or temperature, then it is of interest to compare them separately with the plankton trends. Each of the climate indices discussed here varies across time and, with the exception of Northern Hemisphere Temperature, since that is averaged over space, also have spatial patterns in their influence across space.

1.6.1 The Atlantic Multidecadal Oscillation

The Atlantic Multidecadal Oscillation (AMO) has a period of approximately 60-80 years [90, 141, 62] and was first identified in linearly detrended sea surface temperature data [141], which means it is measured in degrees Celsius. This long term oscillation, which is currently in its warm phase, has also been identified in ice core records and tree ring data [90, 46]. The presence of a 60-100 year oscillation, likened to the AMO, in these records suggest it is probably a natural oscillation and thus not related to climate warming, although evidence for the presence of the AMO during the pre-industrial era is inconclusive [90] and the period is variable over time [46]. What causes the AMO is currently unknown [90]. It is not thought to be driven by solar cycles and there is some speculation that it is the result of currents, with a decrease in the overturning circulation in the North Atlantic being associated with a cool phase in the AMO [141]. The AMO has the strongest influence on sea surface temperature in the subpolar gyre [†], a region in the central North Atlantic where there is a large system of rotating currents [141]. Convection forces drive the movement of the warmer waters outside the gyre, influencing the motion of

[†]The subpolar gyre is a meeting of currents in the central North Atlantic Ocean [118]. Sea surface height is variable in this region and local climate is governed by the overturning circulation.

the currents [118]. Further speculation suggests that the subpolar gyre might move across longitude during the different phases of the AMO [118]. The length of the time series, however, makes this difficult to verify. The AMO is associated with changes in rainfall and temperature over the northern hemisphere and can been seen in the Northern hemisphere temperature (NHT) time series [55]. The AMO is also thought to influence the occurrence of droughts in North America and the European summer climate. It is believed it may also have an impact on the occurrence of hurricanes in the Atlantic [89] and on rainfall and river flows in the continental United States [55]. Since the AMO influences climate events, it is important to remove its influence when studying the effects of the warming trend, as it might either obscure or exaggerate the observed trends [141]. Likewise in order to retrieve the AMO from sea surface temperature data the global warming trend much first be removed, which is typically done by fitting a linear model and removing the trend [90]. Cannaby *et al* [43] identify the AMO to be the most dominant trend in the detrended sea surface temperature over a region extending from approximately 50° W to 1° E and 20 to 70° N.

1.6.2 The Northern Hemisphere Temperature

The Northern Hemisphere temperature (NHT) is a measure of atmospheric temperature in degrees Celsius over the northern half of the globe leading to a single time series. On average the Northern Hemisphere has been warming over the past few decades, a change which has been attributed to an increase in atmospheric carbon dioxide [96]. This change can be seen in the time series. The signal is also subject to an oscillatory effect, which has a similar period to the AMO. As previously discussed the AMO and the general warming trend might obscure the effects of one another, making each separate effect difficult to isolate. Cannaby *et al* [43] remove the general warming trend from the sea surface temperature signal by modelling it as a function of atmospheric carbon dioxide. The trend might also be removed by

linear detrending, which involves fitting a straight line or a higher order polynomial depending on whether the effect of climate warming is believed to be linear and subtracting this from the data. The warming trend is not, however, homogeneous across the entire North Atlantic. Whilst the North Sea, particularly the southern region, seems to be warming faster, the subpolar gyre seems to actually be undergoing a cooling effect [67]. This might be a result of bathymetry or circulation, as the North Sea waters are significantly shallower than the rest of the North Atlantic [123]. The average temperature increase is also believed to affect the behaviour of currents, such as the Gulf Stream, the current that helps maintain the western Europe's relatively temperate climate [114]. Changes in these currents might lead to more complex changes in local climate in different regions of the North Atlantic and this heterogeneity means that the local climate variability can often be difficult to predict [144].

1.6.3 The North Atlantic Oscillation

The North Atlantic Oscillation (NAO) is a measurement of fluctuations in the difference of atmospheric pressure at sea level between the Icelandic low, a permanent low pressure centre, and the Azores high, a permanent high pressure centre located near the Azores in the Atlantic ocean [87, 164, 135]. The difference in pressure between the Icelandic low and the Azores high controls the strength and direction of the westerly winds in to Europe [87, 164]. The NAO also has a strong influence on the climate of the North Atlantic and is a dominant mode of atmospheric pressure variability [94, 77]. This oscillation is believed to have a period of about 8-10 years [76, 87]. This oscillation has been identified in both the sea surface temperature signal and in ice core records [76]. Spatially the pattern of the NAO is dipolar, in that it has two centres where it positively influences sea surface temperature and a centre around which it has a negative influence, having the strongest influence in the North Sea [76]. In the subpolar gyre this oscillation has an inverse effect, with the

temperature signal in this region slightly increasing during the negative phase of the NAO and slightly decreasing during the positive phase. The influence of the NAO in the North Sea may well also be linked to bathymetry [123]. The high phase of the NAO is associated with cool summers and mild and wet winters in Central Europe and slightly warmer winters in eastern North America, in contrast to cold winters in both Europe and eastern North America and Mediterranean storms during the low phase [76]. The extreme phase of the NAO may be influential in the in unusually dry conditions over southern Europe and wetter than average conditions in the north [76]. In the North Atlantic region the NAO is strongly influential in wind speeds and directions, temperature, moisture distribution and the frequency and intensity of storms. The NAO also influences the position of the Azores high, which can effect where storms occur. This oscillation might also effect the abundance of plankton, in particular Fromentin and Planque [59] claim a link between the NAO and the two copepod species *Calanus finmarchicus* and *Calanus helgolandicus*. Whilst a high phase is detrimental to *C. finmarchicus*, *C. helgolandicus* seems to benefit from it. Furthermore the NAO is thought to influence the length of the blooming season of phytoplankton and the behaviour of terrestrial species, including birds and amphibians [122]. During the positive phase there is an increase in numbers of Arcto-Norwegian cod, with temperature being influential at every stage of development from larval growth and mortality, food availability timings to adult growth and survival [122]. The influence of the NAO on wind anomalies also effects ocean circulation [77], meaning it has an impact on ocean currents. In Cannaby *et al*'s study [43] the NAO was found to be the third most dominant mode in sea surface temperature.

1.6.4 The East Atlantic Pattern

Although Cannaby *et al* [43] find it to be the second most important mode of variability in sea surface temperature and it is thought to be the second most domi-

nant mode of atmospheric variability after the NAO [119], the East Atlantic Pattern (EAP) is perhaps less well understood than the NAO. The EAP behaves similarly to the NAO and consists of a north-south dipole in space of anomaly centres spanning the North Atlantic from east to west, as well as showing a strong multidecadal variability [165]. The positive phase is characterized by above average temperatures in northern Europe and heavy precipitation over northern Europe and Scandinavia [165] and impacts sea level pressure [126]. Barnston and Livezey define it as having a centre near 55° N and 20 to 35° W with a strong northwest-southeast gradient over western Europe [8]. Owing to the lack of available atmospheric model reanalysis data in the central North Atlantic it is not possible to extend the EAP index back beyond 1948 based on this definition [43]. This lack of available data makes it more difficult to compare with sea surface temperature records.

1.6.5 Physical Factors

There are a number of physical features which might influence the spatial inhomogeneity of these climate indices and plankton communities. There is an ocean shelf, which extends around most of the British Isles [145]. There is also a west to east gradient in the salinity, with the North Sea having fresher waters than the open ocean [105]. Circulation, currents and mixed layer depth are all factors that influence a region's sensitivity to climate events. Mixed layer depth is a measure of the depth of the water column within which surface water mixes with waters from below and this can influence the plankton [106]. This study will explore the influence of these different drivers on the spatio-temporal variability of plankton in order to better understand which are the most important factors in driving abundance.

Chapter 2

Methods

In this chapter the statistical methods used to analyse the CPR data in this thesis are discussed in further detail. Some of these concepts were introduced in chapter one but in this chapter the technical details of the statistical methods will be discussed in more depth. Recall from chapter one that the CPR data is irregularly sampled in space as the sampling is dependent on shipping routes. This means that interpolation methods are required to transfer the data to a regularly spaced grid, which is useful for interpretation. Approaches to interpolating and smoothing the data in order to accommodate for the irregularity of the sampling and the variability due to the sampling process are expanded upon in this chapter. The main analysis in this thesis makes use of principal component analysis and sparse principal component analysis in order to summarise the complex structure of the data. Previous studies, such as those by Beaugrand et al [25], have also made use of principal component analysis as a technique for visualising the CPR dataset. In this chapter different ways of using principal component analysis with the CPR dataset are focused on. Finally different approaches to using the output of principal component analysis to answer questions of biological importance, such as how plankton respond to climate variables and how the biogeographical regions of the North East Atlantic have changed, are described. This involves adapting k-means clustering and regression

to the PCA output. k-means clustering is a technique for finding groupings of variables within a dataset and so can be used on the output of the principal component analysis of the CPR data to define regions. This chapter will describe how these regions might be interpreted differently depending on whether the clustering is carried out on the common components or the loading vectors. Linear regression can be used to model a relationship between an outcome and one or more predictors. In the case of the CPR data this will primarily be used to investigate the relationship between summaries of the biological data and different climate variables. In this chapter we will discuss how to interpret the output from linear regression. Linear regression models might also sometimes be used for prediction, although this relies on certain assumptions. Another measure of the relationship between variables that is discussed in this chapter is the Pearson's correlation coefficient.

2.1 Smoothing Methods

In the introduction the irregularity of the sampling of the CPR data and the need for interpolation to estimate missing values was discussed. Figure 2.1 shows the abundance of a single species across time at a single location before any smoothing, from which it can be seen that there is clearly missing data at some time points, which forms the motivation for smoothing the data. Results are also more interpretable when presented on a regularly spaced grid.

There are various approaches to interpolating irregularly sampled data. For some of these methods the data can be interpolated and smoothed at the same time, where the aim of interpolation is to estimate missing values and the aim of smoothing is to reduce the presence of noise (see Daubechies and Unser [158] and Stein [147]). One of these is the inverse distance method as used on the WinCPR data [51] and another is kernel smoothing [167]. These methods both involve estimating missing values using nearby data points, which makes use of the spatial structure

of the dataset. In geostatistical methods missing values are estimated by taking a weighted sum of the values at nearby locations [160]. In the case of the CPR dataset the spatial structure is used in geostatistical methods to estimate missing values by assuming that locations closer together will be ecologically similar, e.g. Locations close together will have similar habitats and thus will have similar abundances of particular taxa. Inverse distance smoothing, which is used on the WinCPR dataset, takes the weights to be one divided by the distance between the location at which the abundance is being estimated and the location at which a sample exists. Some smoothing can be carried out using inverse distance interpolation but since the weights are undefined at locations at which observations exist, the smoothing can only be carried out in between observations. Typically the estimate is a weighted sum of all the values at all the locations within a certain distance divided by the sum of the weights. This means that more weight is given to locations closer to where the abundance is being estimated, based on the assumption that the closer the location the more similar it will be. In kernel smoothing instead a function symmetric about zero is used [167], which is dependent on a parameter that controls the amount the data is smoothed. It is based on similar assumptions to the inverse distance method, since closer locations are given more weight. Kernel smoothing also allows for some smoothing to be done at the same time as interpolation.

In the case of the CPR dataset an equation for the interpolated data can be written. Supposing l_i are locations at which observations are made, l is the location at which the abundance must be estimated, t is time and $Y^{(p)}(l,t)$ denotes the abundance of species p at location l given time t, then a general equation for the estimates can be written,

$$Y^{(p)}(t,l) = \sum_i \left(\frac{w_i(t,l)\hat{Y}^{(p)}(t_i,l_i)}{\sum_j \left(w_j(t,l)\right)} \right), \tag{2.1}$$

where w_i are some weights dependent on l_i and t_i. The CPR data is regularly

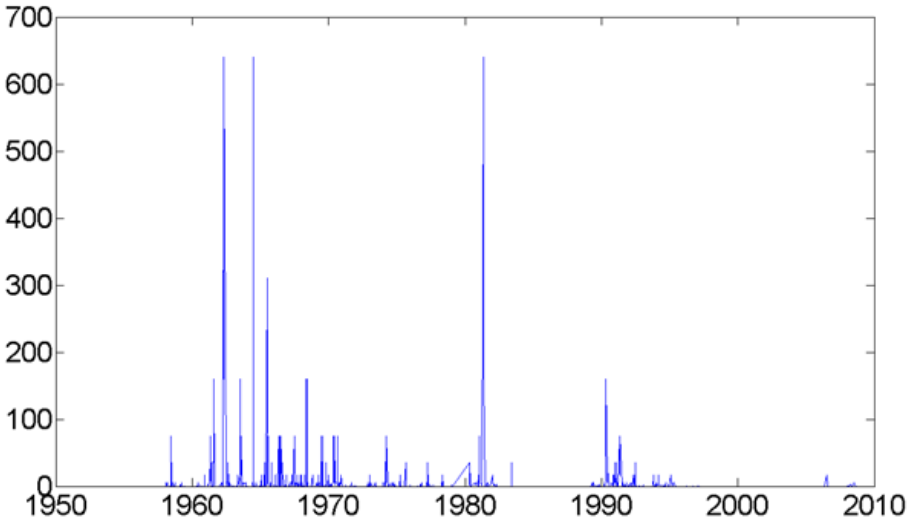


Figure 2.1: Plot of the raw abundances of *Calanus finmarchicus* across months for a single one degree by one degree region.

sampled in time so interpolation will be typically carried out only across the spatial dimension, except where there are too few samples in space in which case some interpolation will be carried out in time as well. It may also be necessary to smooth across space and time for the purpose of noise reduction. In the inverse distance method used on the WinCPR dataset $w_i(l) = \frac{1}{|l-l_i|}$, unless $l = l_i$ in which case the observation must be taken directly and can not be smoothed. It is assumed that locations closer together will be similar in terms of the abundance of different taxa. This follows directly from the assumption ecoregions close together are likely to be subject to similar climate effects and thus will be able to support similar species, which is a reasonable assumption based on plankton physiology and morphology [68]. So the abundance of species p at time t and location l, where p and t are both taken to be fixed, is given by the weighted sum of abundances at nearby locations where observations exist and the weights are the inverse of distance between the locations. Thus locations closer to the location be estimated are weighted more heavily. This can be written as

$$\text{For } l \neq l_i \qquad (2.2)$$
$$Y^{(p)}(t,l) = \sum_i \left(\frac{\frac{1}{|l-l_i|} Y^{(p)}(t,l_i)}{\sum_j \left(\frac{1}{|l-l_j|} \right)} \right).$$

Here l_i are the discrete locations where observations are made and l is the location at which the abundance is being estimated. Where there are insufficient observations in space at a particular time point and near the location that is being estimated some interpolation is instead done in time. For the case $l = l_i$ the function is undefined, meaning it can not be used directly to smooth the data at locations where observations exist but rather to estimate missing values and to smooth the data only at locations where no observations exist.

An alternative to using the inverse distance directly is to take the weights to be kernel functions of the distance [167], which are symmetric functions dependent on a bandwidth parameter. As with the inverse distance method locations closer to the one being estimated are given more weight than those far away, which is done by choosing an appropriate kernel function that will take higher values near zero. This means that the same assumptions about spatial smoothness of the abundances must be made. The main distinction from inverse distance method is that the kernel interpolation method allows one to specify a bandwidth, a parameter which gives more control over how much one smoothes [167]. In contrast to the inverse distance method, which smoothes differently everywhere, kernel methods have a fixed bandwidth across space. A larger bandwidth will smooth the data more by giving more weight to locations over a larger range of distances, whilst a smaller one will smooth less by only highly weighting locations that are very close to the location at which the value is being estimated [167]. This means that the data can be smoothed to reduce the amount of noise simultaneously with the interpolation process and the degree of smoothing can be controlled. Whilst with the inverse distance method

where an observation exists the data can not be smoothed, i.e. that observation is retained because the inverse distance is not defined, in kernel smoothing one can smooth at locations where observations exist. Some care must be taken in specifying the bandwidth parameter [167] as if too large a bandwidth is used the smoothed data will loose some potentially important detail resulting in bias of the smoothed value. Where the bandwidth is large the mean squared error between the observed data and the smoothed data will be relatively large, which means that small scale features of the data will be smoothed out. Conversely if the bandwidth is instead too small the data will be insufficiently smoothed and the resulting data may appear noisy because the variance has not been reduced through the smoothing. Therefore the bandwidth should be chosen in order to minimise the trade-off between the smoothness of the results and the increase in mean squared error. The equation for the kernel smoothed data can be specified if the kernel function is denoted as $K(.)$. If $Y_j^{(p)}$ is the abundance of species p at location l_j and time t_j and H is some bandwidth matrix then the estimated abundance at location l and time t can be written as

$$Y^{(p)}(t,l) = \frac{\sum_j K\left(\begin{pmatrix} l-l_j \\ t-t_j \end{pmatrix}^T H \begin{pmatrix} l-l_j \\ t-t_j \end{pmatrix}\right) \hat{Y}_j}{\sum_i K\left(\begin{pmatrix} l-l_i \\ t-t_i \end{pmatrix}^T H \begin{pmatrix} l-l_i \\ t-t_i \end{pmatrix}\right)}. \tag{2.3}$$

The bandwidth matrix H determines how much the data is smoothed. The larger the entries of H, the more weight is put on distant locations. One potential choice of kernel is a Gaussian function, which satisfies the property of the function taking highest values around zero and smaller values at larger inputs. If H is a bandwidth

matrix then the multivariate Gaussian kernel can be written as

$$K(u; H) = \frac{1}{\sqrt{2\pi|H|}} \exp\left(-u^T H u\right). \tag{2.4}$$

H in this case is a 3×3 matrix. The diagonal entries determine how much one smoothes across each dimension and the non-diagonal entries determine the smoothing between different dimensions, which are non-zero if it is thought that the dimensions are related. In the above model some smoothing is done in both space and time but time might instead be fixed to smooth only in space by fixing t and taking H instead to be a 2×2 matrix. Under the assumption that there is no covariance between the different dimensions and that the bandwidth is kept the same across both spatial dimensions, i.e. the data is isotropic in space, then H is specified to be a diagonal matrix with entries (h_1, h_1, h_2). This will smooth the same amount across longitude and latitude but will allow the smoothing across the temporal dimension to be different. Although other methods will not be discussed in detail, various alternatives to inverse distance interpolation and kernel methods also exist [168], such as: splines, which are smoothing functions based on piecewise polynomials or interpolation methods based on Fourier transforms, which are useful for periodic data.

2.2 Principal Component Analysis

In the introduction it was discussed how PCA might be used as a method of variable reduction on the CPR dataset [86]. Recall that PCA finds directions or maximal variability within a dataset by taking weighted linear combinations of the variables (see section 1.3) and the resulting common components can be used to summarise the structure within the dataset. Although in theory there are as many components as the number of variables, it is possible that most of the variance will be explained in the first few [86], which means the data can be described using a smaller subset

of components rather than using the full set of variables. PCA can be thought of as a data analysis technique which finds an orthogonal representation of the structure in a dataset where the number of variables is large. This is achieved by taking the principal components to be weighted sums of the variables in the dataset. The first component extracted is the sum of variables that explain the largest proportion of the variance within the dataset. The second is the combination that explains the greatest proportion of the variation once the first component has been removed and is restricted to be uncorrelated with the first component. Subsequent components are found in the same way. In other words PCA finds the linear combinator of variables that explain the most variation and removes them and then finds the linear combination of variables that explain most of the remaining variation at each stage [85]. This means that PCA gives an orthogonal representation of the data in which the majority of the variability may be explained by the first few components [85], which is an equivalent to an uncorrelated representation where the error terms are minimised at each stage. This can be useful for interpreting structure in a dataset where the number of variables is large, as when most of the variation is described by the first few components it can be used as a variable reduction technique. The restriction that the common components are uncorrelated means that each component will describe a different mode of variability through the dataset.

If Y is an $n \times p$ data matrix, where n is the number of observations and p is the number of variables then the loading matrix A can be found by doing a Singular Value Decomposition on the covariance matrix Σ of Y, to give

$$\Sigma = A^T S A \tag{2.5}$$

.

The diagonal values of the matrix S, which are the eigenvalues of Σ, give the explained variances associated with each principal component. In theory there are as many components as there are variables, however it is hoped that most of the

variance will be explained by a number of components that are much fewer than the value of p. Therefore the number of components to retain should be chosen so as to explain the variability in the dataset sufficiently well whilst ensuring that the results are interpretable. There are various methods for determining how to choose the number of principal components to retain, which in the CPR survey has the physical interpretation of corresponding to the number of different groups. One of the most frequently used is to threshold on the total explained variance [86].

In the case of the CPR data the number of species or locations is large, which means structure can be difficult to interpret intuitively [31, 25]. It may be easier in this case to reduce the number of dimensions by either looking at joint behaviour over space or 'species communities'. This can save on computational effort, as further analysis can be performed on a smaller number of variables, and can be far more intuitively interpretable than trying to decipher dominant patterns by eye from all the variables. PCA is often described as being 'model-free', in that it is an analysis technique that does not rely on underlying distributional assumptions, but it is possible to view it as an unobserved components model [45, 61], where in the case of the CPR data the unobserved components are responses to different climate indices. In a general model both the loadings and the common signals will be allowed to depend on both space and species and the components are allowed to depend on space, time and species. Hence a model can be written so that each variable is a sum of common signals,

$$Y^{(p)}(t,l) = \sum_{j=1}^{P} a_j^{(p)}(l) z_j^{(p)}(t,l) + \epsilon^{(p)}(t,l). \tag{2.6}$$

Here $z_j(t)$ represents the jth principal component as a function of locational time, $a_j(l)$ are the weights, $Y^{(p)}(t,l)$ is the pth variable, i.e. the pth species, in locational time and $\epsilon^{(p)}(t,l)$ is some random noise, which is assumed to be normally distributed with mean equal to 0. Here the weights are assumed to be dependent on

location and species only but an alternative model could allow the weights to be dependent on time. Each of the signals is calculated, as described by the equation, to be weighted aggregations of the individual variables. There is a trade-off between ease of interpretation and explained variance as a larger number of components will describe more of the variation within the dataset but will be less interpretable. Equation 2.6 can also be written in matrix form.

$$Y = ZA + \epsilon \tag{2.7}$$

The matrix Y contains each variable as a single column and rows contain different observations. Where n is the number of observations and p is the number of variables the dimensions of Y are $n \times p$. So in the case of spatial PCA the columns of Y are different locations and the rows are different time points, i.e. the i, j entry of Y is the abundance of a fixed species at time t_i and location l_j. The rows of A are the loadings for each component and the columns of Z are the principal components, with the rows of Z relating to different observations. In spatial PCA the rows A are spatial locations (see equation 2.8), whilst the columns relate to different components, and the rows of Z are different time points, with each column being a different PC. ϵ is the matrix of random noise.

Typically a more restricted model will be used for the CPR data, where the loadings and common components depend only on particular dimensions of the data. For example the weights might be restricted to depend on space only and the common components will depend on time only, with species being fixed. There are various ways of analysing the CPR dataset using PCA, for example Beaugrand et al [25] use a method called spatial PCA on the CPR dataset to find dominant spatio-temporal patterns of abundance for fixed indicator species. Here the species p is fixed and the weights depend only on the location, whilst the signals depend only on time. In matrix form the data matrix Y is $n \times m$, where n is the number of time points and m is the number of locations. Columns of Y represent locations

and the rows represent observations across time. The signals will then inform us of the common behaviour of a single species over all space. In spatial PCA the dimensions of the covariance matrix would be the same as the number of locations. Spatial PCA can also be written in the same for as equation 2.6. In spatial PCA the abundance of species p at time t and location l might be written as

$$Y^{(p)}(t,l) = \sum_{j=1}^{P} a_j^{(p)}(l) z_j^{(p)}(t) + \epsilon^{(p)}(t,l). \tag{2.8}$$

Whilst Beaugrand et al [25] use PCA to analyse the spatio-temporal behaviour of indicator species, PCA can be used in different settings. There are several ways of viewing PCA in relation to the CPR data, either calculating weights as functions of species, space or time. Whether the weights and signals depend on space, time or species results from how the matrix they are calculated from is written. PCA often is calculated from the CPR data by fixing one of the dimensions and taking the weights to be functions of one dimension and signals to be functions of another. In Spatial PCA, for example, the species is fixed and the weights are taken as functions of space, with signals depending on time. An alternative formulation would be temporal PCA, which would have weights as functions of time and signals as functions of space for a fixed species. Finally species PCA fixes the spatial location and takes the weights to be dependent on species with the signals dependent on time. In general the rows of Y determine what the signals depend on and the columns of Y determine what the weights depend on.

2.2.1 Decompositions of the CPR Dataset

The different formulations of PCA on the CPR data could be viewed in terms of tensor decompositions. Tensors in the field of mathematics and statistics are multi-dimensional arrays [91], such that a first order tensor is a vector and a second order tensor is a matrix. Tensors of order three or higher are referred to as higher order

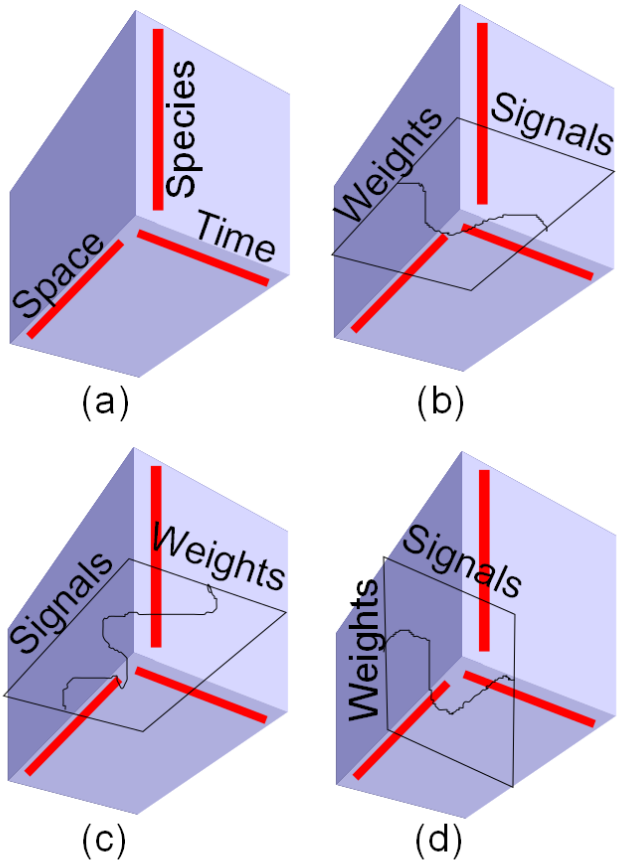


Figure 2.2: The CPR dataset can be thought of as a 3-dimensional space with abundance dependent on time, space and species. The different forms of PCA fix one of these dimensions. The dimensions of the resulting 2-dimensional matrix determine what the weights and signals found using PCA depend on. (a) Representation of the CPR dataset. (b) Spatial PCA. Species has been fixed and so PCA is performed on a time by location matrix. (c) Temporal PCA. Species has been fixed but now PCA is performed on a location by time matrix to give weights dependent on time. (d) Species PCA, where space has been fixed resulting in a time by species matrix. The weights found by PCA depend on species.

tensors. The CPR dataset could be viewed as a third order tensor, if one considered space, time and species as different orders. It could be considered forth order if the two spatial dimensions were considered separately but here we treat them together. Some work on tensor decompositions was done in the 1960's by Tucker [156, 157], who referred to it as three-mode factor analysis. Few previous studies of the CPR dataset have made use of the full multivariate spatio-temporal structure, possibly due to the complexity of such analysis. Some effort was made by Beaugrand, Ibanez and Reid [24] in to developing approaches to analysing the three dimensions simultaneously. They explore the spatio-temporal structure of the CPR dataset across 11 species of Copepod using a method called three mode PCA. In this method PCA is carried out across a species by space-time matrix, which means the columns are functions of species and the rows are functions of both space and time; a time by species-space matrix, and a space by species-time matrix. This returns eigenvectors, which are equivalent to the loadings found using PCA, which are functions of species, time and space respectively. These are then used to find a matrix of the interdependence between these three modes. They refer to this method as three mode PCA. In this thesis the multivariate structure of the CPR dataset is instead studied by first fixing the spatial location and then producing summaries of the structure across the locations.

Figure 2.2 shows the CPR data as a three-dimensional space with the dimensions as time, space and species, along with the tensor decompositions of this space where one dimension has been fixed. The different decompositions can be thought as of different ways of analysing the CPR dataset using PCA. Here (a) shows the CPR dataset as represented by a three-mode tensor and (b), (c) and (d) show three possible tensor decompositions of the CPR dataset, where one of the three dimensions have been fixed. Spatial PCA is the form where the species has been fixed and the weights are functions of space. This means that the principal components represent the dominant modes of variability for a single species across time over

space. This is shown in (b). Figure 2.2 (c) depicts species PCA, where instead of fixing the species the spatial location has been fixed. The weights are functions of species and the common components represent dominant modes of variability in time across species assemblages for a fixed location. Figure 2.2 (d) depicts temporal PCA, where the species has been fixed and the weights are functions of time. The common components then represent dominant modes of variability for a fixed species across space. Principal component analysis can be thought of as a lower order tensor decomposition [91]. The different forms of PCA on the CPR dataset are tensor decompositions carried out when one dimension of the data has been fixed. If the dataset is viewed as a three dimensional space, with the dimensions defined by space, time and species respectively, each different type of PCA is analogous to finding a pathway through a slice of this three dimensional space that describes most of the variability constrained to that slice. Hence when one dimension is fixed then weights can be taken as functions of one of the remaining dimensions and common components as a function of the other.

2.2.2 Spatial PCA

Many previous studies of the CPR data have looked at Spatial PCA [14, 11, 31] (see figure 2.2 (b) and equation 2.8)), which is PCA with weights as functions of space for a fixed species. The abundances of species are modelled using an unobserved components model and in this case the interpretation of the signals are the dominant trends in the species behaviour across time and the weights are interpreted as the spatial patterns of these trends. One might think of the trends as responses to climate signals and the weights as showing how these climate signals influence the species abundance across space, in particular if the common components correlate strongly with a particular climate index. If an individual species is fixed then the underlying unobserved signal in time can be modelled as a weighted sum of the signal at each location. In equation 2.6 p has been fixed and the weights are dependent on l only,

whilst z_j, the signals, are dependent on t only. Previous studies have looked at how the temporal signals $z_i(t)$ can be related to various climate signals, such as the Northern Hemisphere Temperature and the North Atlantic Oscillation index [31, 25], as described in section 1.6. Large weights can be interpreted as regions where the signal is dominant, i.e. the most important locations to that species. If there is a strong correlation between the PCs and a particular climate index they may also represent the pattern of influence of that climate index on the species. The signals are indicative of long term changes in the species behaviour in time, so are subject to climate variability in either long term trends or natural oscillations in climate.

Spatial PCA is typically used for studying the variability of indicator species. Reid and Beaugrand [31] study the relationship between plankton and fish and so use Spatial PCA to find the spatial and temporal patterns of those species which provide a food source for fish [31]. Spatial PCA has also been used to study the CPR data at a seasonal scale [26] with the time signals representing seasonal patterns and the weights representing those locations at which certain seasonal patterns occur. Cannaby *et al* [43] in contrast use Spatial PCA on the sea surface temperature signal instead, as this enables one to find the spatial pattern of the influence of different climate indices on local temperature. The principal components in the study by Cannaby *et al* are matched to different climate indices using linear regression. A similar sort of analysis is carried out on the SST signal in a paper by Beaugrand *et al* [30], where they study the effects of 'regime shifts' on Calanus Copepod species. They identify the temporal signals from the PCA on the SST data restricted to the North East Atlantic with known climate indices, including the NHT warming trend and the NAO. They compare the weights, which they identify as the spatial patterns of these climate indices, with changes in the distribution of the plankton species. This shows how the spatial influence on climate leads to long term changes in plankton assemblages. In this thesis spatial PCA will be used to ex-

plore the behaviour of several indicator species, including two species of copepod; total copepod abundance, which has not been analysed in this way before, and phytoplankton colour. This will then be extended to exploring the behaviour of species communities using species PCA, which have not been studied in depth before. In a similar way to Cannaby *et al* the variability of the sea surface temperature will also be explored using spatial PCA. Sea surface temperature can be decomposed into spatial weights and temporal signals. Here the signals will represent dominant temporal trends in the sea surface temperature and the weights will represent the pattern of the influence of these trends on sea surface temperature across space. The temporal signals can also be compared with climate indices in order to try and better understand what drives sea surface temperature.

2.2.3 Species PCA

Instead of fixing the species when carrying out PCA on the CPR dataset we can fix the spatial location within some regularly sampled grid (see figure 2.2 (c)). The data matrix Y is now $n \times p$, where in n is the number of time points and p the number of species at a fixed location. In this case the weights are allowed to vary across species. This has the advantage that it allows one to study multiple species at a time and to find groupings in the species based around the weights. The signals now represent *joint* temporal behaviour of subgroups of species. Since species that interact with one another or share similar physiology are likely to respond to climate in similar ways, it is reasonable to assume the presence of functional groups of species within the dataset. Referring back to equation 2.6, the weights a are now functions of species, p, and space ,l, only, where l has been fixed. The components z are dependent on t and l but l is treated as fixed and so the components are dependent on t given l. The noise is dependent on p and t given a fixed value of l;

$$Y^{(p)}(t;l) = \sum_{i=1}^{P(l)} a_i^{(p)}(l) z_i(t;l) + \epsilon^{(p)}(t;l). \tag{2.9}$$

$P(l)$ is the number of principal components that are retained at location l, which is determined by the culmative explained variance. The number of components is allowed to vary across locations and can be interpreted as the number of distinct assemblages across space. This being the case the abundance is written as $Y^{(p)}(t;l)$, which is the abundance of species p at time t given location l. Here the signals inform us about the common behaviour of all species at a fixed location. $z_i(t)$ is informative as to how the ecosystem as a whole is changing, compared to spatial PCA where the signals represent dominant modes of variability for a single species or climate index. In species PCA for each PC species that have large weights will have similar functional behaviour to one another and this allows functional groups of species that behave together can be found directly from the data without any prior assumptions about the ecology. The pattern of the weight vector in space can be informative as to how the functional groupings change across locations, i.e. the groups of species that have large weights in certain habitats will be different to those that have large weights in other habitats, which in turn can be used to define ecoregions, for example by clustering on the weight vectors. In the CPR dataset some taxa will be more dominant at certain locations than others, due to the fact that species are adapted to particular ecological niches [75]. Moreover rarer species might be inconsistently sampled and might bias the sample, thus meaning it might be of interest to focus only on the most dominant species at each location. This leads to considering the weight vector as sparse.

2.3 Sparse Principal Component Analysis

Where a dataset contains a few true non-zeros and decaying values a sparse model can be used [171, 44, 6]. Sparsity is a term for when only a few parameters take

true non-zero values. Since in most experimental data true zeros are rarely observed due to the effect of measurement error, sparsity is usually considered in terms of a small number of non-zero values with the rest of the parameters taking small or decaying values. In section 1.3 it was discussed how sparse models might be used for gene expression data, where a small number of genes might be important but that many genes show some level of expression [44]. This concept is being extended to the CPR dataset in this thesis. The spatial heterogeneity of the climate means that not all species will be equally important at every spatial location. Whilst some will prefer warmer climates others will prefer colder waters [12] and the spatial distribution of some species will be driven by wind intensities [56]. As with the gene expression data certain variables at each location will be more important for describing the joint variation. Subgroups of species should be highly correlated, due to the fact they will share similar responses to climate and will all be subjected to the same climate trends. Therefore if it is assumed that at each fixed location only a few species are truly significant (see Helaout and Beaugrand [75]) and that the rest are observed as the effects of noise or represent very rare species that are only occasionally observed then this leads naturally to modelling the loading vector under species PCA as sparse. There are various approaches to incorporating sparsity in to the PCA algorithm [172, 50]. These methods are generally based around some approximation to penalising the cardinality, i.e. the number of non-zeros, in the loading vector. Since it is assumed that only a few variables (e.g. species at a particular or genes in patients with a condition) are truly important the number of non-zeros in the weight vector should be restricted. Penalising on the cardinality directly is computationally intensive and so some approximation to the cardinality penalty is generally required [172, 50]. An alternative might be to do an iterative search. If the number of non-zeros is fixed then the variable that explains the largest proportion of the variability can be added at each stage until the required number of variables has been selected. Such an approach is known as a Greedy algorithm

[116].

Supposing it is assumed that the true data set is generated from some common components with sparse weights plus noise, then the observed principal components will have non-sparse loadings. If it is assumed that those loadings based on noise will be relatively small then one approach might be to simply threshold on the result [172]. The disadvantage of this is that an appropriate threshold can be difficult to select by eye and thresholding can give misleading results [172]. Another approach is to incorporate a penalty directly in to the PCA algorithm. Zou, Hastie and Tibshirani [172] develop a method incorporating a penalty based on the LASSO and D'Aspremont et al [50, 49] use a direct formulation via a semi-definite programming method. Both of these methods use an approximation to the cardinality penalty. Alternatives to these proposed methods include greedy methods [116].

Zou, Hastie and Tibshirani's method [172] is based around the LASSO penalisation, which is a penalty that shrinks small values and therefore tends to favour sparse solutions. By incorporating an additional penalty in to the PCA algorithm they return sparse versions of the loading vectors. The PCA algorithm can be reformulated as a least squares problem, namely minimising $\sum ||\mathbf{Z} - a_j Y_j||^2$ with $\mathbf{Z}$ being the matrix of the principal components, Y_j the variables and a_j the loadings. In order to force the loading vectors to be sparse a penalty must be incorporated in order to shrink small loadings towards zero. Zou, Hastie and Tibshirani [172] incorporate a penalty called the *elastic net*, which is a linear combination of the Tikhonov penalty and the LASSO. Ideally one would incorporate an L_0 or cardinality penalty, which is equivalent to penalising on the number of non-zeros in the solution. This is very computationally slow and so in this approach the LASSO approximates L_0 norm. The LASSO penalty has the advantage that it tends to prefer solutions with few non-zero entries and that penalising on the L_1 norm is roughly equivalent to penalising on L_0 [151].

The ridge or Tikhonov penalty is typically used in ill-posed problems, where

there is no unique solution, in order to regularise the problem. The Tikhonov penalty is based on a Euclidean norm or L_2, which is given by $||y|| = (\sum_i y_i^2)^{1/2}$. In general an L_p norm takes the form $||y||_p = (\sum_i y_i^p)^{1/p}$. By adding the additional constraint of the Tikhonov penalty to a given optimisation problem it is forced to have a unique solution. So where one wishes to optimise $||Ax - b||^2$, which has no unique solution, instead the problem becomes $||Ax - b||^2 + ||\Gamma x||^2$. The regularisation matrix Γ determines the type of solution that is computed. The least absolute shrinkage and selection operator (LASSO) is based around an L_1 norm, which has the form $||y||_1 = (\sum_i |y_i|)$. The LASSO approximates the L_0 penalty but the number of selected non-zero variables is constrained by the number of observations, which is suboptimal in a situation where there are far more variables than observations ($p >> n$). This limitation is overcome by incorporating the Tikhonov penalty [172].

The *elastic net* is dependent on two parameters, λ and λ_1, which control the number of variables that are selected. The two parameters which are included in the optimisation problem control the level of sparsity, i.e. how many variables have non-zero weight. The sparsity level is allowed to vary across different components. The code for computing the SPCA in this way is written in an R package called 'elasticnet' and according to Zou et al [172] is computationally efficient. However the *elastic net* formulisation has the disadvantage that it sacrifices some of the orthogonality of the resulting principal components. According to Zou et al [172] this trade-off is relatively small. A Bayesian version of the LASSO can be formulated by a Laplace prior on the weights and this can be extended to produce a Bayesian version of the *elastic net* [101], in which case both parameters are selected at once avoiding a double shrinkage issues.

Another disadvantage of the *elastic net* formulation is that it results in a non-convex optimisation problem [50]. A function g is convex if $g(\alpha x + (1 - \alpha y)) \leq \alpha g(x) + (1 - \alpha)g(y)$ [168, 169]. Non-convex optimisation problems are compu-

tationally slow to solve because often a global optimum is inefficient to compute [50] and local optima are generally insufficient because they are difficult to analyse [168, 169]. In order to solve this problem D'Aspremont et al [50] incorporate a sparsity criterion in to the PCA problem and then form a convex relaxation, whereby they replace the non-convex constraint with a weaker convex constraint, which takes the form of a semi-definite program. This method is similar to that of Zou et al because the penalty used approximates a cardinality penalty and thus forces some of the loadings to be zero. Like Zou et al, D'Aspremont et al sacrifice some of the orthogonality of the principal components. Although D'Aspremont [50, 49] et al suggest this trade off is relatively small, it is disadvantageous because it will mean that the PCs do not explain all the variance separately and so might make interpretation more difficult.

Initially D'Aspremont et al incorporate a cardinality penalty but as previously described this approach is very inefficient to solve. In order to find a solution they make use of the fact that PCA can also be viewed as an eigenvector problem. If Σ is the covariance matrix of the data matrix Y, which is the matrix of the joint variability between variables given by Y^TY, then the loading vectors are the eigenvectors of Σ. If the variables are orthogonal then the covariance matrix will take zero values everywhere but the diagonal, in which case the PCs are equivalent to the variables, however if the variables are not orthogonal then the covariance matrix will have some non-zero entries other than on the diagonal. D'Aspremont et al write the PCA algorithm in matrix form and introduce a cardinality constraint to the optimisation problem. The loading vectors in PCA can be calculated by finding the eigenvectors a of the data covariance matrix Σ. If $V = aa^T$ then the eigenvalue problem can be shown to be equivalent to minimising $\text{Tr}(\Sigma V)$, where the $\text{Tr}(.)$ is the trace of a matrix given by the vector of the elements along the diagonal. The cardinality of the loading vectors is constrained to be less than or equal to k by adding the constraint $\text{Card}(V) \leq k^2$ to the optimisation problem. D'Aspremont et

al then replace this constraint with a weaker but convex constraint. For any vector u if the cardinality of u is equal to q then $||u||_1 < \sqrt{q}||u||_2$, which means that the cardinality constraint can be replaced by $\mathbf{1}^T\mathbf{V}\mathbf{1} \leq k$, where 1 is the identity matrix. In this formulation sparse principal component analysis can be written as the optimisation problem: minimise $\text{Tr}(\mathbf{\Sigma V}) - \rho\mathbf{1}^T\mathbf{V}\mathbf{1}$. A parameter ρ controls the number of non-zeros k and sparse PCA is defined as an optimisation problem, which is coded in MatLab.

An alternative to both the above methods is to use a greedy algorithm. A greedy algorithm finds the optimal solution at each stage of a problem [162, 80]. In the case of a k sparse solution, which is a solution with k non-zero entries, for example, a greedy algorithm would begin by selecting the optimal 1-dimensional solution and then taking then next best value at each stage until one has a k-sparse solution. In the case of sparse PCA where one desires a loading vector with k non-zeros one might begin by calculating non-sparse loading vectors and selecting the variables with the k largest weights. PCA is then recomputed on just the k selected variables. To compute subsequent PCs one must first remove the variance explained by the previous PC by subtracting the PC times the loading vector from the original data matrix. Subsequent PCs can then by computed by repeating the same process on the residuals. In this study the greedy method is used to find sparse loading vectors. In this case the sparsity k is used directly. The proportion of non-zeros is $\pi = k/p$ and this is termed the sparsity parameter.

Sparse PCA can also be thought of in terms of sparsistency, which is the probability of true zero values being ascribed a zero weight in the computations [93]. Ideally this probability should tend to one. The possible limitations of the Greedy sparse PCA algorithm are that it is not guaranteed to satisfy sparsistency and that it can be computationally intensive. The advantage is that it preserves the orthogonality of the principal components, unlike the other methods described.

2.3.1 Using Mixture Models to Determine the Sparsity Parameter

One remaining issue is how to select an appropriate sparsity parameter so that the right amount of signal rather than noise generated structure is kept. Often it is easier to consider k, the number of non-zero loadings, than to consider λ or ρ directly as they are less clearly interpretable. If it is assumed that the non-sparse loadings consist of noise drawn from one distribution and signal drawn from another then they can be modelled by a so-called mixture distribution [153]. A mixture model is used when a dataset is assumed to have been drawn from two or more different probability distributions [153, 109, 58]. The mixture model is dependent on the parameters of each distribution and a parameter which determines the proportion that belongs to each distribution, which for the purposes of estimating the sparsity level will be the parameter of interest. That is, some of the observations are governed by one process and the rest are governed by another, which may be suitable to consider for sparse datasets that have a mixture of true non-zeros and decaying values [82, 83]. Mixture models are also used in clustered data, where different groups of observations are centred around different points [40, 110]. Supposing the non-zero weights follow a given probability distribution and the noise is drawn from a separate distribution, then the loading vectors computed from the observed data will fit the assumptions of a mixture model. In particular that the data is drawn from two or more different probability distributions. Here a probability distribution refers to a function which gives the probability of an observation taking a particular value. A normal probability distribution takes higher values closer to the mean and lower values further away from the mean for example.

In the case of the CPR data it might be considered that each species abundance can be modelled as a linear combination of climate signals with zero weights in certain locations, where that species is not well adapted. Random noise is then generated from a variety of different sources, including measurement error in the counting process, such as misidentification of species, or from random fluctuations

in conditions. Johnstone and Silverman [82, 83] model sparse vectors as a mixture of two Gaussian distributions, often termed a 'spike and slab' model. In the 'spike and slab' model one of the distributions takes very high values close to the mean of zero and significantly smaller values elsewhere, which means it can be used to model the near zero parameters. In the 'slab' distribution the probability density is spread across a wider range of values, which means it can be used to model the true non-zeros. In the case of the loading vector for the CPR data a mixture of two Laplace distributions appears to capture the structure of the data well, although any appropriate distribution can be fitted to the model.

Supposing the noise is drawn from a distribution with a probability density function $f_1(.)$ with parameters θ_1 and the signal is drawn from a distribution $f_2(.)$ with parameters θ_2 then a likelihood function can be defined. The parameter of interest in this case is π, the proportion of values drawn from the signal distribution and $\pi \times p$ is the number of variables that should have a non zero weight, and θ_1 and θ_2 are nuisance parameters, which is to say unknown parameters which are not the parameter of interest. The likelihood is given by the product of the probability distribution for the mixture model evaluated at each observation. Maximising the likelihood function is effectively finding the parameters for which the data has highest probability [2]. The probability distribution function of the mixture is given by

$$(1-\pi)f_1(.;\theta_1) + \pi f_2(.;\theta_2) \tag{2.10}$$

.

Evaluating this function at each loading will give the probability of that loading being observed under the assumption that the underlying function is this mixture. Assuming f_1 and f_2 are chosen appropriately, π can be found using maximum likelihood or expectation maximisation as follows. The likelihood is

$$f(\mathbf{a}, \theta) = \prod_{p=1}^{P} \{(1-\pi) f_1(a_p; \theta_1) + \pi f_2(a_p; \theta_2)\}. \tag{2.11}$$

Here a_p is the loading for species p, $\mathbf{a}$ is a vector containing the loadings for all species, P is the number of species, π is the proportion of the loadings belonging to the signal distribution, $f_1(.)$ and $f_2(.)$ are appropriate distributions and θ_1 and θ_2 are some parameters. Some care must be taken in choosing the probability distributions f_1 and f_2 in order to properly capture the structure of the data. If these distributions are not appropriately chosen the resulting estimates of the parameters may not be as representative of the data. In the case where f_i is a Laplace distribution the probability densities can be written as

$$f_i(a_p) = \frac{1}{2b_i} \exp\left(\frac{-|a_p - \mu_i|}{b}\right), i = 1, 2. \tag{2.12}$$

Here μ_i is the mean of the distribution and the variance is given by $2b_i^2$. The Laplace distribution is spiked about the mean, which is representative of the distribution of the CPR data because there are a large number of rare species included that have very small weights. Alternative distributions might be used, such as a Gaussian distribution. In the case of the CPR data a Gaussian distribution does not sufficiently capture the probability density of the noise and tends to underestimate the variance of the noise. A mixture of more than two Gaussian distributions could also be used but this would add to the computation complexity because this would mean calculating additional parameters for each distribution.

2.3.2 Temporal Misalignment in Biological Data and Fourier PCA

Another potential problem with fitting sparse PCA to the time courses of different species. In particular whether the assumption of a linear relationship might be violated by the presence of small time lags between species. One problem that occurs often in biological datasets is that of misalignment between time courses. For

example the growth curves of children [140], which is a classical example exhibiting such problems. Although children have similar shaped growth curves, growth spurts occur at different times for different individuals leading to misalignment. When there is misalignment in the data this may cause problems with clustering. Even though the time courses might have similar shapes they are misidentified as belonging to separate clusters depending on the time lag between the different time courses. Another problem with misalignment, relevant to this study, is that PCA does not account for small lags between variables. PCA finds time courses which are linear combinations of variables and so will not naturally group species that behave in a similar way but with a small time lag. One example of this may be predators and prey, which both respond to the same climate signal but the response is seen later in the predator because it is reacting to changes in the abundance of its food source. The predators might respond later to changes in climate because the impact only begins to have an effect after changes have happened to the species they prey upon. One of the assumptions of PCA is that variables will have a linear relationship, which does not hold when there are small time lags between variables. For the purposes of interpretation it is desirable to combine variables that have similar responses together even if there is a time lag between them, due to the assumption that they will be responding to the same climate variable, so it is necessary to accommodate for time lags before carrying out PCA.

One method for accommodating for small time delays in the PCA algorithm is to study vectors in the Fourier domain. Fourier transforms translate data in to a frequency domain, which is often used to find the periods of oscillations in a dataset where these patterns may not be easily spotted in the time domain. A Fourier transform will also transform phase shifts in to multiplicative factors, which the PCA algorithm has no problem with. So long as the data satisfies certain conditions small time lags can be approximated as phase shifts.

If the misalignment is in time, so that $Y^{(p)}(t) = Y^{(q)}(t - \tau)$ for example, then

the misalignment can be approximated as a phase shift so long as the oscillations are concentrated to a small set of frequencies. Supposing the time course of a species can be modelled as an oscillation with amplitude $a(t)$ and frequency $\phi(t)$, then it can be written $Y^{(p)}(t) = a(t)\cos(\phi(t))$. If another species is an oscillation with the form $Y^{(q)}(t) = a(t)\cos(\phi(t)+\phi)$, then the phase is shifted between the two species and in the Fourier domain they will both be transformed to the same representation. The approximation holds if it can be assumed that in the Fourier domain $\tilde{Y}^{(p)}(f)$ is only non-zero for a small set of $f = f_0$, which is equivalent to saying the signal can be represented by a small set of oscillations. If $\tilde{Y}(f)$ is mainly supported near $f = f_0$ then $\tilde{Y}_\tau(f) \approx e^{-itf_0\tau}\tilde{Y}(f)$ and the time shift can over the relevant frequencies be approximated by the phase shift. Supposing a time signal of the form $Y(t-\tau)$, where τ is a time lag, then the Fourier transform can be calculated:

$$\begin{aligned} \tilde{Y}_\tau(f) &= \int e^{-2i\pi ft} Y(t-\tau)dt \\ &= e^{-2i\pi f\tau}\int e^{-2i\pi ft'} Y(t')dt' \\ &= e^{-2i\pi f\tau}\tilde{Y}(f). \end{aligned} \tag{2.13}$$

Since principal components are found by taking linear combinations of variables, PCA is performed on the positive frequencies only. Initially non-sparse PCA is calculated on the transformed data. Since this returns complex loading vectors the sparsity parameter is found by modelling the absolute values of the loading vector as a mixture and the proportion of the variables retained in the sparse components is found in exactly the same way as before. For a *k*-sparse solution those variables with the k largest absolute values are retained. Fourier PCA is then recomputed on those variables with the k largest absolute values on the loading vector. The principal components must be transformed back in to the time domain in order to

be interpretable. In order to return the principal components in the time domain the inverse Fourier transformation of the PCs in the frequency domain is taken. This gives $(1/2)$ Signal $+$ $(i/2)$ Hilbert transform and so the principal component is taken to be twice the real part of the inverse Fourier transform.

In order to find the time delays for each species the values of $\hat{f}$ for which $\tilde{Y}(f)$ is maximised must be computed. τ can be found by taking the angle of the loading associated with the species in question and dividing by $-2\pi\hat{f}$ (see equation 2.14). These values will reveal whether the species is responding before or after the average signal. For each species one can take the time delay in space and then cluster across the different species.

Alternative methods for dealing with misalignment also exist, for example there are several proposed methods for clustering data when misalignment is present. Sangalli et al [140] have developed a *K*-means clustering based method which accommodates for misalignment. Tang and Muller [149] approach this problem via cluster specific warping functions. Liu and Yang [102] take a Bayesian approach to the model by using a B-spline basis incorporating parameters which take in to account the warping and solving using expectation maximisation. Tang and Muller [149] study the same problem in relation to gene expression data and use a time synchronized method in order to cluster the data.

2.4 Cluster Analysis

In order to determine regionalisation of the North Atlantic clustering methods can be used on the output of the PCA. Clustering is a method of finding a number of natural groupings in a data set [168, 81, 60]. One popular method is *K*-means clustering. Here K is used to denote the number of clusters, in order to distinguish from the k that was used to denote the number of non-zeros in a sparse solution previously. Given a pre-set number of clusters K this method uses Euclidean distance to

partition a set of vectors[81] (see section 1.4), where the Euclidean distance is the L_2 norm of the difference between the two vectors. In the first stage a random set of K centres are chosen from the dataset and then each variable is assigned to a cluster based on minimising the Euclidean distance from the centre. The centres are then recomputed by taking the average of each cluster and the process is repeated until the results converge. Since the partition can be dependent on the initial choice of centres the best partition over multiple runs can be chosen by minimising the mean squared error. In general K is defined by the user, although there are methods for choosing the most suitable number of clusters [148].

Other clustering methods include hierarchical clustering, which starts with one variable and selects its nearest neighbour, usually by minimising the Euclidean distance [81]. At each level of the hierarchy the next nearest variable is added, which can be selected by taking the variable nearest the centre of the cluster, the variable with the smallest minimum distance to the cluster or the variable with the smallest maximum distance from the cluster. The choice of distance metric may have a large influence on the results [81] because, for example, the variable with the smallest minimum distance to the cluster may not be the same as the one with the smallest maximum distance or the smallest distance from the centre. This suggests that different distance metrics may lead to different partitions. Alternative methods would include fuzzy clustering, which assigns each variable a degree of membership to each cluster rather than having each variable belonging to a single cluster [81].

Clusters on the output of the principal component analysis can be defined on either the loadings or the principal components. These will have different interpretations in terms of the original biological problem. Clusters on the loadings will identify regions with similar species groupings, whilst clusters on the PCs will determine similarities in long term climate behaviour. Another possible pitfall with clustering on the output of PCA is the issue of mode-mixing [86]. The ordering of

PCs is not necessarily fixed and usually chosen according to their explained variance. Therefore the first PC at one location may have a similar functional behaviour and correlate with the same climate trend as the second PC at another. This is problematic if the explained variance between the first and second PC is similar and does not show a rapid decline. If mode-mixing does occur then this might impact the output of the cluster analysis because the ordering of the components might not be comparable between locations, meaning clustering the output across locations may not give meaningful results. In order to check whether this is an issue the explained variance of the components can be checked to determine whether it does indeed show a steep decline.

K-means clustering might also be used to explore changes in the regions defined by the plankton species over time. This might be done by dividing the dataset in to different sections across time and comparing the output of PCA over the different sections of the data. Under the hypothesis that temperate species have gradually been moving northwards as the waters of the North Atlantic have warmed one would expect to see southern clusters spreading northwards, whilst northern clusters dominated by cold water species would be gradually receding in to polar regions. This can be tested by dividing the data in to two halves: the pre-1985 data and post-1985 data. The analysis can then be carried out on both halves. Regionalisation in both halves of the data is found using k-means clustering on the loading vectors as described earlier.

In order to find the optimal number of clusters it is important to determine the point at which adding more clusters will not significantly improve the explanation of the data. This can be done intuitively by plotting the explained variance or mean squared error against the number of clusters. In general the variance will begin to level off at some point. The point at which the variance levels is chosen as the optimal number of clusters. This is referred to as the elbow method [148]. Other methods for selecting the number of clusters is modelling the data as a mixture of

different distributions and using Bayesian methods to determine the number of distributions, although this may work best for low dimensional vectors [148] because of the computational complexity.

A simulation study is used to test the significance of the results. To test whether the regionalisation found is significant the same methodology can be run on simulated data multiple times. For instance a method similar to bootstrapping may be used to simulate randomised data by selecting time courses and weights at random, sampling from the output of the sparse PCA, to generate data under the same model that has been assumed for the real data and adding random noise. Bootstrapping is a method for determining p-values by sampling with replacement from the true data to get multiple randomised data sets [168]. It is particularly useful when the underlying probability distribution is unknown or difficult to model.

2.5 Regression Analysis

It is important to understand which climate variables are 'driving' the plankton abundance. A climate variable is said to be a driver of abundance if changes in its temporal behaviour have a strong influence on abundance. It is assumed that a climate variable that has a strong relationship with the abundance of plankton then it has a direct influence on the plankton. Evidence for this assumption is found by looking at what is known of the physiology of different species and relating this to the influence of climate variables. For this a simple linear regression model might be used, where the plankton response is thought to be proportional to the climate signals plus a constant [168]. This model can be fitted using least squares estimation. If the plankton abundance is assumed to be equal to a constant plus a number of climate signals multiplied by some unknown coefficients then the coefficients can be estimated by finding the values that minimise the square of the sum of the error terms. The error terms in this case are the true value of the biological variable

minus the abundance estimated from the linear regression model at each time point. The coefficients give a measure of the relationship between the climate variable and the biological variable and a hypothesis can be use to determine whether they are significantly different from zero. This model is used because the abundance of the plankton is thought to be driven by climate [56, 166, 127] due to the way in which environmental conditions interact with their physiology. The response is modelled as linear, as it is reasonable to assume at small perturbations this will hold true, although some care must be taken because in more extreme conditions the relationship may be non-linear. In order to determine whether there is truly a relationship between climate variable and the biological response a hypothesis test can be used [168]. If the correlation at a particular location is significant then the group response of the species at that location is said to be sensitive to that particular driver. A strong correlation will generally be assumed to be indicative of a causal relationship, although this will be justified by looking at whether the causality can be explained by the biology of the assemblage [56, 166, 127]. Typically a causal relationship would also imply a time lag between responses but depending on the measurement scale this is not always possible to observe, e.g. if the lag in response is less than a year for annual data. The relationship between plankton and climate is important to environmental policy makers because it will help them understand what is forcing the ecological behaviour and whether changes are attributable to climate warming trends or to natural oscillations in climate. For signals in time we assume that the temporal principal component at each location can be modelled as a linear combination of covariates:

$$z_i(t) = \beta_{i,0} + \beta_{1,i}c_1(t) + \beta_{2,i}c_2(t) + ... + \beta_{N,i}c_N(t) + \epsilon_i(t), \tag{2.14}$$

where $z_i(t)$ is the ith PC and $c_j(t)$ is the jth physical variable under consider-

ation. Estimates, $\hat{\beta}_{i,j}$, for the coefficients can be found using least squares. The error terms $\epsilon_i(t)$ are assumed to be normally distributed with mean zero. In order to establish whether $\beta_{j,i} = 0$ a hypothesis test with H_0 as $\beta_{j,i} = 0$ and the informal alternative H_a as $\beta_{j,i} \neq 0$ is used. It was previously mentioned that it was assumed that if a variable has a strong relationship with the plankton then it is a 'driver' of abundance. Since climate variables are not necessarily orthogonal, excluding certain physical variables from the model might lead to a stronger correlation with other variables. If a variable is influential on another variable and is influential on plankton abundance, then excluding it from the model could lead to the second variable being incorrectly identified as a 'driver'. This is known as a confounding variable, where a correlation between two variables exists because they both depend on a third. This means that some care must be taken to include all possible 'drivers' in the model and to correctly identify which are important to the plankton. Under the null hypothesis, H_0, there is no relationship between the ith plankton response and the climate driver j in the presence of other drivers, therefore a significant result is indicates that the covariate is an important driver. The test statistic can then be calculated.

$$t = \frac{\hat{\beta}_{j,i}}{se(\hat{\beta}_{j,i})}, \tag{2.15}$$

where se is given by $se(\hat{\beta}_{j,i})^2 = \frac{\frac{\sum_k (z_i(t_k) - \hat{z}_i(t_k))^2}{N-2}}{\sum_k (c_j(t_k) - \bar{c}_j(t_k))^2}$, N is the number of time points and $\hat{z}_i(.)$ being the values estimated from the linear regression model, i.e. $\hat{z}_i(t) = \hat{\beta}_{i,0} + \hat{\beta}_{1,i} c_1(t) + \hat{\beta}_{2,i} c_2(t) + ...$ $\bar{c}_j(t_k)$ is the mean of $c_j(t_k)$. If N is the number of variables then under the null hypothesis t follows a t-distribution with $N-2$ degrees of freedom. Therefore this test statistic can be used to find a p-value. If the p-value ≤ 0.05, say, the null hypothesis is rejected.

In general the correlation will be found across multiple locations by repeating the linear regression on the components at each spatial point. This will result

in a spatial pattern of where each climate driver is important and will help identify 'hotspots' where climate change is having the greatest impact on the ecology. However this means that the problem of multiple testing arises. In the case of multiple testing one must take care to control the false discovery rate [35]. Benjamini and Hochberg [35] attempt to control the false discovery rate by altering the level at which they reject based on the number of tests. P-values are ordered in increasing size and then compared with a function,

$$P_{(i)} < \frac{i}{N}\alpha. \tag{2.16}$$

In this equation $P_{(i)}$ is the ith smallest p-value, N is the number of tests and α is the significance level. This gives a stricter cut-off point under multiple testing, which controls the probability of making a type-1 error. Rejecting a true null hypothesis is referred to as a type-1 error, type-2 being not rejecting a false null. A confidence level of 90% says that if the null hypothesis is true there is less than a 10% chance of the test statistic being statistically significant. Under 90% confidence level the expected number of significant results if the null is true is 10% of the number of tests carried out. A false discovery rate is the rate at which results are found to be significant even though the null hypothesis is true [168], i.e. the rate of making a type-1 error. In the case of a single hypothesis test type-1 errors are only as frequent as the p-value. However when multiple tests are carried out the probability of making a type-1 error rises. In this case it is important to have a stricter rejection rate in order to control for type-1 errors. The Benjamini and Hochberg correction is valid when the tests are independent but also under many cases where there is dependency [36].

The linear regression model can be used for spatial variables, such as the number of components or the sparsity parameter. However since these variables are not continuous a link function must be used to transform them so they can be used as dependent variables in the regression model. Suppose the parameter $p(l)$ is the

number of components at each location. The proportion of the total number of components that are retained at each location is $p(l)/P$, where P is the number of variables (i.e. species), and if it is assumed that $p(l)/P$ is linearly dependent on physical drivers across space them an equation can be written in the form:

$$\text{logit}(p(l)/P) = (\alpha_0 + \alpha_1 d_1(l) + ... + \alpha_M d_M(l) + \mu(l)), \tag{2.17}$$

Where $\text{logit}(x) = \log(x/(1-x))$. The logit link can be used to transform variables that take values between zero and one to scale variables that can take any value from minus infinity to infinity, which means the logit of the proportion can be used as an outcome in linear regression. In this equation $d_i(l)$ are spatial variables. Again this model can be fitted using least squares estimation but maximum likelihood estimation is more commonly used.

Linear regression combined with principal components analysis can be likened to canonical correlation analysis. Canonical correlation analysis is used in a multivariate setting where both the number of responses and the number of covariates can be large [128, 134]. Where the number of dependent variables is large relationships with the independent variables can be difficult to interpret and so this leads to seeking methods that reduce the number of variables. In canonical correlation analysis an orthogonal transform is carried out on the matrix of responses to reduce the number of variables and thus decrease the computational intensity of the regression analysis, since the regression will have to be carried out for a smaller number of variables than the full dataset. In this study the PCA acts in a similar way to this orthogonal transform. Rather than carrying out a linear regression on each species individually, it is carried out on joint responses which are represented by the principal components. Since the number of components is thresholded by the culmative explained variance this means carrying out far fewer computations than would be required for the same analysis on each species.

The relationship between the physical variables and the biological variables

might also be described using correlation coefficients between two variables and the amount of variance explained in one variable by the other. One such measure of linear dependence is Pearson's correlation coefficient, which is calculated from the covariance of the two variables divided by the standard deviations of both variables multiplied together. It takes values between -1 and +1, with a negative value indicating a negative linear relationship, so that one variable increases as the other decreases, and a positive value indicating a positive relationship, with both variables increasing together. The Pearson's correlation coefficient can be seen as a measure of how much two variables vary together. The proportion of the variance explained can also be calculated from the squared value of the Pearson's correlation coefficient, meaning that when the absolute value is of the correlation coefficient is close to 1 a large proportion of the variation is explained by the model. Using these measures it is possible to determine the strength and nature of the relationship between the physical and biological variables, which will allow us to interpret the effects of these climate indices on the ecosystem.

2.6 Modelling Vulnerability to Climate and Regime Changes

One of the goals of this study is to investigate the vulnerability of plankton to climate across space and species. It is known that the response of plankton to temporal variables is not uniform in space [22]. The linear regression model can be used to make predictions about the behaviour of the biological variables based on changes in the climate variables. This can be done directly with the principal components to estimate changes in joint species behaviour over space or using the principal components to estimate changes in individual species. In the unobserved components model each species abundance was given as a weighted sum of the principal components. The principal components are modelled as linear combinations of the climate trends and so this can be substituted to give estimates of the relationship

between the climate indices and individual species.

The linear regression equation for the principal components modelled as responses to climate and the model for the number of components can be substituted back in to the equation for the smoothed species abundances to give an estimate of how the species abundance is related to spatial and temporal climate signals. The advantage of regressing against the principal components rather than the species time courses directly is one of computational efficiency. Since a few principal components will explain the variation at each location this means that one a few regression coefficients need to be calculated at each location, rather than repeating for every species. If $\mu_i(t;l)$ is some mean signal an equation for the species abundances in terms of the physical variables can be written.

$$\begin{aligned} Y^{(p)}(t;l) \quad &= \textstyle\sum_{i=1}^{\hat{p}(l)} a_i^{(p)}(l)(\beta_{i,0} + \beta_{i,1}c_1(t;l) \\ &+ ... + \beta_{i,n}c_n(t;l) + \mu_i(t;l)) + \epsilon_p(t;l) \end{aligned} \tag{2.18}$$

The values of $\beta_{i,n}$ can be estimated using least squares and $\hat{p}(l)$ is an estimate of $p(l)$ given by equation 2.17. The above model can be used to estimate the behaviour of species p as the physical variables change. A species is regarded to be vulnerable to climate change if a large change in the physical variables results in a significant change in the behaviour of the species, i.e. the magnitudes of $\beta_{i,n}$ are large for principal components where $a_i^{(p)}$ is non-zero. The vulnerability of different regions to different climate variables also can be assessed by the magnitude of $\beta_{i,n}$. If the absolute value of the normalised $\beta_{i,n}$ is large then this indicates that a large change in the related physical variable will result in a large change in the joint behaviour of all species at that location, which is interpreted as the sensitivity of the assemblage associated with that component to the climate variable.

2.7 Modelling Trends and Oscillations

Principal component analysis is often seen to be model-free but this is not strictly speaking true. The principal components can be seen as either deterministic or stochastic processes. A deterministic process is one with an underlying function, whilst a stochastic process is a random signal generated by some probability distribution function. In the case of the CPR data the principal components are believed to be dependent on climate variables. The regime shift is well documented in the data but is unclear whether this is a gradual change in time or an abrupt stepwise change in the mean of the signal [17] and this can influence how the change is modelled.

Climate change might be modelled as a linear increase over time or as a polynomial [43] and different climate signals can be viewed as combinations of linear trends and oscillations. Using linear regression models can be fitted to the data to capture these oscillations [168] and to estimate their period. For long term trends a linear regression model can be fitted in the same way as previously described, using a hypothesis test to determine which terms should have non-zero coefficients. Frequencies of oscillations might also be estimated by fitting a sinusoidal model, as per the method of Rice et al [133], allowing one to estimate the period of each oscillation.

$$z(t) = A\sin(\omega t) + B\cos(\omega t) + \epsilon_t \tag{2.19}$$

where the noise terms ϵ are assumed to follow a Gaussian distribution with mean zero [133]. In this case simple linear regression can not be used to estimate the coefficients because the frequency, ω, must also be estimated from the data. A, B and the period ω can be found using least squares, i.e. by taking the product of $|z(t) - A\sin(\omega t) + B\cos(\omega t)|^2$ over all time, with minimising this function being equivalent to minimising the errors. The product is transformed to a summation by

taking the logarithm of the likelihood, since minimising the logarithm of a function is equivalent to minimising the function itself. Rice et al [133] show that the period ω can be found by minimising a function given by $S_n(A, B, \omega)$.

$$S_n(A, B, \omega) = \sum_t \left(z(t) - A\sin(\omega t) - B\cos(\omega t)\right)^2 \tag{2.20}$$

This model can be used to estimate the period of natural oscillations [133]. One reason that one may wish to model the climate trends and oscillations is for the purpose of making predications. The fitted values of the climate trends at future time points might be used in combination with the linear regression for the plankton signals as a response to climate trends to estimate how the plankton abundance might continue to change.

Chapter 3

Multidecadal Oscillations

3.1 Overview

Cannaby *et al* [43] explore the influence of natural oscillations on the sea surface temperature (SST) in the North Atlantic. Since these oscillations vary in their importance across space they use spatial principal components to break the sea surface temperature signal down into spatial and temporal representations. Cannaby *et al* begin by removing the climate warming trend from the SST data at each spatial location, which is assumed to be functionally dependent on atmospheric carbon dioxide. Weights are then computed as a function of space and signals as a function of time. The first three signals correspond to the AMO, the EAP and the NAO respectively. In the region of the North Atlantic studied by Cannaby *et al* [43] these three signals account for 23, 16 and 9% of the variance in the detrended data respectively. In this study one of the main aims is to explore the relationship between climate and the spatio-temporal structure of plankton abundance. Before studing the plankton dataset it is therefore useful to apply an exploratory analysis in order to understand the influence of different climate variables on sea surface temperature. In this chapter the spatial patterns of climate indices are studied in a similar way to Cannaby *et al* on both detrended and non-detrended data, with the region of interest being

extended east to include the North Sea and the Meditteranean. The inclusion of the North Sea is particularly important as this is where the plankton data is best sampled [160] and hence will be useful when we come to draw comparisons. However the result of this is some change in the principal components and their importance compared to Cannaby *et al*'s results. The goal of this analysis is that understanding the spatial variability of climate will enable a better understanding of the spatial variability of the plankton.

Since the linear trend can obscure any oscillations [141], it is important to investigate the effects of detrending or not detrending the data on the resulting spatio-temporal decomposition. In the first instance the sample median is subtracted from the time course at each location and then the time courses are re-standardised by dividing by the sample variance. The median is used rather than the mean in case the climate data is skewed (ie. there are a small number of extreme values recorded). This is done so that the first principal component will not just be a representation of the average [85]. The weights in space and time signals are then found. It is predicated that the first component will be dominated by climate warming trend, as this is the greatest source of variability in the sea surface temperature trend [130], with locations with a positive weight being those were the sea surface is warming and those with a negative effect being undergoing a cooling effect, as the local effects of climate change might vary across space. Although the sea surface is on average warming, it can behave different locally and understanding these local differences better is an important part of understanding the impact of climate change, due to the fact that climate is a dynamic system and so local responses can be highly non-linear [144]. A small change in the average temperature might for example have a much more significant effect at a local scale and so local changes in climate are not always easy to predict. Subsequent trends are expected to represent oscillations, although as the linear trend is so dominant the spatial patterns of these oscillations may be more difficult to identify. The analysis is then repeated on the detrended

data in order to isolate the effects of these natural oscillations without them being obscured by the climate warming trend [141]. The trend is removed by fitting a line to the data at each location using least squares to estimate the coefficients of the linear trend and then subtracting the fitted trend from the time course at each location. This linear trend is thought to approximate the recent warming trend that has been observed in the sea surface temperature over the past few decades and so after subtracting it what is thought will remain are shorter term oscillations. The same normalisation is also carried out on the detrended data. For the purposes of comparing with the plankton data the sea surface temperature is also analysed over a region restricted to the same spatial area and temporal period over which the plankton data will be analysed.

In order to show the inhomogeneity of the warming trend the average temperature change is calculated for each location. Figure 3.1 shows the difference between the average temperature up till 1960 and the temperature in 2009. For the most part the sea surface temperature has increased by between 1 and 1.5 degrees Celsius. An exception is around the subpolar gyre, a combination of four currents near to the coast of Greenland, where a cooling effect was observed over the past few decades. The reason for this cooling might relate to changes in currents, as the overturning circulation of water in the gyre has an important effect on local temperature [118]. Contrastingly the southern North Sea has been warming faster, perhaps due to the waters there being particularly shallow [161]. Similar patterns have been observed in previous studies. Beaugrand *et al* [18] show that the sea surface temperature from the 1960s to the 2000s increased more in the southern North Sea than in the rest of the North East Atlantic, whilst little change was observed in the subpolar gyre region. The Convention for the Protection of the marine Environment of the North East Atlantic (the OSPAR Convention') in their 2010 status report [115] also show that the southern North Sea is the region in the North East Atlantic where the most warming has occured over the past few decades.

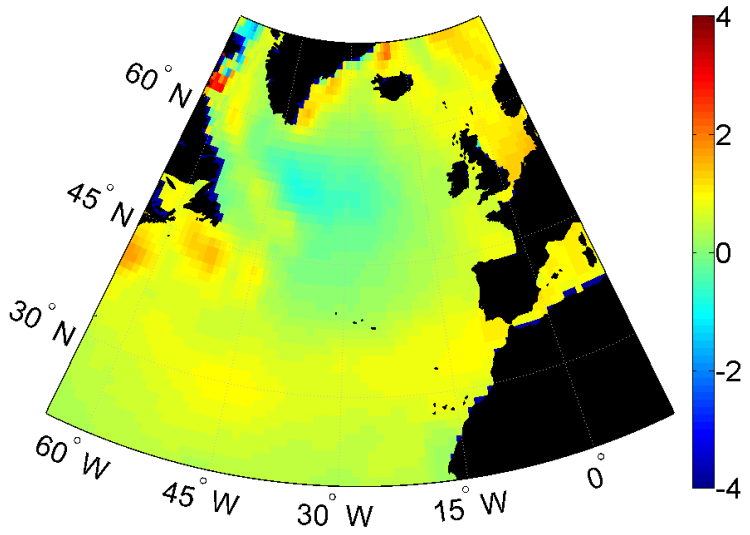


Figure 3.1: Plot of the change in sea surface temperature measured as the sea surface temperature in 2009 minus the average for 1890 till 1960 at each grid point.

3.2 Methods Used in this Chapter

In this chapter the main method used is spatial PCA, which is used to find the dominant spatial and temporal patterns in the sea surface temperature. Recall from section 2.2 that PCA is used to find dominant trends that represent the structure in a dataset by taking linear combinations of the variables and that spatial PCA, where weights are taken as functions of space and signals as functions of time, has been previously used to analyse both SST and plankton datasets. Equation 2.8 shows how the implicit assumptions are that the variables are linear combinations of common temporal trends weighted across spatial locations. In this chapter it is assumed that the sea surface temperature signal in time at a given location is given by a linear combination of underlying climate trends, which have a different influence on the local temperature at different locations. The Pearson's correlation coefficient, as described in section 2.5, is used as a measure of the relationship between the common component of the sea surface temperature and the different climate indices, as described in section 1.6. The correlation coefficient is then used

to identify the common components with known climate indices.

3.3 Climate Trends Across the Ocean Basin Scale

3.3.1 Sea Surface Temperature with Median only Removed

In the first instance the sea surface temperature is analysed including the effect of the warming trend. This will allow the spatial pattern of the warming trend to be explored. When the sea surface temperature, having only the median removed, is broken down in to its component trends, the most dominant component is the warming trend. The first component also contains some oscillatory behavior with a period similar to the AMO. In fact the AMO accounts for just over 30% of the variation in this trend. A consequence of this residual oscillation being present in the first component is that the AMO is not represented by a subsequent component due to the orthogonality constraint of PCA. The first component behaves very similarly to the NHT signal [150], which is a combination of a linear warming trend and an oscillation in time. On the first principal component the spatial pattern of the loadings is positive in most regions but is higher in the southern North Sea, in the south east of the North Atlantic and around the coast of America. The first component has much smaller loadings in the Subpolar Gyre, where the cooling effect was observed.

The weights on the second component represent the spatial pattern of the NAO, with a dipole in space [76]. They are positive in the North Sea and the south west and negative around the Subpolar Gyre. The time signal is a short term oscillation, ie. having a period of approximately ten years, appearing similar to the NAO. Subsequent trends are more difficult to identify with climate signals. The third component has positive weights around the gyre and near zero weights elsewhere. The time signal associated with this trend is oscillatory with a period of only a few years. The AMO is not identified as a separate trend when investigating the sea

surface temperature data with the median only removed because the NHT and the AMO are not orthogonal [150, 141] and PCA identifies orthogonal trends [85]. This means that the spatial pattern of the AMO can not be found without first detrending the data. The average warming trend is removed in this case by fitting a linear trend at each spatial location and subtracting this from the time course. This is referred to as linear detrending.

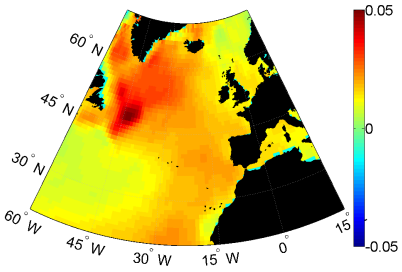
3.3.2 Detrended Sea Surface Temperature

The linear trend is probably one of the strongest sources of variability across the North Atlantic, as it is present across all regions. By removing this trend one expects there will be a change in the importance of subsequent components and their ordering. Geographical patterns of various climate signals also become more clearly defined when not obscured by the warming trend [141]. Figure 3.2 shows the breakdown of the sea surface temperature once the linear trend has been removed and table 3.1 shows the Pearson's correlation coefficients between the principal components and certain climate indices. The first component accounts for 34.63% of the variance, with the second and third accounting for 13.68% and 10.54% respectively.

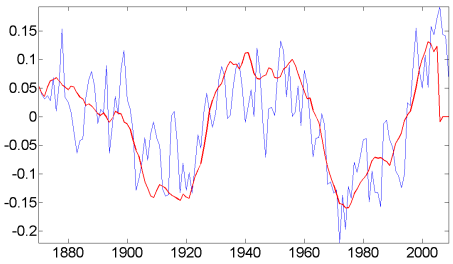
Component	AMO	NAO	EAP
1	0.7	/	/
2	/	0.336	/
3	/	0.5	/

Table 3.1: Table summarising the Pearson's correlation coefficients between the linearly detrended sea surface temperature data over the ocean basin scale and different climate indices.

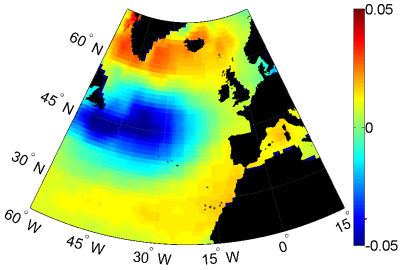
The AMO is present across all regions but its influence is centred upon the subpolar gyre, which it is hypothesised is because it is a result of current circulation in the



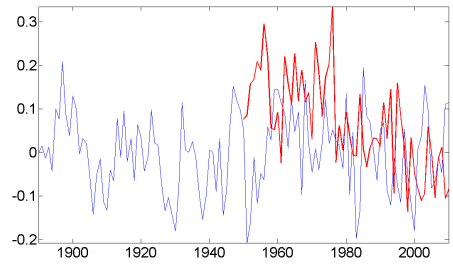
(a) Loadings in space for the first principal component.



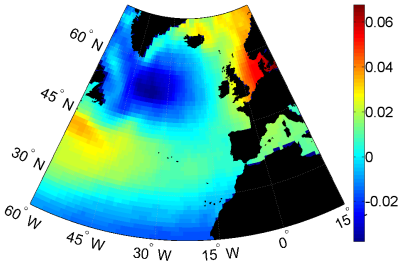
(b) The first principal component (blue) and the AMO (red).



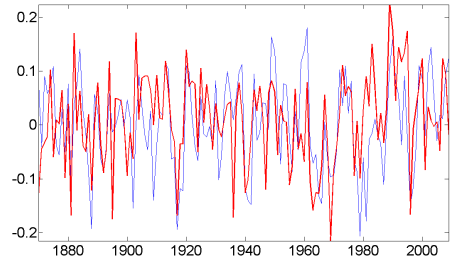
(c) Loadings in space for the second principal component.



(d) The second principal component (blue) and the EAP (red).



(e) Loadings in space for the third principal component.



(f) The third component (blue) and the NAO (red).

Figure 3.2: Loading vectors plotted in space and principal components for the linearly detrended sea surface temperature.

North Atlantic [141]. Spatially the first component is strongly positively weighted around the subpolar gyre. The weights are positive but much smaller about the rest of the North Atlantic, meaning the first component represents the average behavior of the detrended sea surface temperature.. The first component shows a strong oscillatory behavior, with a period of close to 60 years. This means that both in its spatial distribution and its temporal structure it resembles the AMO [141, 90, 43]. The correlation between the first PC and the AMO is also very strong, with the AMO accounting for over half the variation and a Pearson's correlation coefficient of over 0.7. Since the first component also accounts for a large proportion of the variation in the detrended sea surface temperature, it can be deduced that the AMO is an important driver of sea surface temperature. Thus it can be concluded that once the average warming trend has been removed the AMO becomes the most important pattern in sea surface temperature for the purpose of understanding the variability.

The second largest source of variability is not correlated with any known climate indices, meaning it may be some as yet unknown climate oscillation or it may an aggregation of different oscillations. Unlike the first component the second principal component is negatively weighted across some regions of space and positively weighted across others. This indicates that some regions are responding in an opposite way to others to this trend. It shows a sinusoidal behavior in space, with negative weights around the subpolar gyre and extending across in to part of the region off the coast of North America and positive weights elsewhere. The period of this spatial oscillation can be approximated as twice the distance from the maximum point to the minimum, which in this case is about 40 degrees in latitude, which is approximately 4450 kilometres. Although the time signal has only a weak correlation with the EAP, the weights do resemble its spatial pattern [165, 43, 8]. The time course is more noisy but also contains a shorter oscillatory trend. In the frequency domain the second PC has an oscillation with a period of about 18 years.

One possible cause of this approximately 18 year oscillation might be the lunar cycle, which was reported by Yndestad [170] to have an influence on the NAO, although it is difficult to verify whether this is what is driving the second component. The second component has a weak positive correlation with the NAO signal, with a Pearson's correlation coefficient of 0.336. There are different possibilities for what the second component may represent. It may be an aggregation of different trends, perhaps including the EAP, or might also consist in whole or part of the lunar cycle or it may be some as yet undefined signal. Since this trend contributes a large amount of variation to the sea surface temperature signal this is an area that warrants further investigation.

The North Atlantic Oscillation appears to also be an important influence on the sea surface temperature, although this is positive in some regions and negative in others. The third principal component has a dipole in space, with negative weights in the subpolar gyre. There is a region with positive weights in the southwest and the signal is very strongly positively weighted in the North Sea. This corresponds to the spatial pattern of the NAO [76]. The period of the oscillation in space is about 90 degrees in longitude at a height of about 52 degrees latitude, which is about 6170 kilometers accounting for the curvature of the Earth. The signal exhibits a short term oscillation and when regressed against the NAO signal has a positive correlation. The Pearson's correlation coefficient is close to 0.5, which is reasonably strong. The correlation with the NAO is stronger post-1965 than it is prior, with an explained variance of 22.12% and a Pearson's correlation coefficient of 0.4670 beforehand and an explained variance of 32.69% and a Pearson's correlation coefficient of 0.5705 afterwards. This can be seen from the plot (figure 3.2 (f)), as the two signals 'behave together' more frequently in the latter part. This could indicate that the third component is actually an aggregate of climate trends, with the NAO being the most important over later years. Spatial PCA produces weights that are fixed in time but in the case of this trend it might be appropriate to

have time varying weights, as the influence of the NAO on the SST signal appears to change over time. In the frequency domain the third PC has oscillations with a period of 8-10 years like the NAO [87, 76, 65] but also has a smaller amplitude oscillation of period 14 years, which is not present in the NAO signal. It is not clear what this oscillation represents and so this is another area that might warrant further investigation. The NAO signal has a residual oscillation with period similar to the AMO but because of the orthogonality of the principal components [85] this is not present in the third component, which means the correlation is not perfect. Summarising how much of the variation in sea surface temperature can be understood across different regions using these known climate indices it is possible to see that the AMO is an important effect in certain regions, whilst the understanding of other regions is greatly improved by account for the effect of the NAO. There are some coastal regions with complex local climate that are poorly explained by these indices alone. Figure 3.3 shows how much of the variation at each location can not be explained by the principal components as responses to climate signals. The first component of detrended sea surface temperature is modelled as a linear function of the AMO. The Euclidean distance between estimate of PC 1 given by the AMO in the model times the weight on PC 1 at a given location and the observed detrended time course at that location is calculated for each spatial location. This is normalised by the Euclidean norm of the observed values at that location. The Euclidean distance between two signals is the L_2 norm of one signal minus the other. The Euclidean norm of a time course is given by the square root of the sum of the value at each point squared, which is the L_2 norm of the signal. The estimate of the first component from the AMO is given by a linear regression equation, where the component is assumed to be proportional to the AMO plus a constant [168] (see equation 2.14). This equation can be estimated and the AMO substituted in to given the approximation of the principal component. This gives an estimate of the amount of variation explained by the AMO index across space. A dark red

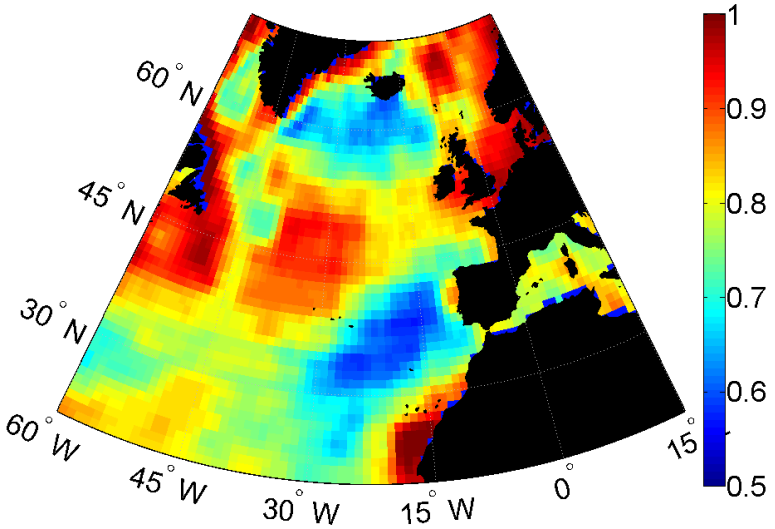
pixel shows that a large proportion of the variance is not explained by the AMO, whilst a blue pixel shows that a smaller proportion is unexplained. This shows that the AMO explains the least amount of variance in coastal regions [141], near the Labrador sea and the North Sea. In these regions there must be additional factors which influence the sea surface temperature. The third component appears to be a response to the NAO. This means that the variance explained jointly by the AMO and the NAO can be modelled. The biggest difference is that the North Sea is now much better explained [76], given how important the NAO is to this region. Regions which are well explained become larger. One region where the observed sea surface temperature remains poorly explained by these two climate signals is the region near the coast of North America. Since this is an area where there are a large number of small eddies and the ocean circulation here is complex [104], then the behavior of the sea surface temperature signal is harder to predict. Coastal regions also seem less well explained, in particular around the coast of Greenland, suggesting that local effects accounted for by higher components might influence the SST in this region.

3.4 Multidecadal Oscillations across the North East Atlantic

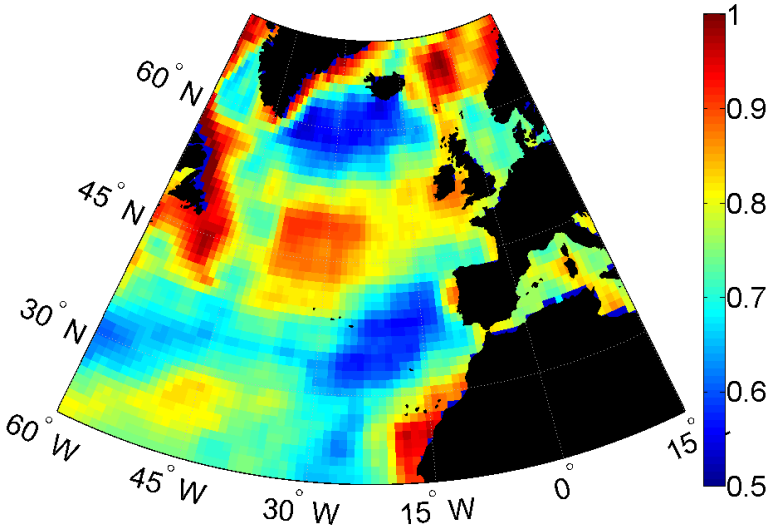
Component	NHT	AMO	NAO
1	0.5477	/	/
2	/	0.6054	/
3	/	/	0.5265

Table 3.2: Table summarising the Pearson's correlation coefficients between the principal components of the sea surface temperature with the median only removed over the CPR region and the climate indices.

When comparing with the biological data the sea surface temperature must be restricted to the same spatial and temporal region, however this will result in a different set of components because of the spatial variation in the importance of



(a) Plot of the amount of variation unexplained at each location based on the first detrended principal component modelled as a response to the AMO.

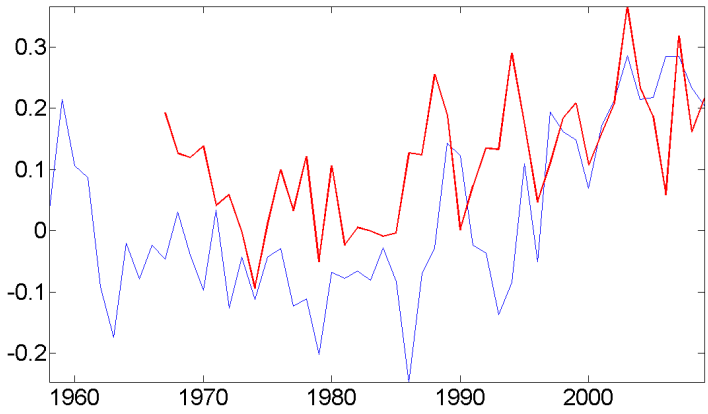


(b) Plot of the amount of variation unexplained at each location based on the first three detrended principal components modelled as a response to the AMO and the NAO.

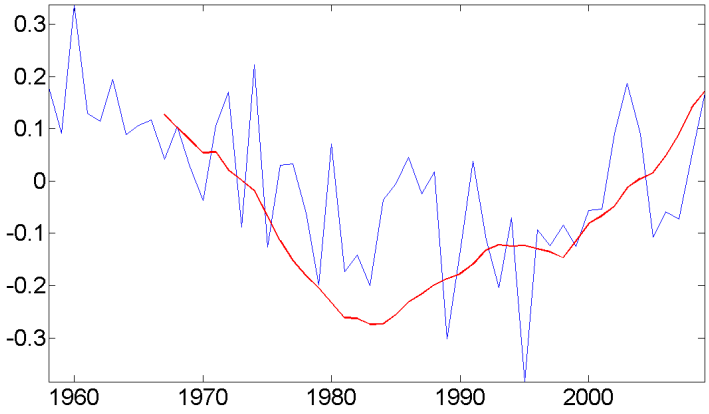
Figure 3.3: Plots of the amount of variation in sea surface temperature left unexplained by the principal components modelled as responses to climate indices culminatively.

different signals. Due to the irregular spatial sampling the best spatial coverage is achieved by restricting to the North East Atlantic [160] and temporally the most reliable data dates from about 1958. There are some changes in the components when the data is restricted to this region only, see figure 3.4. Pearson's correlation coefficients between the components and different climate indices are shown in table 3.2. Without detrending the first component resembles the general warming trend and is most strongly weighted in the southern North sea. The weights may be larger in the southern North Sea because of the fact that there are shallower waters in this region [123], which has meant temperature has changed more dramatically in this region compared with the rest of the North Atlantic (see figure 3.1). With a lag of 9 years, where 9 years is the lag at which the correlation coefficient is highest, the first component of the sea surface temperature over the CPR space-time region correlates with the NHT with a Pearson's correlation coefficient of 0.5477 and the lagged NHT explains 32.07% of the variation. This once more shows that the average warming trend is the most important driver of climate across a large scale [19].

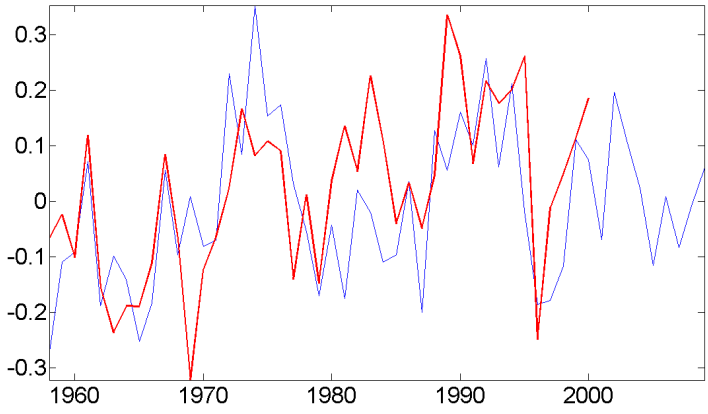
The second resembles the AMO, having strong weights in the north of the region. With a time lag of 9 years, which is again the lag where the correlation coefficient is maximised, the Pearson's correlation coefficient between PC 2 of the sea surface temperature and the AMO signal is 0.6054 and the lagged AMO explains 36.71% of the variation. The NAO signal contains an oscillation of similar period to the AMO with a time lag, suggesting the AMO has some interaction with the NAO. The second component might also represent this component of the NAO. Since the lags are the same for both the NHT against the first component over this region and the AMO and the second component, it appears that the data in the north east of the North Atlantic responds later than the average signal over the northern hemisphere. Across the ocean basin scale the AMO was obscured by the average warming trend [141] but across the North East Atlantic this is not the case. This



(a) Plot of the first principal component on the sea surface temperature with only the median removed (blue) and the NHT signal lagged by 9 years (red).



(b) Plot of the second principal component on the sea surface temperature with only the median removed (blue) and the AMO signal lagged by 9 years (red).



(c) Plot of the third principal component on the sea surface temperature with only the median removed (blue) and the NAO signal with no time lag (red).

Figure 3.4: Plots of the first three principal components on the sea surface temperature over the CPR region.

may be because the AMO is less of a dominant influence on the SST in this region (see figure 3.2) and so does not influence the first component to the same extent. Alternatively this may be a feature of restricting the data to a shorter time period. Since the AMO appears as the second component it is clear that although it is less influencial than in other regions it still plays a role on climate over this region.

The third is again positively weighted in the North Sea, particularly the south, and the temporal signal is similar to the NAO. The NAO without any time lag accounts for 28.03% of the variation in the third PC and the Pearson's correlation coefficient is 0.5265. The North Sea is the region in which the NAO signal has the strongest influence, which can be seen clearly from the analysis of the entire North Atlantic (see figure 3.2). Previous studies have also found the NAO to be influencial in the North Sea [65].

3.5 Discussion

In this section it has been demonstrated that both natural oscillations in pressure centres and the overall trend of climate warming are associated with the sea surface temperature. The Atlantic Multidecadal Oscillation and the North Atlantic Oscillation were identified as significant sources of variability in the sea surface temperature. Most importantly it is apparent that these trends vary in their influence spatially. Whilst the AMO is positively weighted across all regions, albeit some more strongly than others, the NAO has positive weights in some regions and negative in others, meaning it is an oscillation in both time and space. This implies that local variation must be taken in to consideration in any study of the impact of climate on the ecosystem and that regional climate patterns can vary in ways that are non-intuitive compared to average trends. This is seen in the average warming trend, for example, where an increase in sea surface temperature is seen on average but certain regions are undergoing a cooling effect. In the following chapters,

where the focus shifts to the ecology, it remains necessary to consider this spatial heterogeneity, as this will increase the understanding of how the plankton distribution changes over space. The effect of the average warming trend is particularly strongly represented in the North Sea, especially the southern part, which is the region where much of the plankton data is collected. It can also be seen that detrending or not detrending influences the ordering of principal components and that the principal components on the sea surface temperature can change depending on the spatial region that is under consideration.

Chapter 4

The Interpolated Data Modelled using Sparse PCA

4.1 Overview

In this section the sparse Species PCA model is used on the pre-processed WinCPR dataset, which contains data for the North Sea across 110 different species. This will demonstrate how well the analysis performs at identifying spatio-temporal structure. A number of previous studies have used Spatial PCA on the WinCPR dataset with various indicator species, including those by Beaugrand *et al* [25, 160]. Species PCA adds new insights to previous work because the diversity across space, ecoregions defined by species and joint functional responses can all be studied together. In this chapter both zooplankton and phytoplankton are studied together. Those variables which are measured differently, such as Phytoplankton Colour Index, are not included. The WinCPR dataset also contains a number of aggregations of species, such as total copepod abundance, which are also omitted as this may produce misleading results. The data covers the time period from 1958 till 2001 and the North Sea is gridded in to 183 locations. Both yearly averages and monthly data are explored in this section. The yearly trends are then compared with possible

climate 'drivers' across space in order to show that there is spatial heterogeneity in responses to climate. The monthly time courses are analysed in the frequency domain, as this can be used to show that seasonal patterns dominate the data at this scale.

4.2 Methods used in this Chapter

In this section the pre-processed WinCPR data is used in order to assess the applicability of the methodology to understanding the spatial temporal behaviour of plankton communities across assemblages. Recall that this dataset is a gridded database of log-abundances of over 100 species of plankton that has been interpolated using the inverse distance method as described in section 2.1. The CPR data is irregularly sampled in space and so the inverse distance interpolation method is used to transfer the data to a regularly spaced grid. The WinCPR dataset is described in more detail in section 1.2.

In order to find the joint behaviour of species assemblages sparse species principal component analysis is used on both monthly and annual data (see section 2.2.3 and section 2.3 for a full description of these methods). This method finds dominant joint behaviours of assemblages for keystone species across space by estimating a sparse weight vector. The output of this analysis can be interpreted as finding the dominant species across space and their joint responses over time. It was proposed that in order to estimate the number of species that should have non-zero weights on each component a mixture model might be used (see section 2.3.1) and the WinCPR dataset can be used to investigate this model. K-means clustering is then used to define ecoregions, which can be based on either the weight vectors, representing regions defined by species assemblages, or on the principal components, representing regions defined by functional behaviours (section 2.4). Finally linear regression and the Pearson's correlation coefficient (section 2.5) are used to analyse the re-

lationship between the principal components and different climate indices, which will be informative as to which climate indices might be responsible for driving the joint behaviour of species assemblages. Through investigating the relationship with the climate indices it is also shown that taking in to account time lags is important (section 2.3.2). In chapter 2 it was discussed how small time lags that might be present in the CPR dataset could be accounted for using Fourier transforms. In this chapter it is shown that taking in to account these time lags is important with regard to investigating the response to climate indices, suggesting that not all species respond to climate forcing at the same rate. Furthermore Fourier transforms can also be used to estimate the frequency of oscillations, which can be seen in particular in the monthly data as it is dominated by seasonal cycles.

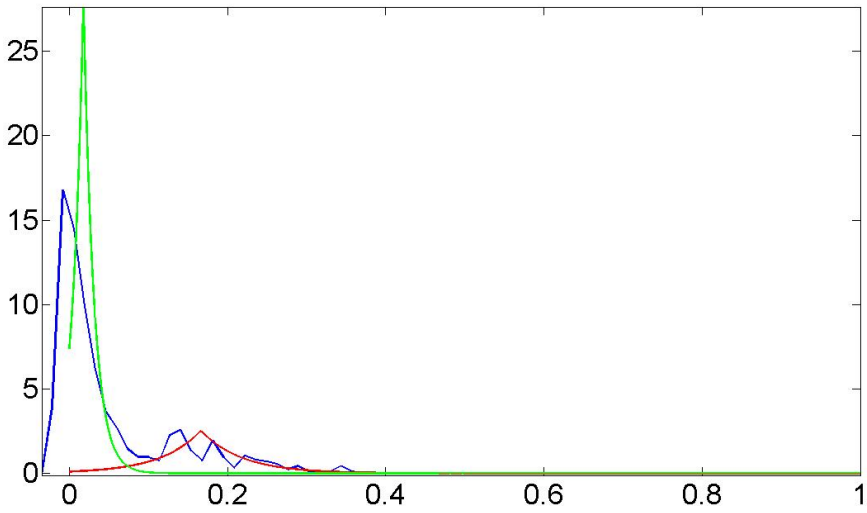
4.3 Sparse Principal Component Analysis on the WinCPR Data

4.3.1 Verifying the Mixture Model using the WinCPR Data

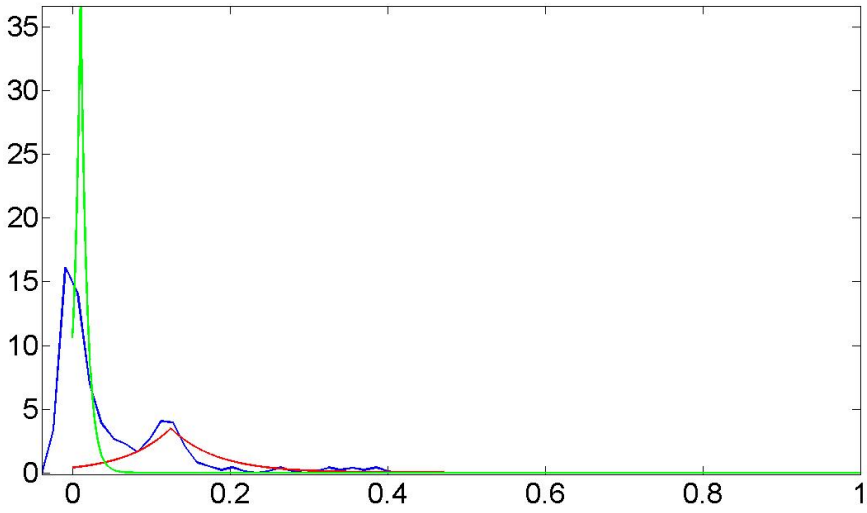
A model is placed on the absolute values of the loading vectors, comprising of a mixture of two Laplace distributions [109, 168]. Figure 4.1 shows the Kernel density estimates of the absolute values along with the density estimated by the mixture model (see equations 2.11 and 2.12) for two locations. The plots show that the Laplace density is needed to capture the distribution of the noise near zero. These results suggest that the mixture model is appropriate, thus will give good estimates of the sparsity parameter. It also aids in selecting the right choice of distribution.

4.3.2 Measures of Diversity

Figure 4.2 shows the sparsity parameter, i.e. the number of species with non-zero weight divided by the total number of species, for the first component in space as



a) Location 2.



b) Location 82.

Figure 4.1: Plot of the density of the absolute values of the loadings along with the densities estimate from the mixture model. The kernel density estimates of the true signal are shown in blue, the estimated probability density for the noise in green and the estimated probability density for the signal in red.

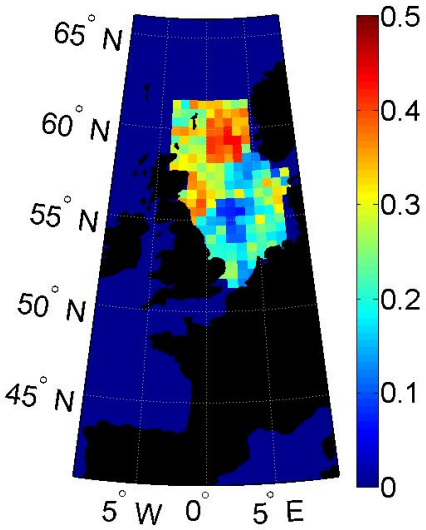
estimated by the mixture model and the number of principal components. For the first principal component the solution is less sparse in the north North Sea than in the south, indication a larger group size. One possible explanation for this is the influence of oceanic species from the rest of the North Atlantic, which have a tendency to show more variability and thus must be retained in order to describe most of the variance. Hence more species must be included in this mixing section of the North Sea. Previous studies have noted the importance of mixing regions as influences on plankton diversity [84]. For instance the Celtic shelf, which is the region of the ocean shelf south of Ireland, is more diverse because of the increased mixing of nutrients with the surface waters due to the currents in this region. Part of the ocean shelf also runs north of Scotland [123], suggesting this region too may be defined as a mixing region. For higher principal components the pattern begins to change, although there is something of a North-South division. In the second component there are slightly more species retained in the north of the region. In the third principal component the highest sparsity parameter is found at the southern most tip of Scandinavia. Possibly the third component is representative of species that are found more frequently in this region. In the fourth component there is slightly more of an east-west division in the sparsity pattern. Around the coast of Norway a large number of the sparsity parameters are set to zero because only three components are required to explain most of the variation in this region. Near the coast of the United Kingdom the sparsity parameter is very slightly higher. This may mean this component is largely represented by coastal species, although higher components may also be comprised of a lot of noise.

The number of principal components, which can been seen as a descriptor of the number of different functional groups, is higher in both the south and the north west, whilst it is markedly lower in the north east. This could be due in part to the higher diversity in the south, as more species may be able to survive in the shallower and more temperate environment, and again the influence of oceanic species in the

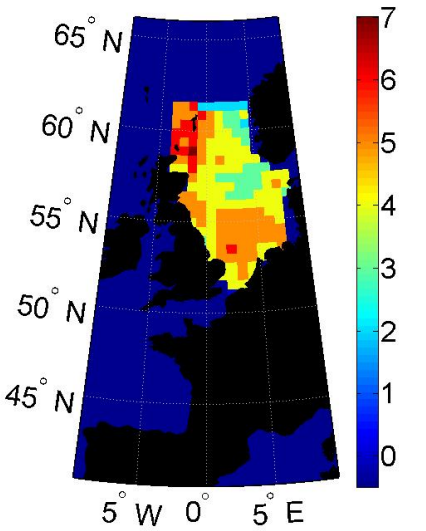
north east. What is certainly clear is that there is a definite spatial structure in both the group size and the number of groupings. Spatial heterogeneity in the number of components and the sparsity parameter suggest spatial variability in the diversity of the plankton ecosystem. Since diversity can be seen as a measure of the stability of an ecosystem [20] this implies that some regions will be more vulnerable than other to changes in the physical variables. As was shown when investigating the spatio-temporal structure of the SST the influence of the warming trend is heterogeneous across space [67] (see figure 3.1) and as many species of plankton are influenced by the climate [9] this appears to lead to spatial heterogeneity in the diversity of plankton species. The diversity is influenced by physical features, such as the bathymetry, as well as climate variables, with the ocean shelf region being important as a mixing region.

4.3.3 Ecoregions Defined by the Loading Vectors

In figure 4.3 K-means clustering has been carried out on the first two loading vectors. The cluster analysis was also carried out on the third and fourth loading vector but these are not shown. There is clear regionalisation in space, with the northern, central and southern North Sea being clearly defined. From a biological perspective clustering on the loading vectors is akin to finding regionalisation based on species groupings. On the first loading vector cluster one represents the coastal regions around Scotland and the northern part of the continent. Dominant species there include a number of diatoms: *Paralia sulcata*, a species found both in the water column and in the sediment [64]; *Chaetoceros [phaeoceros] spp.* and *Thalassiosira nitzschiodes*. Diatoms are the most abundant of the phytoplankton [68, 111] and so it is to be expected that they would have large weights. The dinoflagellates *Ceratium fusus* and *Ceratium macroceras* are also present in this region. Hardy [68] suggests in coastal regions there is a greater mixing of bottom-dwelling diatoms with the plankton, which may explain the prevalence of *P. sulcata* in this region. *P. sulcata*



a) Sparsity parameter for principal component 1.



b) Number of principal components.

Figure 4.2: Plot of the sparsity parameter for the first four Fourier transformed principal components across space. Here a red pixel depicts a large value of π (so a very non-sparse solution) and a blue pixel depicts a small value. The limits are set between 0 and 0.5. The number of principal components across space is also shown, with red indicating a higher number and blue a lower value.

tends to have weights with positive real parts across coastal regions on the first loading vector and zero weights elsewhere, suggesting it shows most variability in these regions, which agrees with pre-existing knowledge that this species may respond to water column mixing. *Thalassiosira nitzschiodes* has positive weights over most of the southern and central North Sea but these are especially high around the coast of Scotland. This suggests that it is better adapted to warmer and shallower waters. *Ceratium fusus* has positive weights over the entire North Sea, which suggests it shows strong variability across all regions. For this species the physical features are less of an influence on its spatial distribution and this is shown by the fact that it has significant weights across all three clusters. The cluster analysis shows more diatom species occurring in coastal regions. McQuatters-Gollop *et al* [111] study spatial patterns of Diatom and Dinoflagellate abundance across the North East Atlantic. Spring blooms of Diatoms occur earlier in the year than for dinoflagellates and begin in shallower waters, due to the increased penetration of light in these regions, before spreading out to the rest of the region. In this analysis yearly averages have been taken, which means that the earlier occurrence of Diatom blooms in shallower waters might explain why there are more Diatom species strongly weighted on average in these coastal clusters. McQuatters-Gollop *et al* [111] note that the Shetland and Orkney Island regions of Scotland and German Bight, which are regions covered by cluster one, are particularly productive areas for phytoplankton. Diatom abundance is particularly high in the German Bight in August. Whilst most species of Diatom live pelagically, some are known to exist as surface films in benthic communities. Diatoms are thought to be able to survive in most climates and so are not vulnerable to changes in temperature [56]. By contrast they are affected by currents and wind speeds, which may give some explanation as to why they are slightly more common in the coastal clusters. In particular the mixing of nutrients with the water column in these regions in thought to positive influence Diatom abundance [68, 64], as well as increased mixing of sediments with the water column

causing benthic species of Diatom to be found amongst the plankton [68].

The second cluster occupies the northern North sea and is characterised by dinoflagellates: *Ceratium fusus*, *Ceratium furca* and *Ceratium tripos*. This suggests a higher degree of variability of dinoflagellates in this region. A study by Leterme *et al* [99] suggested that there was an increasing contribution from dinoflagellate species to the phytoplankton colour index in the northern North Sea, although they stress that there is some discrepancy between cell counts and colour index meaning that it may not directly reflect the abundance. They do, however, suggest that dinoflagellate abundance is increasing in the northern North Sea along with phytoplankton colour index. Whilst in their study Diatom abundance did not show a significant global trend in the northern North Sea, its contribution to the phytoplankton colour index has decreased. These results suggest that whilst Diatom abundance is relatively stable in this region, dinoflagellate abundance is changing. From this it can be concluded that dinoflagellate species are on average strongly weighted in the northern North Sea because of the increased variability in their abundance in this region, making it an important region for change. This can be seen from the average temporal signal in the northern North Sea (see Figure 4.4), which is increasing.

The third cluster, which is represents the central region is also dominated by dinoflagellates, in particular: *Ceratium macroceras*, *Ceratium fusus* and *Ceratium furca*. The study by Leterme *et al* [99] claims an increase in phytoplankton colour index in the central North Sea but no significant change in either the Diatoms or the dinoflagellates. They suggest that the ecosystem is relatively stable in this region. This implies that the first principal component is predominantly made up of phytoplankton species across all spatial regions. An explanation for why the phytoplankton dominates most of the variability could be the difference between count and biomass [68]. The CPR survey contains raw abundance data, which is not converted in to biomass, except for the phytoplankton colour index [160]. Phytoplankton are smaller in size than the zooplankton but occur in larger numbers

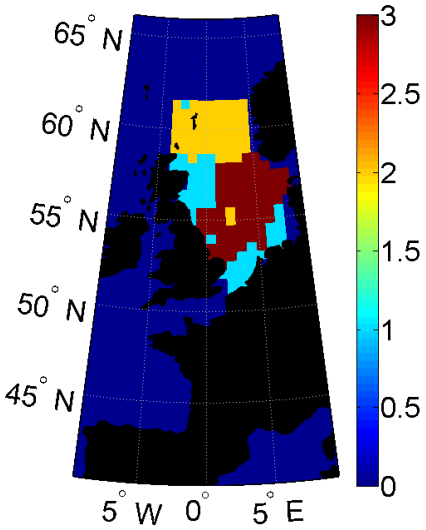
[68].

The clusters on the second loading vector are less clearly defined, however they are roughly separated in to the northern, central and southern coastal regions. The first cluster mostly covers the central region. The most dominant species for the second loading vector in this cluster are the dinoflagellates *Ceratium furca*, *Ceratium fusus* and *Ceratium tripos*. Leterme *et al* [99] do not find there to be significant changes in the Diatom or Dinoflagellate communities in the central North Sea, although there is an increase in phytoplankton colour index. The average time course for the second component is oscillatory in time and so this may represent decadal trends rather than long term changes. Cluster two, which covers the southern coastal region, is dominated by the dinoflagellate *Ceratium furca* and two species of copepod: *Para-Pseudocalanus* and *Temora longicornis*. The latter species is known to be prevalent in the southern North Sea and dominates the copepods in this region [68], so its importance in this cluster corresponds with pre-existing ecological knowledge. This supports the hypothesis that the weight vectors are representative of important species across space. Dinoflagellates are less important than the Diatoms in the southern North Sea [99] but *Ceratium furca* does have a strong weight on the second loading vector suggesting that it is an important phytoplankton species, albeit less so than the Diatom species, in this region. *Pseudocalanus* is a small species of Copepod which shows a high degree of variability of different years [68]. McGinty *et al* [108] determine that *Para-Pseudocalanus* is adversely affected by increasing temperatures and so the strong weight on this species in this region might be associated with a decline. The third cluster, which mostly covers the very north of the region, the region where the most mixing with oceanic waters occurs, is dominated by the dinoflagellate *Ceratium macroceras* and the diatom *Chaetoceros [Hyalochaete]*. Hardy [68] claims that where two waters mix, such as where coastal waters meet oceanic waters, there is an increase in the reproductive rates of species due to the interaction of the two systems replenishing constituents

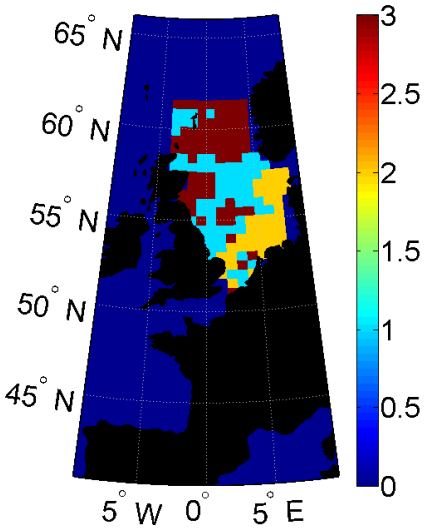
that are deficient in either system due to slightly different plankton compositions. In the northern North Sea the shallow waters of the North Sea mix with waters from the rest of the North Atlantic and so it might be defined as a mixing region.

The third loading vector is dominated by two clusters, with cluster three being very small. The first and the third cluster occupy the southern North Sea. In terms of dominant species the diatom species *Thalassiosira nitzshiodes* and *Odontella sinensis* both are strongly weighted in cluster one. In cluster two, which occupies the northern part a number of diatoms and dinoflagellates are strongly weighted including: *Ceratium furca*, *Ceratium fusus*, *Chaetoceros [Hyalochaete]* and *Psuedonitzshia seriata*. In the third cluster dinoflagellates *Ceratium furca*, *Ceratium fusus* and *Ceratium horridum* are the most strongly weighted species.

The clusters on the fourth loading vector have a similar spatial structure to those on the third, although the southern region is decreasing in size which suggests more spatial homogeneity in species distribution at this level. The first cluster occupies most of the region, whilst the second and third occupy the continental coastal area in the south of the north sea. Cluster one is dominated by a number of diatom species and the dinoflagellate *Ceratium macroceros*. Clusters two and three are mainly dominate by dinoflagellates in the *Ceratium* genus but a couple of copepod species, namely *Acartia spp.* and *Para-Pseudocalanus* do also have a presence in cluster two. The second and third loading vector show less spatial structure than the first two. There is a rapid decline in the amount of variation explained by the principal components, with the first two components explaining more than 50% of the variation at most locations. This may imply that most of the structure has been subtracted by the third component, meaning that higher components are in general less structured or are selecting only single species.



a) Clusters on the first loading vector, where cluster one is blue.



a) Clusters on the second loading vector.

Figure 4.3: Plot of the clusters on the absolute value of the Fourier transformed loading vector. On the spatial plots of the clusters: cluster 1 is shown in light blue, cluster 2 in orange and cluster 3 in dark red.

4.3.4 Regions Defined by Functional Behaviour

Figure 4.4 shows the result of a similar cluster analysis performed on the first principal component, which is equivalent to defining the regionalisation on the joint behaviour of species. These show spatial structure but are not completely identical to the clusters on the weights. The clusters on the first principal component show a clear division of the North Sea in to north, central and southern regions. The average trends over the first and third clusters, which represent the southern and central regions show a decline over time. The north by contrast is showing a slight increase. This may be due to an increase in warm water species, which with the increase in sea surface temperature have become more able to survive further north. This cluster might represent a northwards migration of certain species [12, 5]. Recalling from the analysis of the loading vectors the first component is defined primarily by phytoplankton species. One problem that may arise is that phytoplankton and zooplankton have been aggregated together, even though they are measured in different ways. This could be problematic for those species that are very variable in time, as a small weight could still relate to a large contribution to the component [172]. The first component seems to represent the general warming trend but not all species of phytoplankton are primarily driven by changes in temperature [56], which suggests the weights may in fact be misleading. Since different types of species are measured in different ways and have different biomasses [99] which leads to the observations being recorded on different scales, this implies that it might be advisable to consider different subgroups separately. This issue will be addressed when investigating the raw data.

It is important to be aware of the sign of the principal components, as species can have positive or negative weights. If a trend is increasing but most variables have a negative weight then the trend in fact represents a decline. For the most part the most strongly weighted species on the first principal component have a positive real part or a weight which is zero because of the sparsity constraint. This means

that there is a general decline of most species, except in the north, where there is a general increase. The cold water copepod *Calanus finmarchicus* is one of the few species to have negative weights on average. The real part of its weight is zero in most regions, except for the north North Sea, where it tends to be negative. This suggests that this species is declining especially in the northern North Sea. *C. finmarchicus* is known to be sensitive to changes in temperature [117]. In general it seems that the most important source of variability in joint plankton abundance across the North Sea is the climate warming trend. Another species that has negative weights in the north is *Euphausiacea*, which are commonly known as Krill. These are important species because they provide a food source to both fish and whales [68]. Letessier *et al* [100] find a negative relationship between *Euphausiid* abundance and SST, which implies that the negative weights in this region are also a response to climate warming.

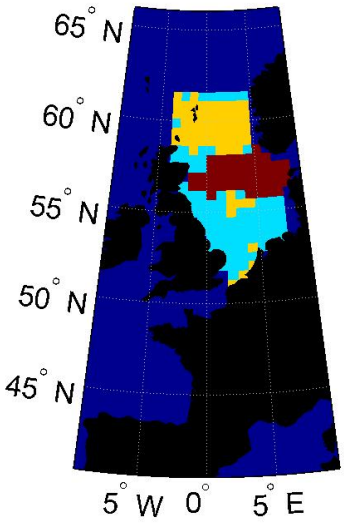
Figure 4.5 shows the regionalisation based on the second time component. The clusters on the second component are less well defined, although there is some indication of an east-west distinction. Cluster one occupies the British coastline, cluster two the south east and cluster three the north east. For the most part the second component seems to be an oscillation, perhaps a response to the AMO. This oscillation is most clearly seen in cluster one, the coastal cluster, perhaps indicating a relationship between the AMO and the plankton being stronger along the coast line. The AMO is known to influence wind intensities [141], which in turn influence mixed layer depth, which is a driver of the abundance of certain phytoplankton species [56, 64]. The time courses for clusters two and three appear like time-lagged versions of the oscillation in cluster one, with the minimum occurring earlier in cluster two and later in cluster three. The AMO is a secondary driver of plankton abundance, being particularly important for Diatom species [56], and from the ordering of the principal components it can be concluded that it is the second most important driver of plankton abundance across all species in the North Sea. On the third and

fourth principal components the spatial structure is even less clearly defined, so two clusters are sufficient. Since components one and two explain most of the variation across most locations in the data it may be that there is little spatial structure in the higher order components.

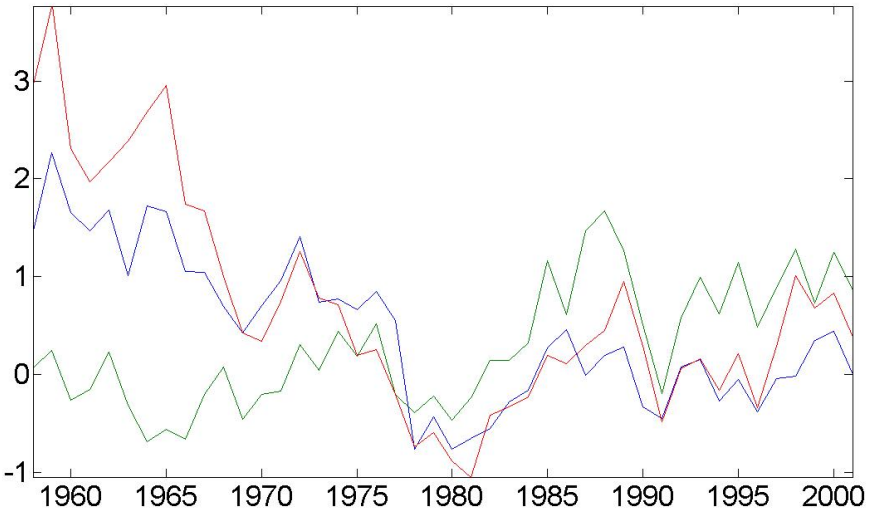
4.3.5 Time Delays Across Space

Fourier PCA accounts for small time delays between the individual species and the average signal (see equation 2.14). Figure 4.6 shows the time delay in years for two different species across time for the first PC. A positive time delay indicates that a given variable is ahead of the average signal at a particular location and a negative delay indicates that it lags behind. *Ceratium fusus*, a diatom species, is in general lagged behind the average signal. There is an east west divide in the time lag, with it responding earlier than the average signal in the west at certain locations. The regions that it responds ahead tend to be shallower coastal locations, where mixing occurs [68]. *Para-Pseudocalanus* is a small species of copepod and can be seen to be responding ahead of the average signal in the north of the region. Further south it responds behind the average signal. It is thought to be adversely affected by rising temperatures [108] and so may be moving to higher latitudes, causing the change in abundance in the north North Sea to occur more rapidly than the average signal.

The time delays can also be taken as a function of space for each species and clustered as in figures 4.7. The first cluster contains copepod species *Calanus finmarchicus* and total abundance of Calanus copepods. These species seem to be responding ahead of the average signal in the north region, where mixing with oceanic waters occurs. *C. finmarchicus* is a cold water copepod species [117], which is thought to be moving northwards as sea surface temperature changes [74]. It is particular sensitive to temperature changes in the North Sea [75], which may explain why it responds ahead of the average signal. Cluster two contains species which have a faster than average response in the north and respond behind the average

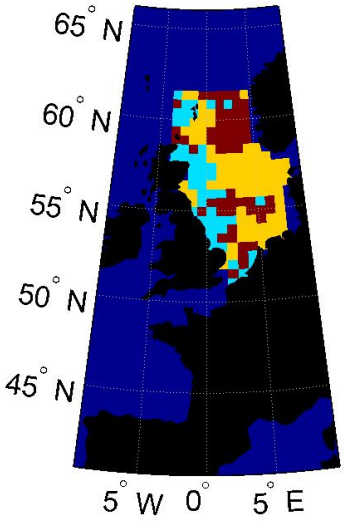


a) Clusters on the first time course, where cluster one is blue, cluster two is orange and cluster three is red.

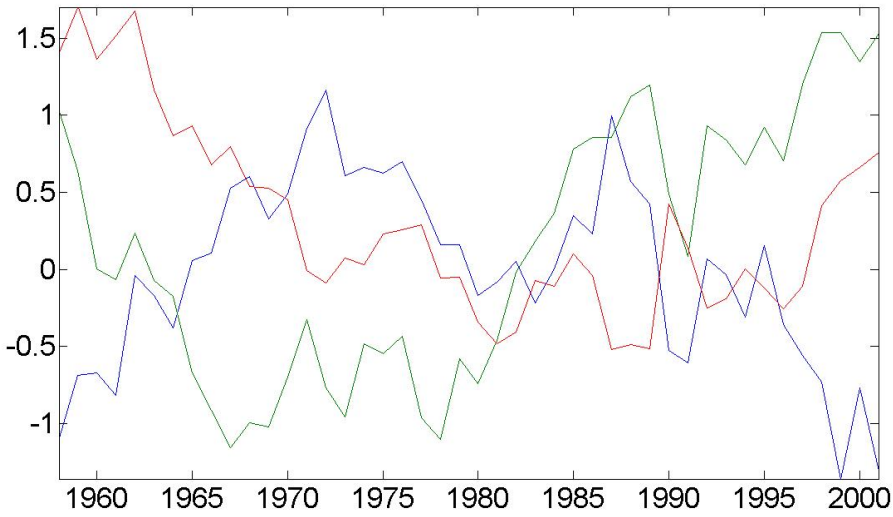


b) Centres for the first time course.

Figure 4.4: Plot of the clusters on the first principal component for the Fourier transformed data. In the plots of the averages the centre of cluster one is shown in blue, cluster two in green and cluster three in red.



a) Clusters on the second time course, where cluster one is blue, cluster two is orange and cluster three is red.



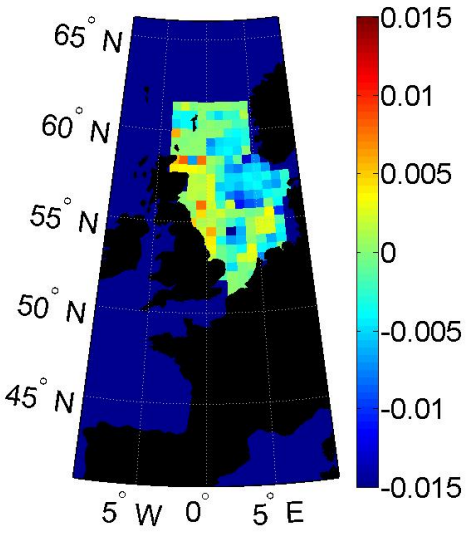
b) Centres for the second time course.

Figure 4.5: Plot of the clusters on the second principal component for the Fourier transformed data. In the plots of the averages the centre of cluster one is shown in blue, cluster two in green and cluster three in red.

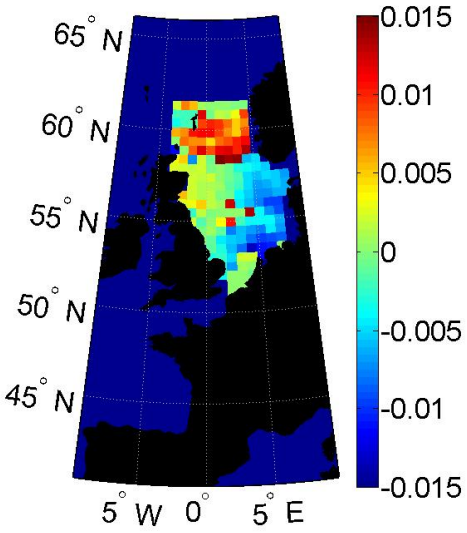
signal in the south, particularly in the German Bight. This group contains a mixture of dinoflagellates and a few diatom species. Dinoflagellates are thought to be increasing in the northern North Sea and slightly decreasing in the south [99]. The time delays show that the increase is occurring faster than the average signal, whilst the decline is occurring more slowly, suggesting that changes in the behaviour are not uniform in space. Where the dinoflagellates are responding behind the average signal it may be possible to use the average signal to predict their future behaviour. Cluster three is largely uninteresting as the species have close to no time delay, which means it may consist of those with mostly zero weight under the sparsity constraint.

4.3.6 Relationship Between the Plankton and Climate

Figures 4.8, 4.9 and 4.10 show Pearson's correlation coefficient between the principal component and the Atlantic Multidecadal Oscillation, North Atlantic Oscillation index and Northern Hemisphere Temperature across space respectively. Even when the false discovery rate is controlled the correlation is significant at a large number of locations for all three climate measures. The correlation with the AMO is statistically significant for the first component at 131 of the 183 locations and on the second component at 136 of the 183 locations. This indicates a strong influence of this trend on behaviour particularly for the second component. The Pearson's correlation coefficient is positive for the first component across most regions, except for the very north. The coefficient lies in the range -0.6827 to 0.8036. On the second component the correlation with the AMO is positive across most regions, being particularly strong in the south coast and negative across the coast of the British Isles. The difference in influence between coastal and non-coastal waters might be explained by the relationship between the AMO and currents [141], as it is the influence on the circulation that is important to the plankton [56]. The correlation for the second component is even higher, lying in the range -0.8555 to

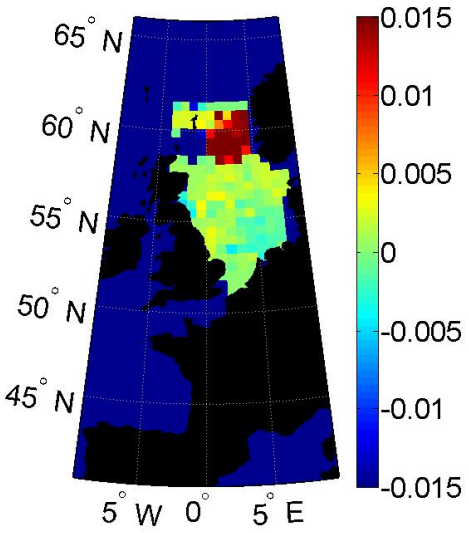


a) Time delay for Ceratium fuscus.

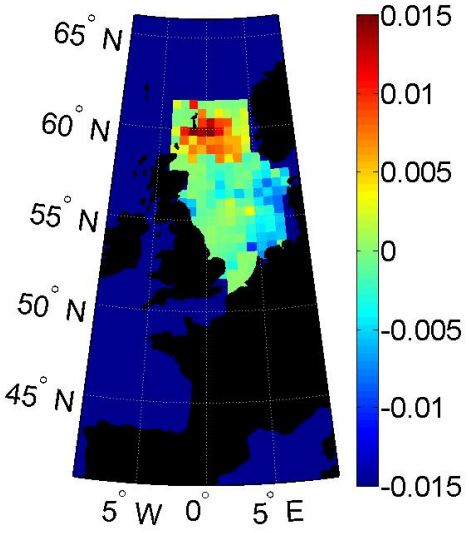


b) Time delay for Para-Pseudocalanus spp.

Figure 4.6: Time delays in for the first principal component for two frequently non-zero species



a) First cluster.



b) Second cluster.

Figure 4.7: Centres of the clusters for time delays as a function in space clustered on species and the mean squared error for the distance from the centre of the cluster plotted with species on the x-axis for principal component1.

0.8814, which indicates a very strong relationship in some locations. The correlation can be negative or positive due to the fact that there are a mixture of types of species, some of which will have a positive response and others having a negative response to climate effects. The sign of the loadings is chosen so as ensure the mean is positive, so that where the correlation is negative this indicates that there is more weight given to those species that are responding negatively to the climate index. In these regions, where the correlation is negative, species that respond positively will have a negative weight. As previously discussed the AMO is thought to be influential on the abundance of certain phytoplankton species [56], perhaps in some cases even more so than temperature, and the first and second component are mainly comprised of phytoplankton species.

The first principal component has a positive correlation with sea surface temperature in the north. Since the SST is a spatio-temporal signal the Pearson's correlation coefficient can be calculated for local sea surface temperature at each location. This relationship changes for higher principal components, with the second principal component showing a more positive relationship in the south east. This correlation is significant for a large number of locations and is positive in the east. Since the first principal component accounts for the largest proportion of the variance this could be an indicator of how different functional groups with different joint behaviours dominate in different regions and how consequently these groupings have different responses to climate change. for the first component the correlation with the NAO is highest in the north North Sea, whilst for the second principal component it is highest in the central North Sea. The NHT has the strongest correlation with the first principal component in the northern North Sea and the second principal component has a positive correlation with the NHT in the central and southern North Sea. The correlation with the NHT is positive where the average signal is increasing over time and negative where it is decreasing, as the NHT signal shows an increase across time. This correlation is significant at 88 of the 183 locations and the

Pearson's correlation coefficient lies between -0.5903 and 0.5582. For the second principal component the correlation is less significant. From this it can be deduced that the first component is comprised of a mixture of responses to the AMO and the NHT and the second component is primarily a response to the AMO. Temperature is known to be a major driver of the abundance of many zooplankton species [12, 117, 21] and may influence certain phytoplankton too [155, 99] and so this explains why the NHT signal is influential in driving the first component. The NAO has a far weaker correlation with the first two components, suggesting it is a driver of lesser importance. There is still some spatial structure in its relationship with the first component, with the Pearson's correlation coefficient taking more positive values in the north North Sea. This is a strong indication that the joint behaviour of the species is influenced by long term changes in climate over time. The plots show that there is spatial variation in how the signal responds to climate, which supports the hypothesis that there is spatial heterogeneity in plankton responses to climate [108].

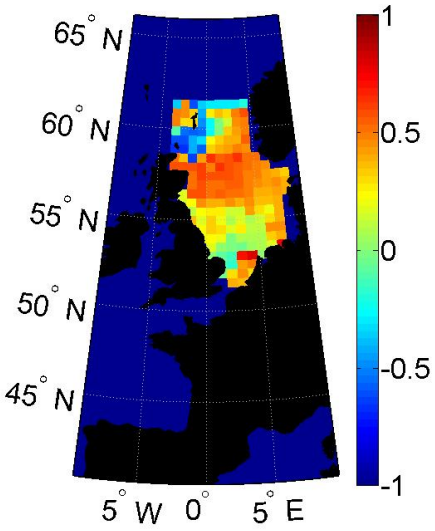
In order to show why taking in to account small time delays is important, it is useful to look at the correlation with physical variables. Initially the analysis was performed without taking in to account time delays. The spatial pattern of the sparsity parameter and the principal components was repeated across higher components, suggesting that without taking in to account misalignment the method was far less effective at separating out different functional groupings. Furthermore when regressing against physical variables the correlation was shown to be significant at only a small proportion of the locations. Figure 4.11 shows the number of locations where the regression between the Northern Hemisphere Temperature and the first principal component is significant when controlling the false discovery rate for both the Fourier and non-Fourier analysis. The red line shows the p-value with locations ordered by increasing p-value across the x-axis. Those with a p-value below the blue line are statistically significant. It is clear to see that the Fourier principal

components correlate far more strongly with the physical variables. This may be because they are more effective at separating out different groupings. It is therefore necessary to take in to account time delays between species, which indicates that not all species respond to climate variables at the same time.

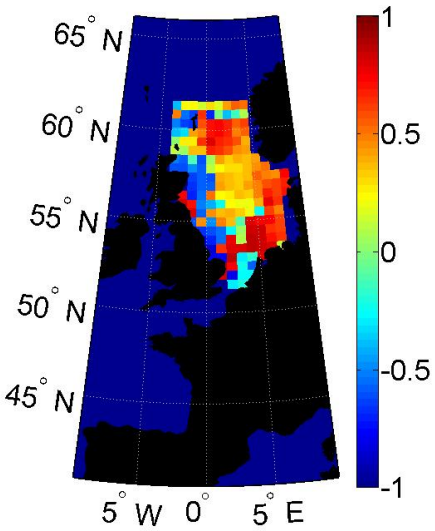
4.4 Modelling Seasonal Data using the WinCPR

Whilst studying yearly averages can give an indication of how the behaviour of the plankton has changed in the long term, studies of the monthly data can be used to determine how seasonal cycles have changed in time. Figure 4.12 shows the sparsity parameter as estimated from the monthly data. There is less spatial variability in the sparsity parameter for the first principal component than there was for the yearly data, although there is a clear north-south divide in the plot showing the number of principal components in space. The dominant model of variability for the monthly data is the seasonal cycles. The seasonality of the plankton is discussed in a number of other studies [14, 25, 79]. From the sparsity parameter on the first component and the number of components, there is a clear north-south divide. This indicates more variability in the southern region, due to the larger number of functional groups, indicating a higher diversity in the seasonal behaviour. The warmer and shallower waters in the southern North Sea [123] may have an impact on seasonality, as seasonal blooms can be driven both by light penetration [111] and by temperature [79]. This north-south divide is seen in subsequent components, although the sparsity pattern has less structure in space. The monthly data tends to have a higher sparsity parameter on the first component than the yearly data. This may be because the seasonal patterns obscure long term trends [25] and so differential responses to long term climate trends are not apparent compared to shorter term responses to seasonal cycles, which may show less differentiation between species.

Figure 4.13 shows the first principal component and the first loading vector clus-

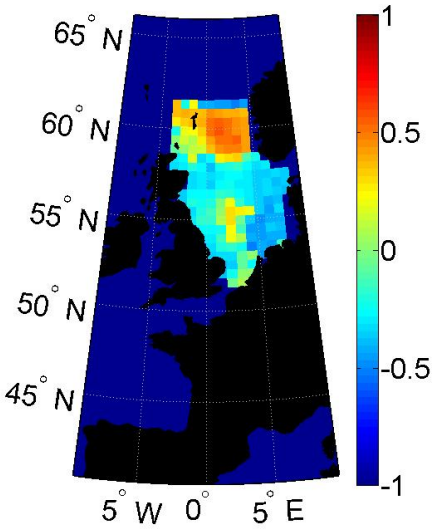


a) Pearson's correlation coefficient for the first principal component.

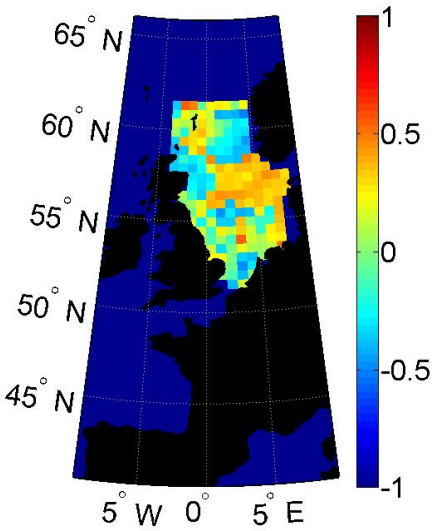


b) Pearson's correlation coefficient for the second principal component.

Figure 4.8: Plot of the correlation coefficients between the principal components for the Fourier transformed data and the Atlantic Multidecadal Oscillation.

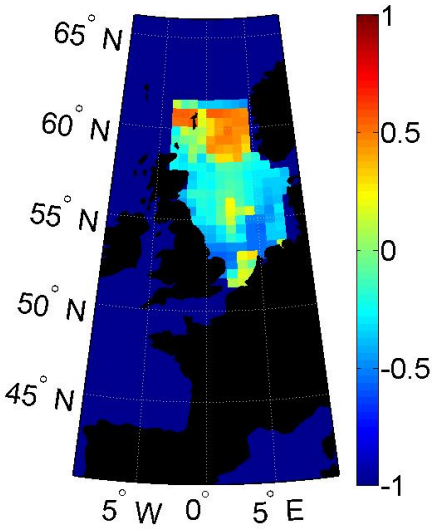


a) Pearson's correlation coefficient for the first principal component.

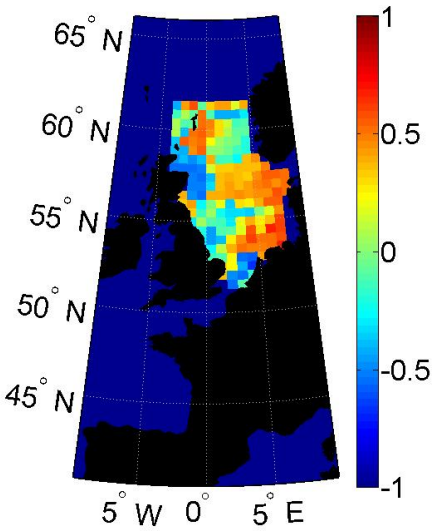


b) Pearson's correlation coefficient for the second principal component.

Figure 4.9: Plot of the correlation coefficients between the principal components for the Fourier transformed data and the NAO.

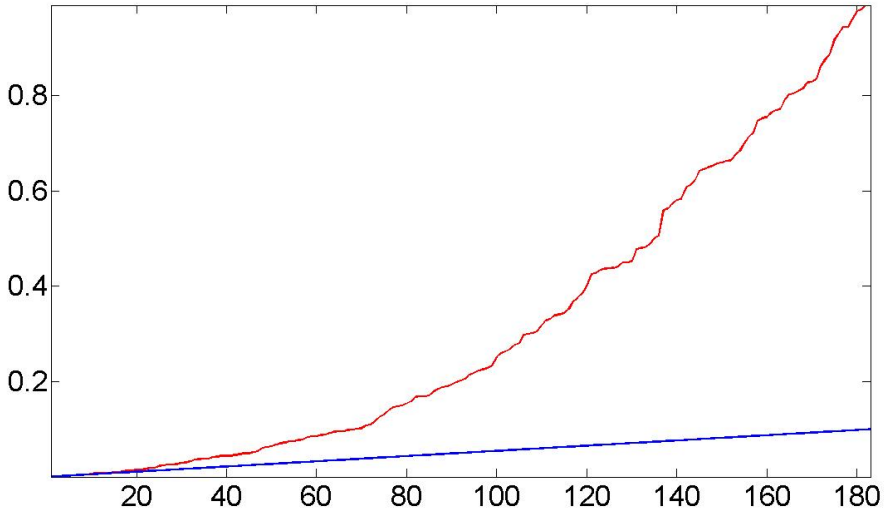


a) Pearson's correlation coefficient for the first principal component.

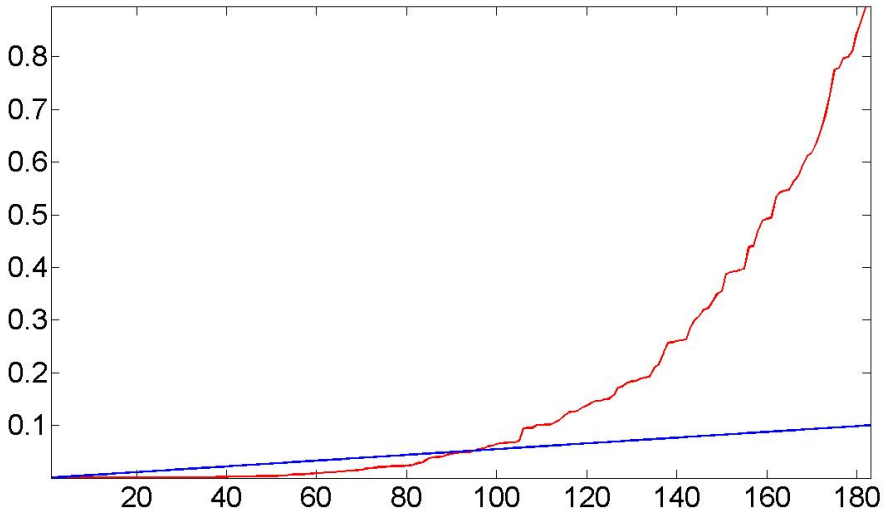


b) Pearson's correlation coefficient for the second principal component.

Figure 4.10: Plot of the correlation coefficients between the principal components for the Fourier transformed data and the Northern Hemisphere Temperature.



a) The first non-Fourier principal component and the NHT.



b) The first Fourier principal component and the NHT.

Figure 4.11: Plot showing the number of locations where the correlation between the first principal component and the physical variables is significant when controlling the false discovery rate for both the Fourier and non-Fourier PCA. The x-axis is the locations ordered by increasing p-value and the p-values are plotted on the y-axis.

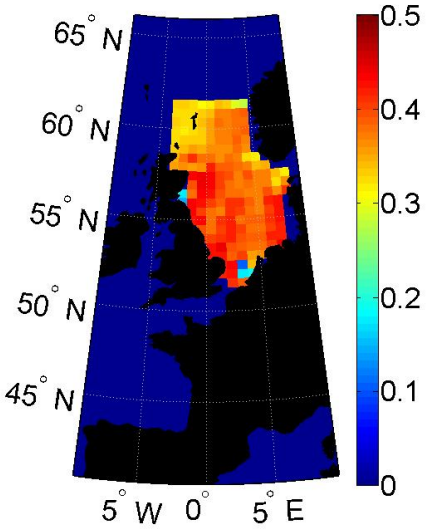
tered across space. There is a clear regionalisation here, although this is differently defined depending on whether one considers the signal or the loading vector. On the time courses the third cluster dominates the central region, whilst the second dominates the north and the south. Cluster one is defined on the north east corner. The time courses are almost entirely dominated by seasonality and are difficult to distinguish, although it appears that the maxima of the oscillation occur slightly later for the time course on the third cluster, the one defined on the central North Sea. The least intuitive result is that the southern North Sea and the north west have been clustered together. The clusters on the loading vectors define a clear regionalisation in to southern, central and northern North Sea. This suggests that the spatial patterns in the seasonal cycles of the plankton are mostly driven by the differences in temperature between the north North Sea and the southern North Sea, which corresponds with existing knowledge that suggests that plankton blooms may occur earlier at warmer temperatures [79, 26].

In order to represent changes in the seasonal cycle one can look at the instantaneous frequency of the signal. The cycles are not perfectly sinusoidal, which is seen in the instantaneous frequencies (see figure 4.14). This means that abundance may be increasing or decreasing at different rates throughout the year. The median value of the instantaneous frequencies tends to lay around 1/12, which suggests that the signals are dominated by yearly cycles. Figure 4.15 shows the first and second principal components for locations 2 and 167 in the Fourier domain. Across all components there is a peak around 0.081, which corresponds to an oscillation with period 12 months. This agrees with what can be intuitively learned from looking at the time courses, that the components on the monthly averages are dominated by the yearly cycles. There is also a peak in component one at location 2 and both components at location 167 around 0.166. This corresponds to an oscillation with period around 6 months. This may be due to the fact that some species have two blooms in the course of a year [111], one in spring and another in summer, and so

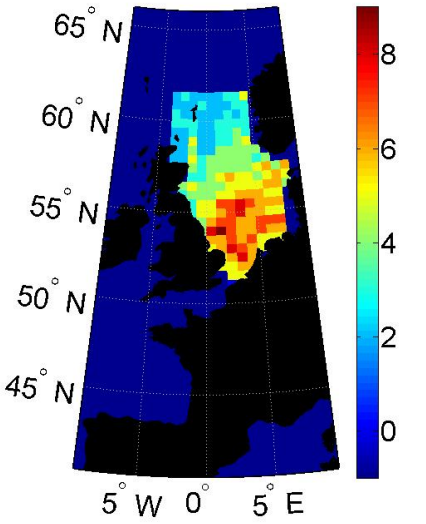
this oscillation captures that behaviour. Some care must be taken in interpreting this however, as the second oscillation may be due to harmonics in the data.

4.5 Discussion

This chapter shows the effectiveness of sparse species PCA as a tool for decomposing spatial, temporal and species structure in the CPR data in a way that is interpretable. The sparsity parameter and the number of components, which can be viewed as measures of diversity, show spatial structure. In particular they take higher values in regions where mixing occur. There is also spatial structure in the species groupings, which show the distinction between north and south and between coastal and open regions. It is important to take in to account the time delays, particularly when investigating the influence of climate variables. When the time delays are not taken in to account the correlation with the climate variables is generally weak, whilst the correlations are significant at more locations when the delays are accounted for. This suggests that species are responding to the climate variables at different times. Long term climate variation is shown to have a statistically significant relationship with the WinCPR data, as does the AMO. Comparing this with previous studies of the physiology [48, 56], which have shown responses of certain species to temperature in experiments and have investigated the effect of water column mixing on diatom species, this can be interpreted as a causal relationship. Spatial structure is also seen in the monthly data, with increased diversity in the southern part of the region. In the Fourier domain the data can be show to have oscillations with a 12 month period and a 6 month period, the first being interpreted as those species with annual blooms and the latter those species which have two blooms per year. The methodology allows for further insights in to the complex structure, as it produces summaries of behaviour in the different dimensions. In this chapter it has been shown that this allows an understanding to be gained that would

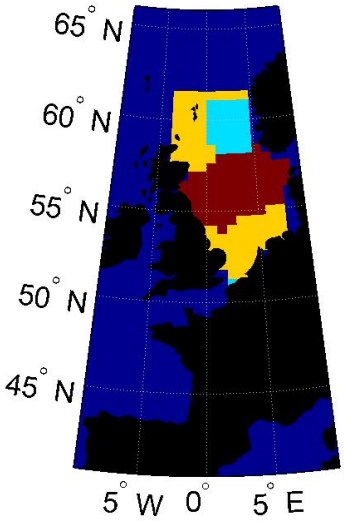


a) Sparsity parameter on the first component.

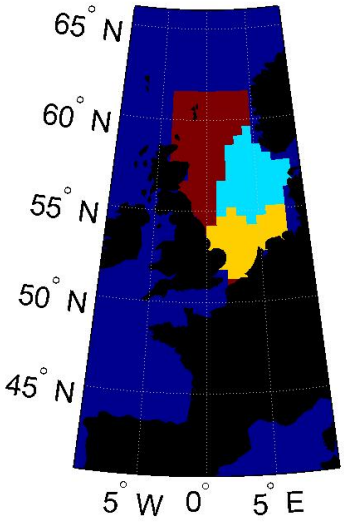


b) Number of components.

Figure 4.12: Values of the sparsity parameter and number of principal components for the monthly data.

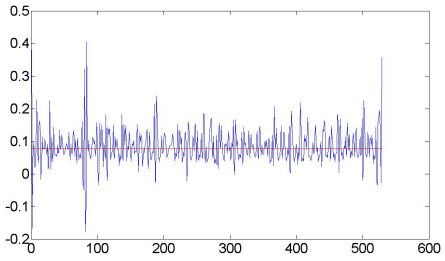


a) Clusters on the first principal component.

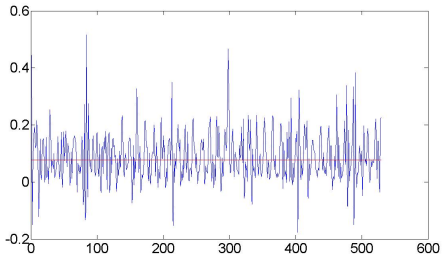


b) Clusters on the first loading vector.

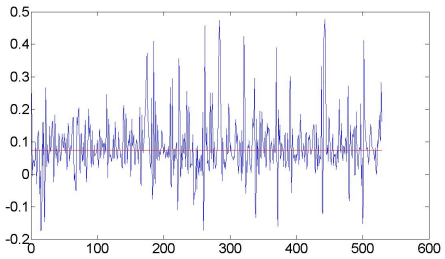
Figure 4.13: Clusters on the first PC and the first loading vector for the monthly data. For the centres of the clusters, cluster one is shown in blue, cluster two in green and cluster three in red.



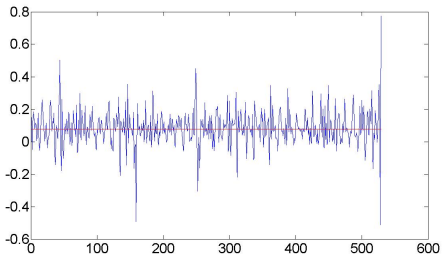
(a) First principal component at location 2.



(b) Second principal component at location 2.

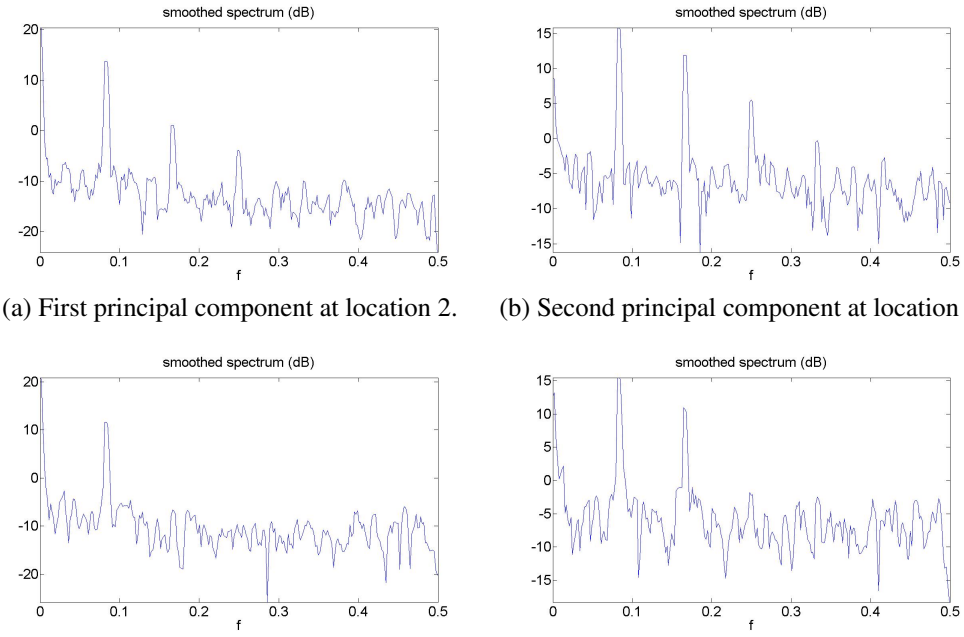


(c) First principal component at location 167.



(d) Second principal component at location 167.

Figure 4.14: Plot showing the instantaneous frequencies in blue and the median of those frequencies in red for the first and second principal components at two locations.



(a) First principal component at location 2. (b) Second principal component at location 2.

(c) First principal component at location 167. (d) Second principal component at location 167.

Figure 4.15: Plot showing the principal components in the Fourier domain for the first and second principal components at two locations.

not have been possible without the statistical tools. It is shown that sparse PCA can be used to investigate structure across space and time whilst studying multiple species, which has not been done using the CPR data before [25].

Chapter 5

Modelling the Raw CPR Data

5.1 Overview

In order to better capture the spatial structure a larger spatial region than the North Sea is required. Raw data for the 109 species included in the WinCPR dataset is then gridded using Kernel smoothing (see equation 2.3), which is used to grid the data by interpolating the value at the centre of each grid square, for the North East Atlantic. The spatial grid is chosen to be 1 degree by 1 degree and extends from 20 degrees west to 10 degrees east and from 42 to 65 degrees north. The raw CPR data contains abundances for species by year, month, longitude and latitude. In order to carry out the analysis the data is interpolated in space and smoothed in both space and time using Kernel smoothing methods [167] then gridded in space. The spatial grid is shown in figure 5.1. Spatial PCA is used by Beaugrand et al [31, 25] to investigate spatio-temporal structure across single species at a time. Since it is impossible to look at multiple species at once with this kind of analysis they investigate indicator species only, which are species thought to be indicative of the behaviour of the system as a whole. In this thesis spatial PCA is carried out on *Calanus finmarchius*, which is a cold water copepod; *Calanus helgolandicus*, a warm water copepod and phytoplankton colour, which is thought to be an indicator of total phytoplankton

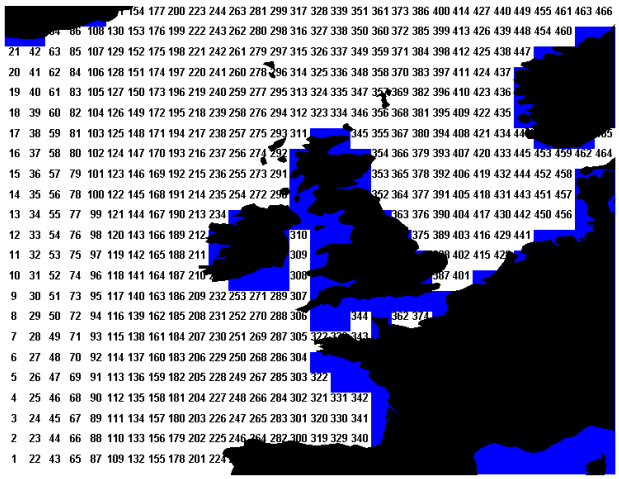


Figure 5.1: Plot of the pixel locations for the gridded data across space.

biomass. Whilst such an analysis can give a general indication of the behaviour of different species groupings, the drawback is that the choice of indicator species has to be drawn from various assumptions about the data. In order to explore functional groups of species sparse PCA (see section 2.3) is carried out first on spatially averaged data to give an overview of their joint behaviour across the whole region and then at each location in order to determine spatial structure, with the weights as functions of species and the components as functions of time (see figure 2.2).

5.2 Methods Used in this Chapter

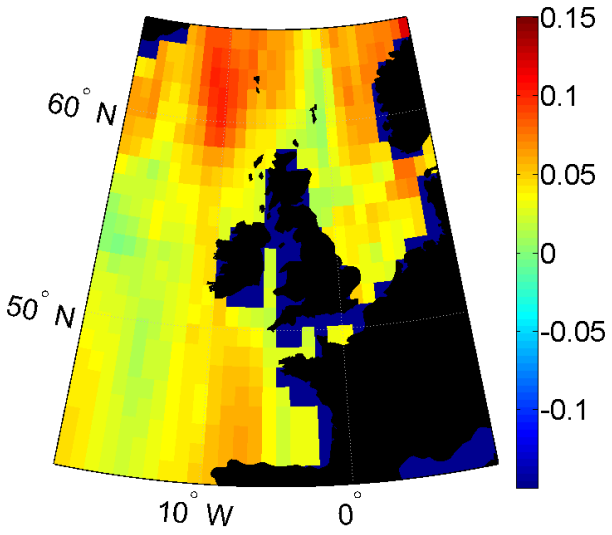
In this chapter the raw CPR data has been gridded and smoothed for each month separately using Kernel Smoothing (section 2.1) and then yearly averages are taken. Recall that Kernel Smoothing can be used both for interpolation and smoothing, as missing data is estimated based on weighted sums of the observed data at nearby locations, where the weights are a function of the distance which is maximised at zero. Spatial PCA, where the weights are functions of location and the com-

ponents are functions of time (section 2.2), is used in this chapter to analyse the spatio-temporal behaviour of several indicator species. The components represent the dominant trends for the indicator species and the weights show the spatial pattern of these trends. Sparse species PCA (sections 2.2 and 2.3) is then used on spatially averaged data to find species assemblages across the entire region. Recall that this analysis can be used to select keystone species by forcing the weights of rarer species to be zero and that it will also find the joint behaviour of these keystone species. This analysis is also repeated across each location to determine how both dominant species groupings and common trends vary across space, which is assessed using cluster analysis to find regions where the signals are similar (section 2.4). Spatial variability is also explored using measure of diversity, which in sparse species PCA is represented by the number of components, i.e. the number of assemblages, and the sparsity parameter, the size of these assemblages. Finally the relationship between the dominant trends at each location and various potential climate drivers is explored using Pearson's correlation coefficient (section 2.5), which as previously described is a measure of both the strength and the direction of the relationship between two variables. In this case a strong relationship between a component and a climate variable might be seen as an indication of causality if there is a biological mechanism that might explain the relationship.

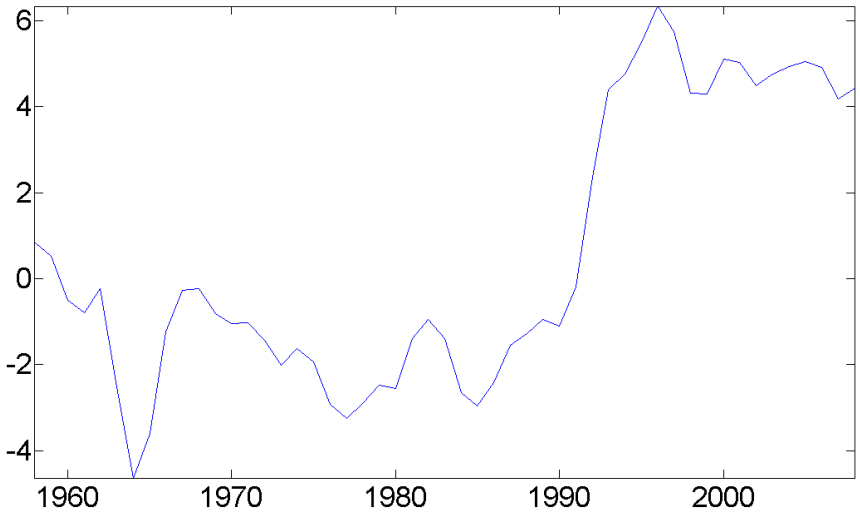
5.3 Spatial PCA on Indicator Species

5.3.1 Spatial PCA for Phytoplankton Colour Index

Figure 5.2 shows the first principal component and loading vector for phytoplankton colour. The first component shows a stepwise increase post 1990, in that the average signal is greater after 1990 than before, perhaps corresponding to the 'regime shift'. This can be interpreted as a significant change in biomass of phytoplankton on average recorded after 1990 and as can be seen from the component this

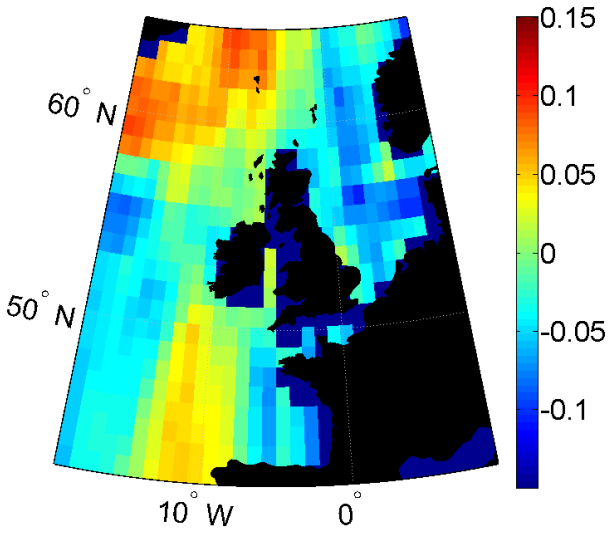


a) The first loading vector.

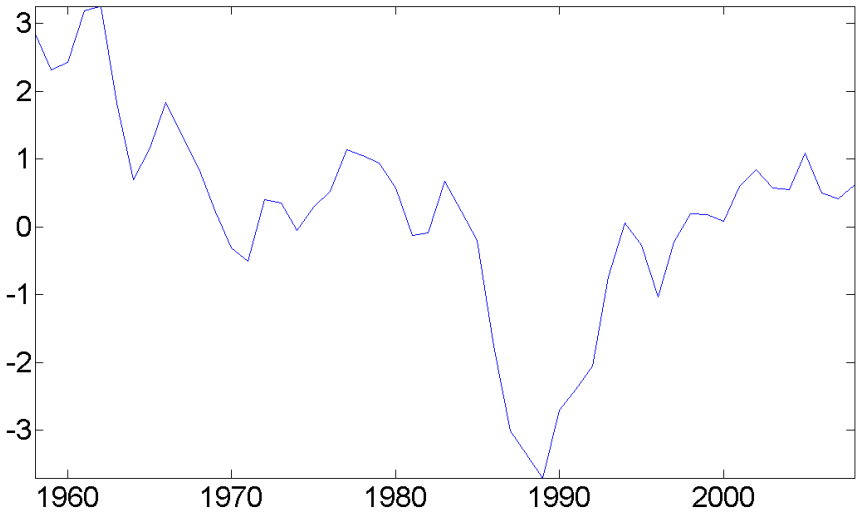


b) The first principal component.

Figure 5.2: Plot of the loading vector and the first principal component for phytoplankton colour.



a) The second loading vector.

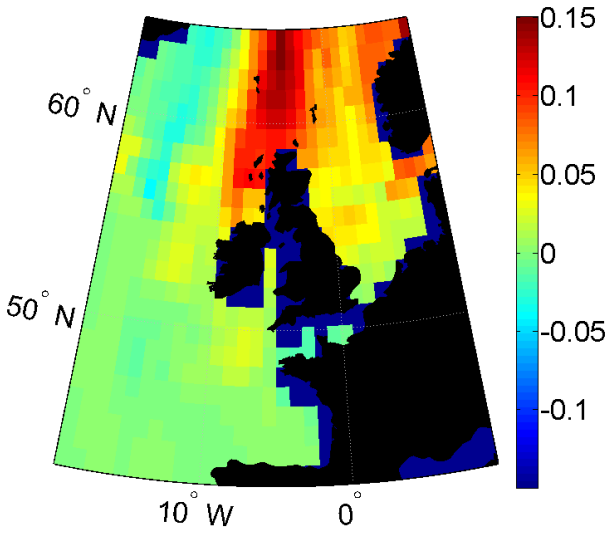


b) The second principal component.

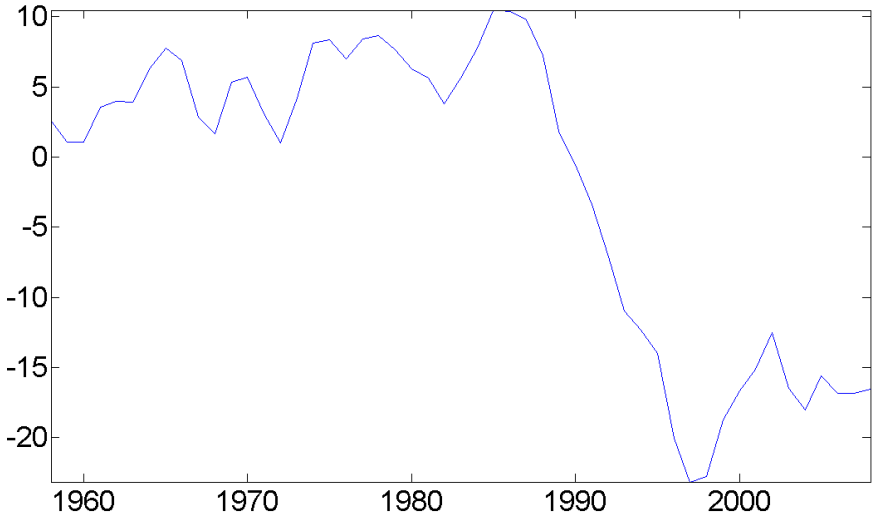
Figure 5.3: Plot of the loading vector and the second principal component for phytoplankton colour.

change is not gradual. Since the loading vector is positive across the entire region, it can be concluded that the phytoplankton colour index is on average increasing across the entire North East Atlantic, indicating a total increase in phytoplankton abundance. The increase in total phytoplankton biomass has been reported in other studies [13, 99]. However since this index is a representation of biomass only this is not informative as to how the structure community as a whole is changing [99] and it may be that certain species are decreasing or showing no long term change in abundance. Small species of phytoplankton may also be underestimated by the CPR survey due to the size of the mesh [68], as cells that are small than the mesh are less likely to be collected. This means that although this index can be used as an indication of the overall change in phytoplankton, some care must still be taken in the interpretation of this result.

The second component is an oscillation in time (figure 5.3) and is positively weighted in the North West corner, whilst having negative weights in the North Sea and the south west. The oscillation has a minimum just before 1990. One possibility is that it may correspond to the AMO or a time-lagged version of the AMO, since it has a similar period to the AMO and there is a possibility that this index might be influential to certain phytoplankton species. Diatom species may be sensitive to the AMO [56] and so this may explain why the AMO is a driver of phytoplankton abundance in general. Comparing with the spatial pattern of the second principal component on the sea surface temperature data, which is identified with the AMO (figure 3.4), shows the AMO is positively weighted in the north of the region. This may explain why the second principal component on the phytoplankton colour data has positive weights in the north. The third principal component represents a shorter term oscillation and has positive weights in the bay of Biscay and negative weights north of Scotland. It is not clear if this is a response to a particular climate driver.

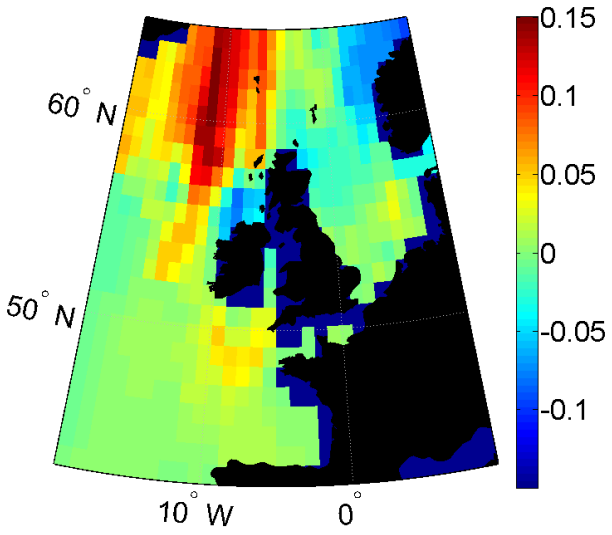


a) The first loading vector.

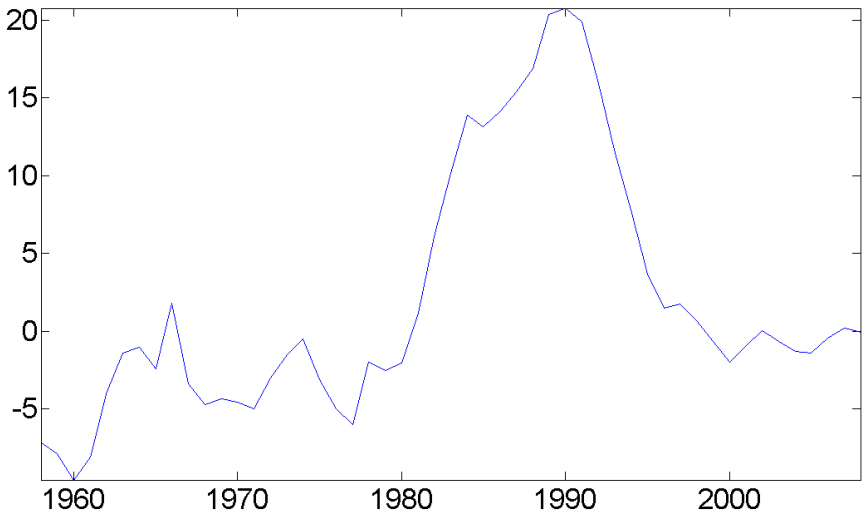


b) The first principal component.

Figure 5.4: Plot of the loading vector and the first principal component for *Calanus finmarchius*.



a) The second loading vector.



b) The second principal component.

Figure 5.5: Plot of the loading vector and the second principal component for *Calanus finmarchius*.

Component	NHT	AMO	NAO
1	-0.7057	/	/
2	/	-0.4473	/
3	/	/	0.3830

Table 5.1: Pearson's correlation coefficients between the first three principal components of *Calanus finmarchicus* and the climate indices.

5.3.2 Spatial PCA for *Calanus finmarchicus*

Figures 5.4 and 5.5 show the first and second principal component for *Calanus finmarchius* respectively and table 5.1 shows the Pearson's correlation coefficients between the principal components for *Calanus finmarchius* and different climate indices. In the first principal component the cold water species [117] is most strongly weighted in the north and shows a rapid decline over time. It is believed that one of the effects of climate change is that warm water species have moved northwards with rising temperatures, whilst cold water species begin to loose out due to the effects of climate change [12]. The trend has positive weights across the region, suggesting this decline is happening across the whole North East Atlantic, agreeing with pre-existing knowledge about the behaviour of this species in response to climate change [74]. It has a negative correlation with the NHT warming trend, with a Pearson's correlation coefficient of -0.7057, again showing that as temperatures rise the abundance is decreasing across the North East Atlantic.

The second principal component shows an oscillation in time and is negatively correlated with the AMO with a Pearson's correlation coefficient of -0.4473, suggesting some link between the AMO and *C. finmarchicus*. The third principal component is an oscillation with a peak just before 1990 and has positive weights in the central north of the region. It correlates positively with the third component of sea surface temperature, the NAO signal, with a Pearson's correlation coefficient of 0.3830. The NAO is thought to have an influence on *C. finmarchicus* [59], although this is far less important than the general warming trend [74].

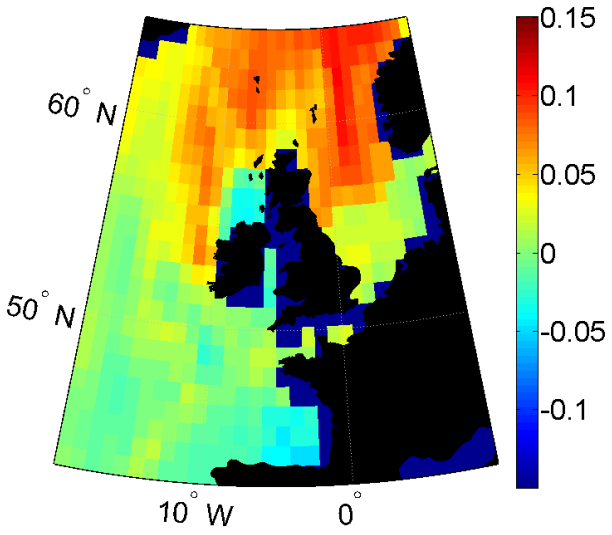
5.3.3 Spatial PCA for *Calanus helgolandicus*

Component	NHT	AMO	NAO
1	0.7273	/	/
2	/	/	/
3	/	-0.3819	/

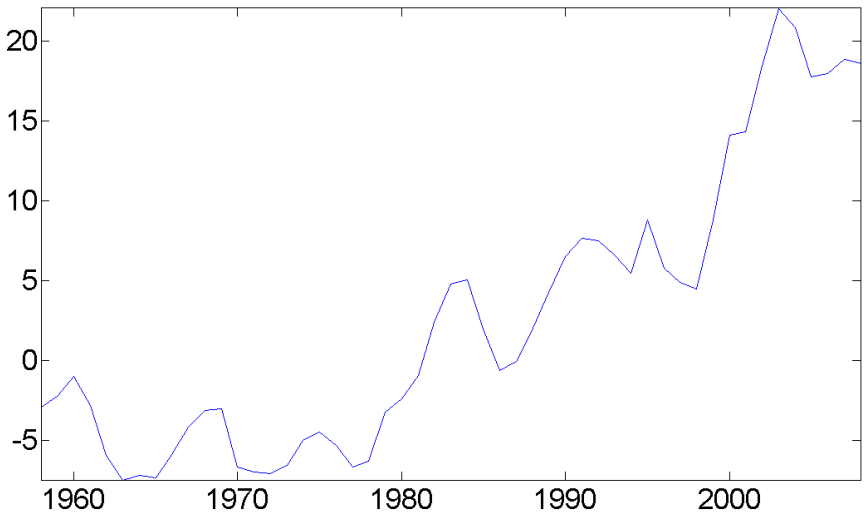
Table 5.2: Pearson's correlation coefficients between the first three principal components of *Calanus helgolandicus* and the climate indices.

The first principal component on *Calanus helgolandicus*, which is a warm water copepod [117], shows a general trend of increasing abundance across the North East Atlantic (figure 5.6), particularly in the north where it is strongly positively weighted. This trend also has a significant correlation with the first principal component on the non-detrended sea surface temperature data over the North East Atlantic, which was interpreted as being the NHT warming trend. The Pearson's correlation coefficient is 0.7273 and the correlation is significant with a p-value of less than 0.05. The Pearson's correlation coefficients between the principal components for *Calanus helgolandicus* and different climate indices are shown in table 5.2. The relationship between the first component and the NHT indicates that whilst *C. finmarchicus* is decreasing across the entire North East Atlantic, *C. helgolandicus* is increasing. The rate of increase in the northern North Sea supports the theory that warm water species are replacing cold water species at higher latitudes [12].

The second principal component (figure 5.7) is short period oscillation, most strongly weighted in the North Sea. It has some correlation with the third sea surface temperature component, which was identified with the NAO. As with *C. finmarchicus* the NAO is thought to influence the abundance of *C. helgolandicus* [59], although it has an opposite effect. This demonstrates that shorter term oscillations are important to plankton abundance as well as long term warming trends. The third principal component is also an oscillation but is most strongly weighted in the south west. It correlates negatively with the second component of the sea surface temper-

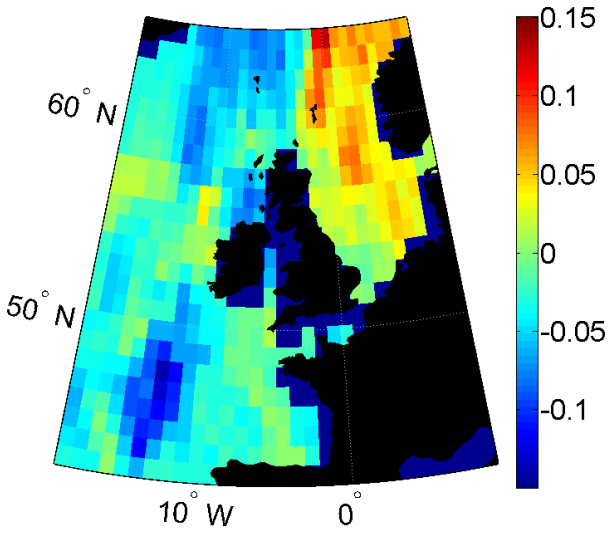


a) The first loading vector.

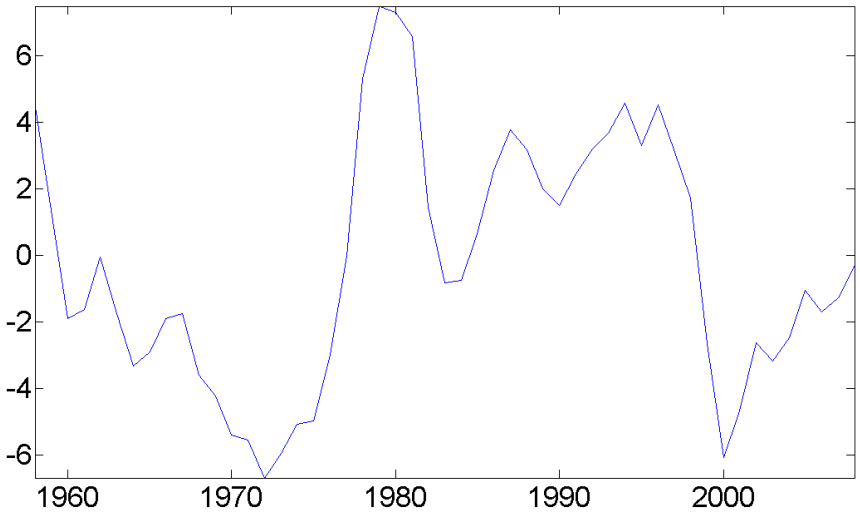


b) The first principal component.

Figure 5.6: Plot of the loading vector and the first principal component for *Calanus helgolandicus*.



a) The second loading vector.



b) The second principal component.

Figure 5.7: Plot of the loading vector and the second principal component for *Calanus helgolandicus*.

ature, which was identified with the AMO, with a Pearson's correlation coefficient of -0.3819, suggesting the AMO has a lesser influence on *C. helgolandicus*.

5.4 Species Principal Component Analysis for Spatially Averaged Data

5.4.1 Averaged Zooplankton Data

Component	NHT	AMO	NAO
1	-0.7017	/	/
2	/	/	/
3	/	-0.5202	/

Table 5.3: Pearson's correlation coefficients between the first three principal components of the zooplankton data averaged over space and the climate indices.

In using spatial PCA significant species must be chosen by making assumptions about the behaviour of the dataset. The strength of sparse PCA is that species are automatically chosen from the data without any prior assumptions about their importance. Rather than considering all species together, which can be problematic given differences in the ways in which different groups are counted, two large groups are considered separately: the zooplankton and the Diatoms.

Figure 5.8 shows principal components on the zooplankton community for data averaged over the North East Atlantic. The first component has a strong negative correlation with the NHT warming trend, with a Pearson's correlation coefficient of -0.7017, see table 5.3. This trend accounts for a large portion (50.87%) of the total variation in the abundance across all species averaged in space. Amongst those species whose loadings have a positive real part are *Chaetognatha traverse*, *Echinoderm larvae* and *Calanus helgolandicus*. *Chaetognatha* are a species of marine worms and are found both in cold and in temperate waters [68], they often prey upon Copepods and next to Copepods are the most numerous species of zooplank-

ton [68]. They are also found exclusively amongst the plankton [68]. *Echinoderms* are a group of marine animals, which includes both starfish and sea urchins. Whilst the adults live in primarily benthic communities, the larvae are pelagic and form part of the plankton [88]. These species are increasing along with rising sea surface temperatures. *Calanus finmarchicus* has instead a weight with a negative real part, indicating its abundance is on average over the whole North Atlantic declining in time, which agrees with the individual species analysis. Other species with weights with negative real parts include *Para-Pseudocalanus spp.*, *Acartia spp.*, *Oithona spp.* and *Pseudocalanus spp. (adult atlantic)*. *Para-Pseudocalanus* and *Acartia* are both small species of copepod, which are more abundant in the northern North Sea than in the southern North Sea [68], suggesting a negative relationship with temperature. *Oithona* and *Pseudocalanus* are also species of copepod. Since the first component accounts for a large proportion of the variation it is clear that the climate warming trend is an important driver for zooplankton species and the difference in the sign of the weights between different species suggests that increasing temperatures will lead to a dramatic change in the composition of the zooplankton community [132, 30, 22, 21].

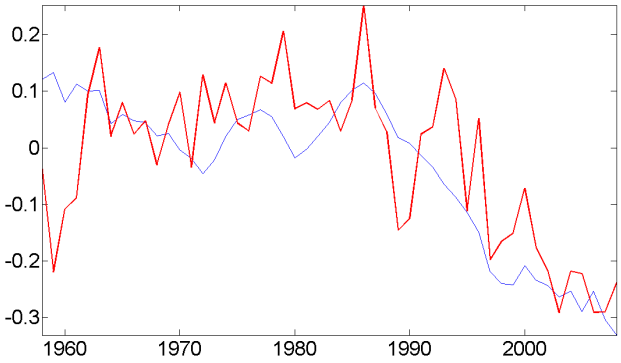
The time signal for the second component on the spatially averaged zooplankton abundances is an oscillation with a minimum around 1970 and a peak in the mid-1990's. The taxa *Centropages typicus*, *Podon spp.*, *Evadne spp.*, *Chaetognatha* abundance and *Echinoderm larvae* have strong weights on the second component all with negative real parts. *Centropages typicus* is a calanoid copepod. Although both *Podon* and *Evadne* can be found in open oceans they are most common in coastal regions [68] and the spatial pattern of their abundance has a tendency to be patchy [68]. *Chaetognatha* and *Echinoderm larvae* are found in many regions. One possibility therefore is that the second component could relate to some coastal effect not determined by sea surface temperature.

The third component on the zooplankton community accounts for 11.37% of the

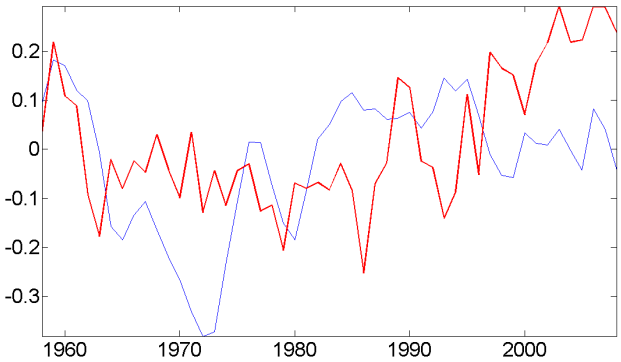
total variation and is correlated with the second sea surface temperature component with a Pearson's correlation coefficient of -0.5202 (table 5.3), which is driven by the AMO. *Centropages typicus*, *Evadne* and *Chaetognatha* eyecount all have weights with positive real parts. *Echinoderm larvae* and *Calanus finmarchicus* conversely have weights with negative real parts. In the spatial PCA for *Calanus finmarchicus* it was shown that the second component on the sea surface temperature was negatively associated with the second species component, suggesting an inverse relationship between the AMO and *C. finmarchicus*. The simple benthic zooplankton taxa *Copepod nauplii* has a very strong negative weight on the third component. As previously discussed the AMO can have an influence on water column mixing [56, 43] and so might have an effect on plankton though its influence on the mixed layer depth.

5.4.2 Averaged Phytoplankton Data

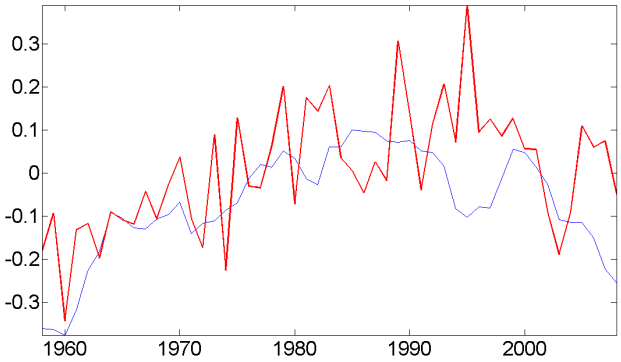
Figure 5.9 shows species principal components for diatom species averaged over the North East Atlantic. The most important influence seems to be the AMO. Although the results of the SST analysis showed that the AMO was not a particularly important driver of SST in the North Sea, it does have an influence on the diatoms in this region. This implies that there is an indirect effect of the AMO on local climate in this region, perhaps linked to its influence on currents and wind speeds [141], which in turn influences the behaviour of the diatoms. This suggests that diatoms are not affected by temperature, aside perhaps from a secondary effect cause by changes in their predators, so much as hydro-climatic variability [56]. The first component of the phytoplankton communities accounts for 43.19% of the total variation across the different species averages and is positively correlated with the second sea surface temperature principal component. The time signal resembles an oscillation with a period of about 50-60 years and a minimum around 1980. Almost all of the species in the phytoplankton group are diatoms and have positive weights with respect to



a) First principal component (blue) and the first principal component on the sea surface temperature multiplied by minus one (red).



b) Second principal component (blue) and the second principal component on sea surface temperature (red).



c) Third principal component (blue) and the first principal component on sea surface temperature (red).

Figure 5.8: Plots of the first three species principal components on the zooplankton subgroup averaged over the North East Atlantic.

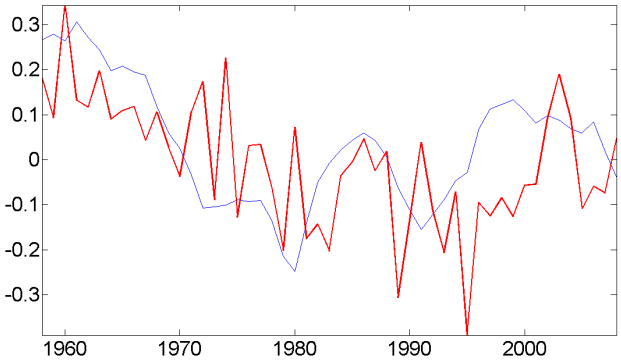
the first component, such as types of *Pseudo-nitzschia* and *Skeletonema costatum*. The only negatively weighted species are the two diatoms *Bacillaria paxillifer* and *Gyrosigma spp.*. This suggests for the most part diatoms are more abundant in the positive phase of the AMO and that the average behaviour across Diatom species is influenced by the AMO.

The second component on the phytoplankton is an increasing trend in time and accounts for 20.8% of the variation. About two thirds of the species have weights with positive real parts, amongst those are diatom species *Paralia sulcata* and *Gyrosigma spp.*. Species with negative weights include: *Thalassiothrix longissima*, *Fragilaria spp.* and *Navicula spp.*, which are also all diatoms. This may represent the response of Diatom species to the warming trend, which may be a secondary driver of their abundance. Other studies have shown that this change is heterogeneous across the North Sea, with a general increase in Diatom abundance in some regions, a decrease in others and no change in some areas [99]. The averaged trend will not capture differences in spatial responses, which shows why it is important to analyse the data at a finer resolution too. The third component explains 7.19% of the total variation and has a positive correlation with the fourth temperature component. The time signal shows a small increase over time, along with an oscillation of period around 15 years.

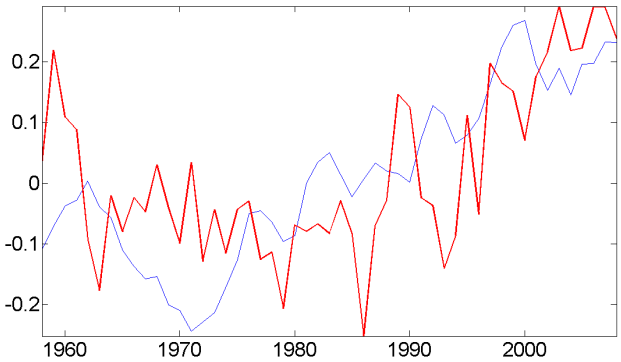
5.5 Multivariate and Spatio-Temporal Structure Modelled by Sparse PCA

5.5.1 Modelling Across all Species

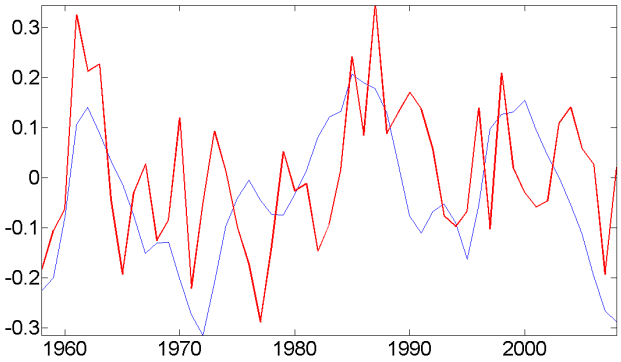
The analysis can be carried out over each location separately rather than spatially averaged data in order to determine spatial variation in responses. Figure 5.11 shows the values of the sparsity parameter for the gridded data across all species. There is an east west gradient in the first sparsity parameter, which might be at-



a) First principal component (blue) and the second principal component on the sea surface temperature (red).



b) Second principal component (blue) and the first principal component on sea surface temperature (red).



c) Third principal component (blue) and the fourth principal component on sea surface temperature (red).

Figure 5.9: Plots of the first three species principal components on the phytoplankton subgroup averaged over the North East Atlantic.

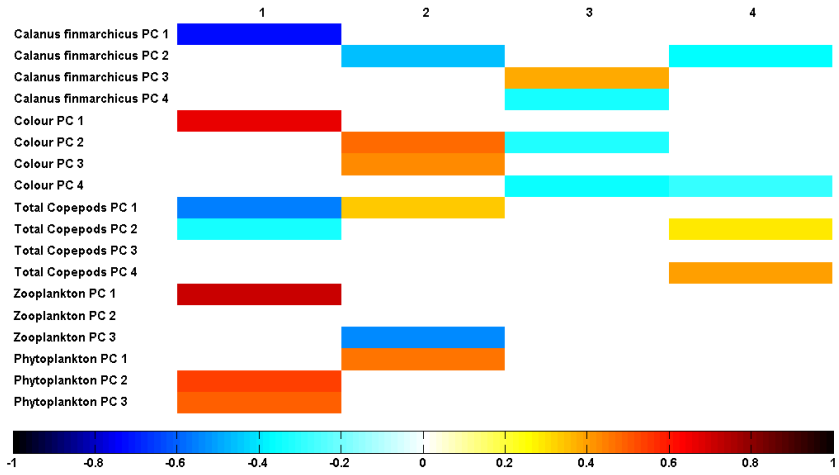
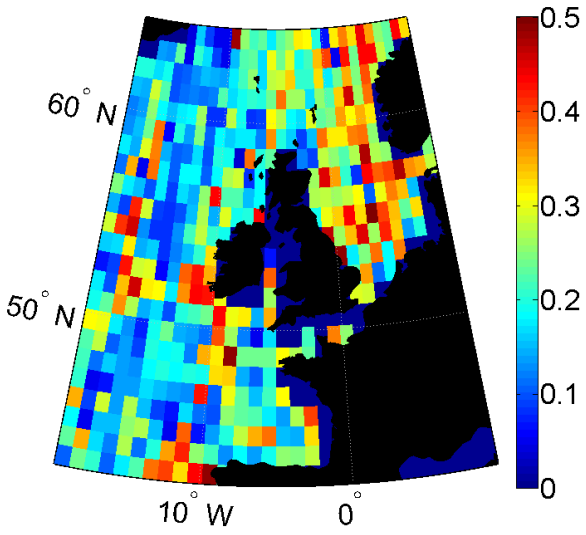
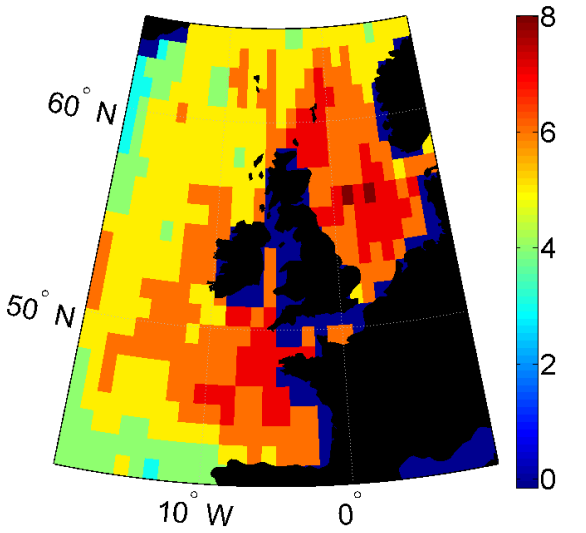


Figure 5.10: Plots of the Pearson correlation coefficients between the principal components for the plankton data and the principal components on the sea surface temperature data, see figure 3.4.

tributable to physical variables including currents, bathymetry and salinity and the influence of these variables on the plankton [56, 105, 9]. The salinity has an east-west gradient [9], as does the bathymetry [9], and so either of these variables may be influencing this structure. The sparsity parameter on the first principal component is in general higher in the shallower regions, indicating a greater degree of diversity here. Subsequent sparsity parameters have less spatial structure but the number of components is structured, with more groups being needed in the ocean shelf region [123], where waters are shallower. Figure 5.12 shows clusters on the first principal component and the centres associated with each cluster. The interpretation of this is that the regions are defined by species assemblages with similar functional behaviour, possibly because they are subject to similar climate effects. There is a north south divide in the spatial regions. In the north, covered by cluster one, there is an increasing trend. In the other two clusters the change is less dramatic but there seems to be a slight decline in the trend. These regions might be a response to the differences in general abundance trends across the north and south



a) Sparsity parameter on the first component.



b) Number of components.

Figure 5.11: Values of the sparsity parameter and the number of components plotted for the whole of the North East Atlantic.

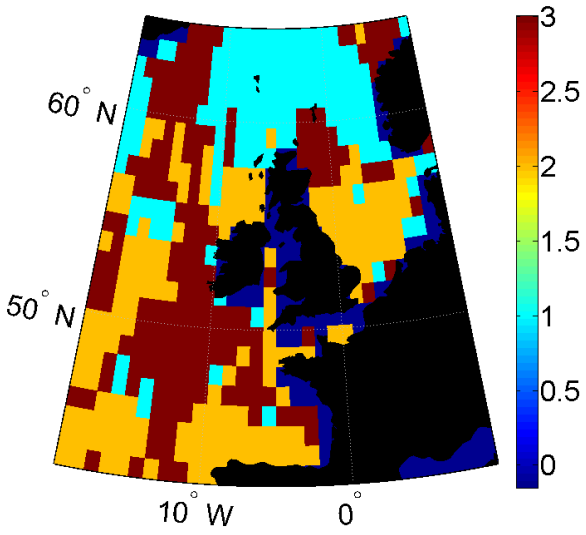


Figure 5.12: Clusters on the first principal component for all species.

of the region. This is seen in the phytoplankton, where although total abundance is increasing there are differential responses between diatoms and dinoflagellates across space [99], and in the zooplankton as changes in the composition across different regions [26, 32, 75]. As these trends are aggregations of different types of species they represent the average changes in abundance.

Spatial patterns might also be determined by the first loading vector. These clusters might be interpreted as defining ecoregions because they are determined by the dominant species. Figure 5.13 shows the clustering on the real part of the first loading vector. The first cluster is mostly dominated by phytoplankton species. A few zooplankton species have non-zero weights, for instance the cold water copepod *Calanus finmarchicus* that has a negative weight in cluster three that covers mostly the North Sea, where it is declining [74]. When clustering on the species instead of a north south divide the regions are determined on an east west divide. This may be due to the effect of bathymetry, as the North Sea is shallower than the rest of the North East Atlantic [123] and so may support different subgroups of species. Figure 5.14 shows the Pearson's correlation coefficient between the first principal

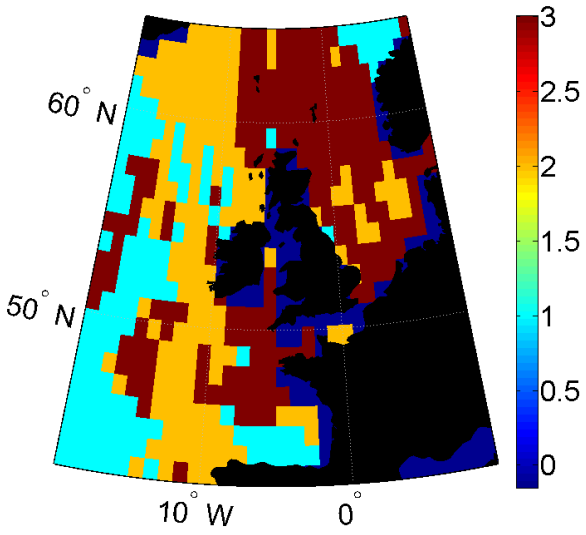


Figure 5.13: Clusters on the real values of the first loading vector for all species.

component on all species compared with the NHT warming trend and the AMO. In the North Sea region there is a strong positive correlation with the AMO and the warming trend. The correlation with the NHT is positive in much of the North Sea but far weaker than the relationship with the AMO. From this it can be concluded that the first principal component is driven by a mixture of the warming trend and the AMO, as it was for the WinCPR dataset. This is due to having combined species that are sensitive to temperature [9, 98, 117] with those that are more sensitive to wind intensities [56].

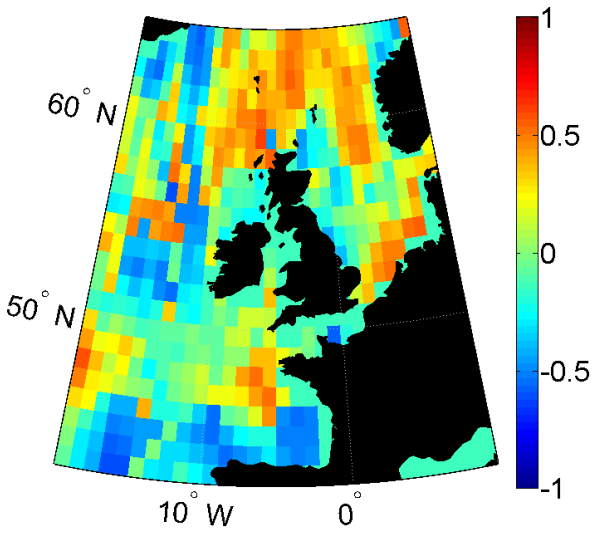
Clusters on the second principal component for the data across all species are less well defined than on the first principal component, although the North Sea seems to be separated slightly from the rest of the region, since most of cluster one is confined to the North Sea. The time signal in cluster one shows a slight increase. Cluster three also has an increasing time signal and there is a decline in cluster two. For the clusters on the real values of the second loading vector cluster two is mostly confined to the North Sea. Of the zooplankton in the second cluster *Calanus helgolandicus* and *Echinoderm larvae* are on average positively weighted, whilst

Calanus finmarchicus is negatively weighted. This means that those zooplankton that respond positively to temperature have positive weights in this cluster and those that respond negative have negative weights [26, 32, 75]. The most strongly weighted phytoplankton species in this cluster are *Paralia sulcata*, which is influenced by changes in temperature [64], and *Cylindrotheca closterium*, which both have weights with positive real parts. The other two clusters are mostly dominated by phytoplankton species. For the second principal component the North Sea region has a slight positive correlation with the NHT warming trend, which agrees with the distribution of positive and negative weights between cold and warm water species. This suggests that as the temperature warms a change in species is observed. There is a positive correlation between the second principal component across most of the open ocean and the AMO. In this region the phytoplankton species tend to have strong positive weights, which can be influenced by currents [56]. The correlations are, however, weaker than for the first principal component.

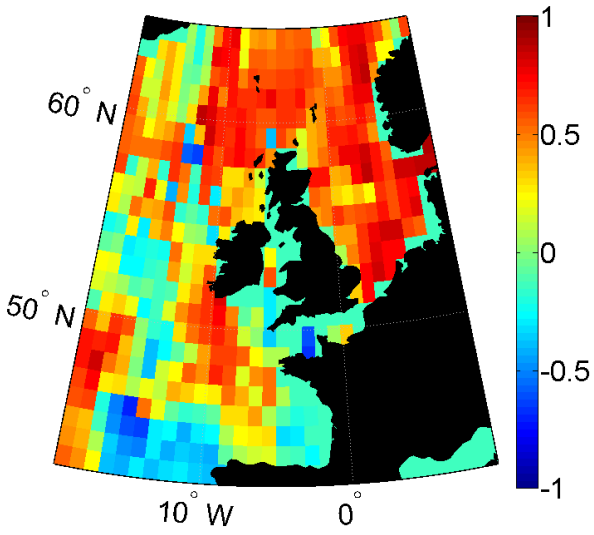
5.5.2 Results for the Zooplankton Group

There are a number of reasons that it might be advantageous to study different species groups separately. Amongst these is the fact that zooplankton and phytoplankton may respond to different trends and that they have different biomasses [68], which makes it difficult to compare them.

Figure 5.15 shows the sparsity parameter for the first four principal components across the entire time period from 1958 till 2009 for the zooplankton communities and the number of principal components found by thresholding the culmative explained variance. Whilst the number of components can be seen as representing the number of distinct assemblages, or the diversity of assemblages, the sparsity parameter is a representation of the number of different species in each assemblage. The sparsity parameter on the first component (see figure 5.15 a.) follows the pattern of the bathymetry of the North East Atlantic. Where the oceanic shelf is higher,



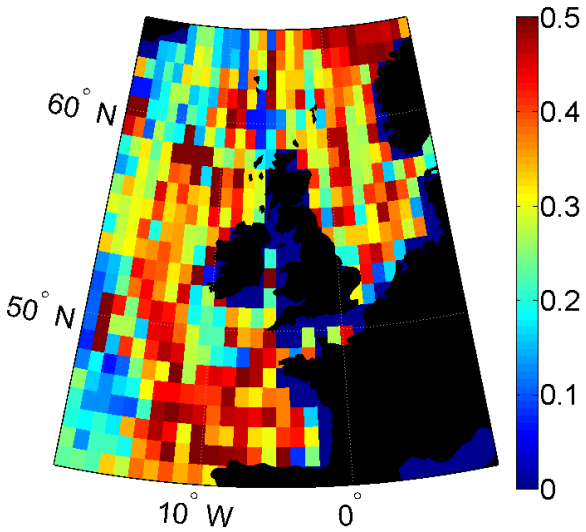
a) Regression against the NHT warming trend.



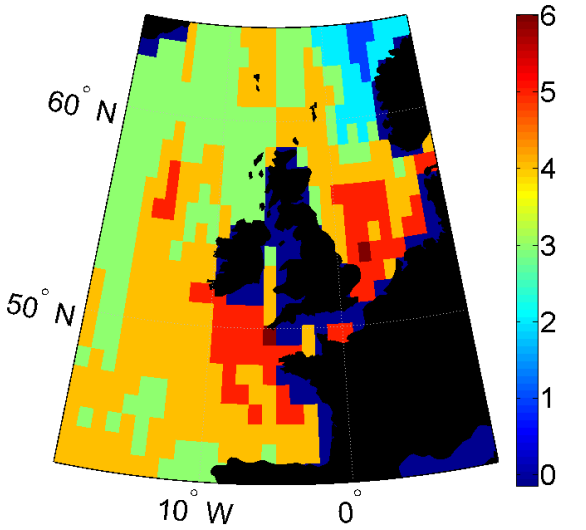
b) Regression against the AMO.

Figure 5.14: Plots of Pearson's correlation coefficient between the first principal component at each location and the climate indices.

which follow a curved line from about 5° west in the north to about 15° west in the south, the sparsity parameter is also larger. By contrast in the open ocean where waters are deeper the sparsity parameter is lower, meaning fewer species are included when calculating the first component. One explanation for this might be that the first component includes a number of species which are only found in shallower waters, such as species of planktonic worms [68]. The sparsity parameter on the second component is less structured in space. The number of principal components is higher in the south than in the north, particularly in the southern North Sea and south of the United Kingdom. The Bay of Biscay, which is known to be a diverse and fertile region [62] since it is a mixing region [57], also has a large number of principal components. This could be viewed as a measure of the number of functional groups of species, which could in turn been seen as a measure of diversity. This indicates a spatial correlation between sea surface temperature, which is higher in the south, and diversity [9]. Figure 5.16 shows clusters based on the real values of the loadings for the first principal component on the data for the full time course restricted to zooplankton species. Cluster one in the spatial plot is light blue, cluster two is orange and cluster three is dark red. In the spatial plot cluster two covers most of the north of the region, cluster one covers the edges between clusters two and three and cluster three is mostly restricted to the southern part. In cluster one, the mixing region, species such as *Temora longicornis*, a small copepod [68]; *Centropages typicus*, an oceanic copepod [68], and *Chaetognatha Traverse*, a marine worm [68], have positive weights. *Oithona spp.* and *Evadne spp.*, both of which are copepods, meanwhile are negatively weighted. In cluster two *Podon spp.* and *Evadne spp.* are positively weighted and *Calanus finmarchicus* has a negative weight. The average time course in this region is increasing, suggesting that the first two species are increasing in abundance and *C. finmarchicus* is declining. Positively weighted species in cluster three are *Acartia spp.*, *Oithona spp.*, *Evadne spp.* and *Pseudocalanus spp. (Adult Atlantic)*, all species of copepod [68]. The

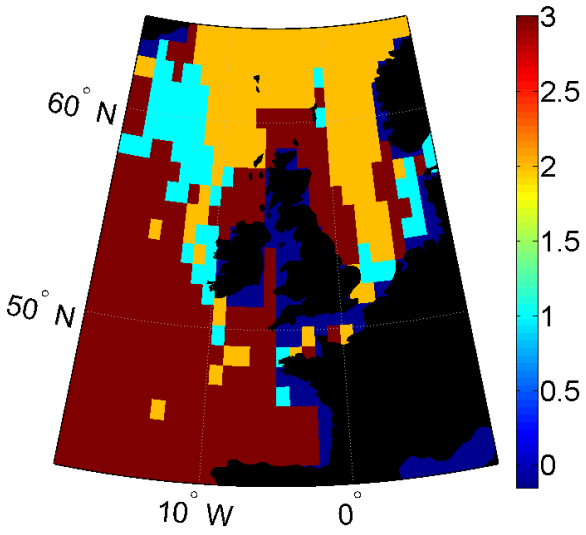


a) Sparsity parameter on zooplankton species and across all time for the first principal component.



b) Number of principal components for zooplankton species and across all time.

Figure 5.15: Sparsity parameter and number of principal components across zooplankton species for the entire time course from 1958 till 2009.

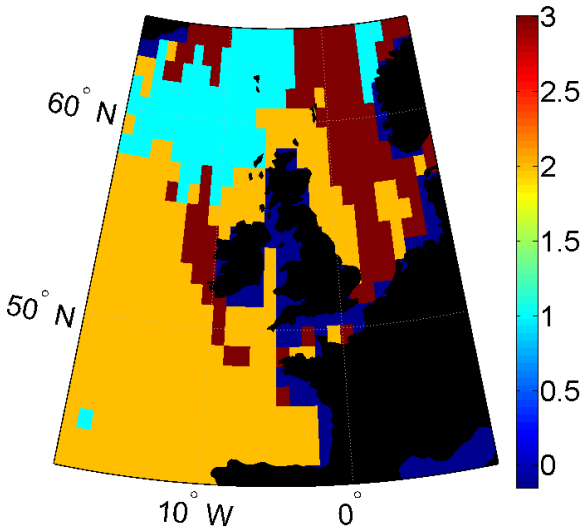


Regions based on zooplankton species over all time.

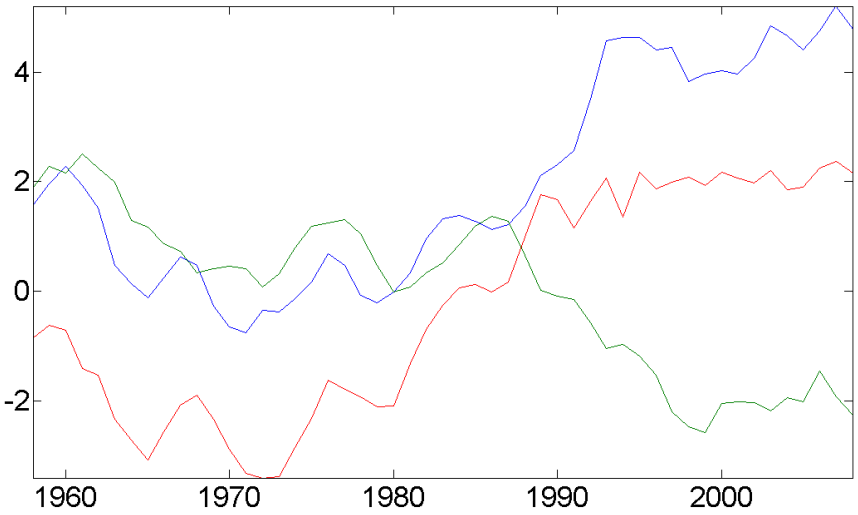
Figure 5.16: Plots of regions over all time based on clusters on the first loading vector. Cluster one is blue, cluster two is orange and cluster three is dark red.

regions defined on the zooplankton seem to be governed both by temperature and sea depth [123], suggesting both are drivers of species composition. Figure 5.17 shows the clusters on the first principal component, with cluster one in blue, cluster two in orange and cluster three in dark red. Clusters one and three on the signals overlap with clusters one and two on the loading vector, covering the north of the region (see figure 5.16). The difference in the spatial pattern on the signals is that there is a east west division in the north of the region. Cluster one corresponds to an increasing trend in the signal, therefore those species with positive weights in this region are increasing and those with negative weights are in decline. In cluster three the trend is also increasing, although the increase seems to begin earlier and the overall trend is shifted downwards. Cluster two exhibits an upwards trend and the spatial region overlaps with the spatial region for cluster three on the loading vectors, which implies those positively weighted species, such as the temperate species *Acartia spp* are increasing in abundance in the oceanic part of the North East Atlantic. The temporal trend also seems to be governed by latitude and sea depth

[123]. This could be related to the rate of warming, which varies across the North East Atlantic, and seems to be more pronounced in the southern North Sea than the rest of the region (see figure 3.1). the clusters on the first principal component, with the colour scheme as before. Clusters one and three on the signals overlap with clusters one and two on the loading vector (see figure 5.16), covering the north of the region. The difference in the spatial pattern on the signals is that there is a east west division in the north of the region. Cluster one corresponds to an increasing trend in the signal, therefore those species with positive weights in this region are increasing and those with negative weights are in decline. In cluster three the trend is also increasing, although the increase seems to begin earlier and the overall trend is shifted downwards. Cluster two exhibits an upwards trend and the spatial region overlaps with the spatial region for cluster three on the loading vectors, which implies those positively weighted species, such as the temperate species *Acartia spp* are increasing in abundance in the oceanic part of the North East Atlantic. The temporal trend also seems to be governed by latitude and sea depth [123]. This could be related to the rate of warming, which varies across the North East Atlantic, and seems to be more pronounced in the southern North Sea than the rest of the region (see figure 3.1). Figure 5.18 shows the regions based on the clusters on the second loading vector. They are less well defined in space but there is still a north-south divide. Clusters one and two cover most of the south of the region. *Echinoderm larvae* and *Oncaea spp.* have positive weights in cluster 1. In cluster 2 *Centropages typicus* and *Pseudocalanus spp. (Adult Atlantic)* have positive weights. Cluster 3 covers the north of the region and *Temora longicornis*,*Acartia spp.*, *Evadne spp.* and *Calanus finmarchicus* are all positively weighted in this region. The clusters on the time courses for the second component are poorly defined in space. There is some evidence of a east west divide. There is a slight upward trend in cluster two, a downward trend in cluster one and cluster three is dominated by an oscillation, perhaps attributable to the AMO, as the oscillation has a similar period to the AMO.

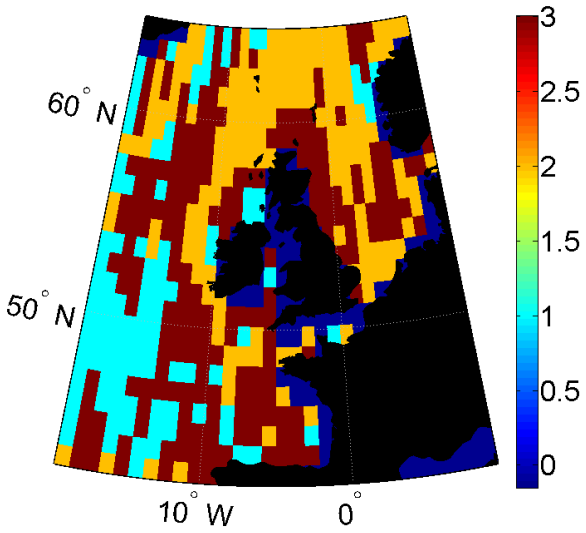


a) Regions based on the time courses for zooplankton species. Cluster one is blue, cluster two is orange and cluster three is dark red.



b) Time courses for each cluster, with cluster one shown in blue, cluster two in green and cluster three in red.

Figure 5.17: Plots of regions based on the time courses for the first component over all time.



Regions based on zooplankton species over all time.

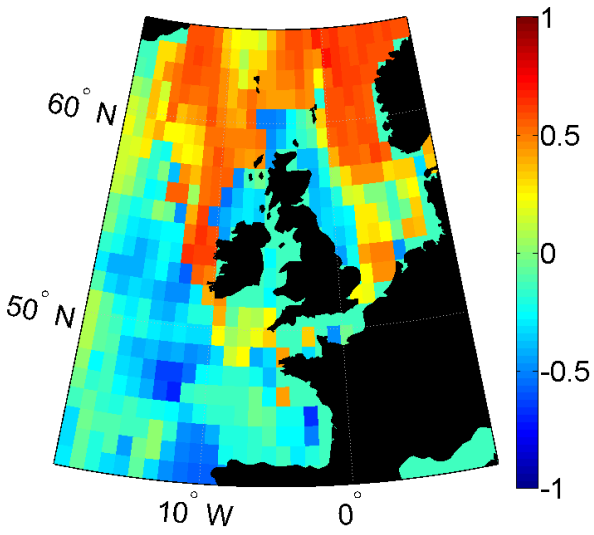
Figure 5.18: Plots of regions based on the second component over all time. Cluster one is blue, cluster two is orange and cluster three is dark red.

Figure 5.19 shows the Pearson's correlation coefficient between the first principal component and the NHT trend and the AMO. There is a strong positive correlation, in particular in the North, between the first principal component and the NHT. Comparing with the weights on two indicator species (see figure 5.20), it is possible to see that *Calanus finmarchicus* and *Calanus helgolandicus* show greatest variation in the northern North Sea, where the NHT has a positive relationship with the joint behaviour. *Calanus finmarchicus* has negative weights in this region, implying a negative relationship with temperature (i.e. the species is decreasing in abundance as temperature increases). The warm water copepod *Calanus helgolandicus* has positive weights, indicating the opposite relationship. This is supported by knowledge of the physiology of these two species [117]. As with the averaged trend, the NHT is a dominant driver of temporal behaviour of zooplankton species over most regions. The correlation with the AMO is also positive in the northern North East Atlantic and is strongest in the mixing region in the north along the oceanic shelf. This suggests that where mixing occurs the AMO has more of an influence

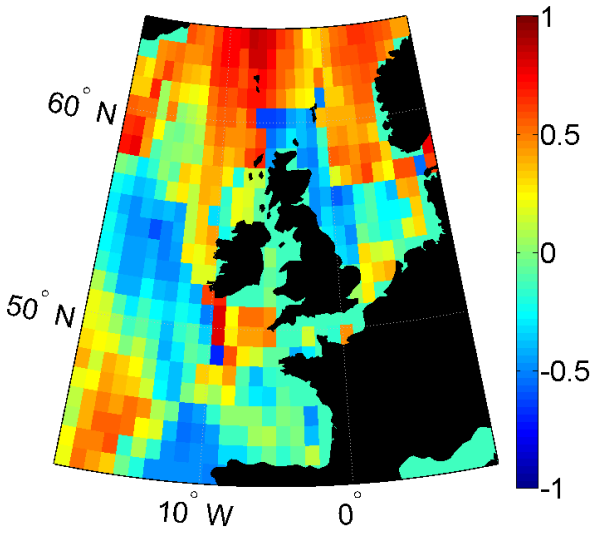
on the zooplankton group. Since phytoplankton production can be higher in mixing regions [68] and certain species of phytoplankton are known to respond to the AMO [56], this relationship between the zooplankton and the AMO in this region may be a response to changes in their food supply. For the second component the correlation with the NHT is weaker than for the first principal component, although there is still a positive relationship in the North Sea. There is also a positive relationship between the second principal component and the AMO and NAO across some regions, although the Pearson's correlation coefficient rarely has an absolute value greater than 0.3 suggesting these are only weak correlations.

5.5.3 Results for the Diatom Group

Figure 5.21 shows the sparsity parameters for the first principal component on the full time series restricted just to the Diatom data and the number of components. On principal component 1 the sparsity parameter is highest in the north east, whilst across subsequent principal components it shows little spatial structure. The number of components is highest in the southern North Sea and near coastal regions, indicating a higher degree of diversity in these regions. Some of the species of phytoplankton will prefer shallower waters and so will be more abundant in these coastal regions [56]. Figure 5.22 shows clusters on the real values of the loading vectors for the first principal component across the entire time course for the phytoplankton species. The real parts of the weights are almost universally positive, meaning that all species are tending towards behaving in the same way. The clusters divide the North East Atlantic in to three regions across longitude. The first cluster covers primarily the shallower waters, including coastal regions and the North Sea. Strongly weighted species in cluster one include the diatoms *Gyrosigma spp.*, a bottom-dwelling plankton that is often carried near to the surface in coastal waters [68]; *Guinardia delicatula*, which prefers temperate coastal waters [78]; and *Dactyliosolen fragilissimus*, which prefers northern temperate regions [78]. This

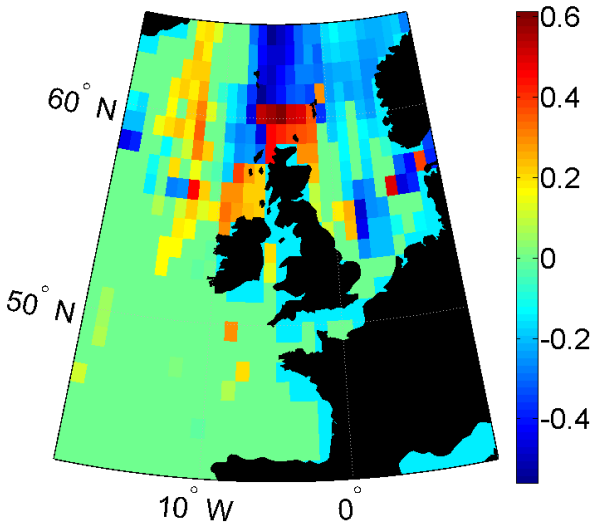


a) Regression against the NHT warming trend.

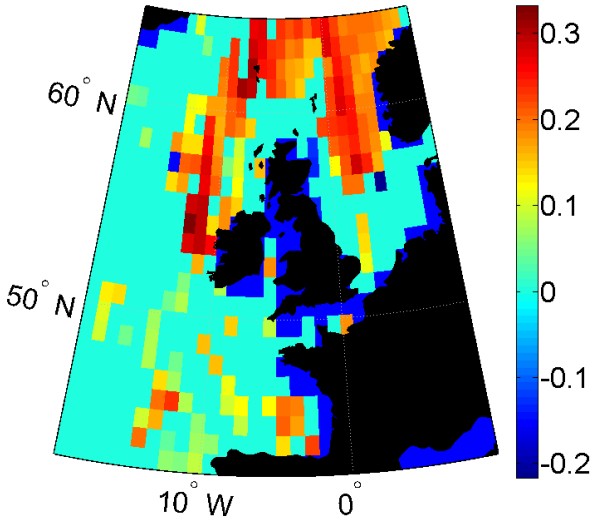


b) Regression against the AMO.

Figure 5.19: Plots of Pearson's correlation coefficient between the first principal component on the zooplankton species at each location and the climate indices.

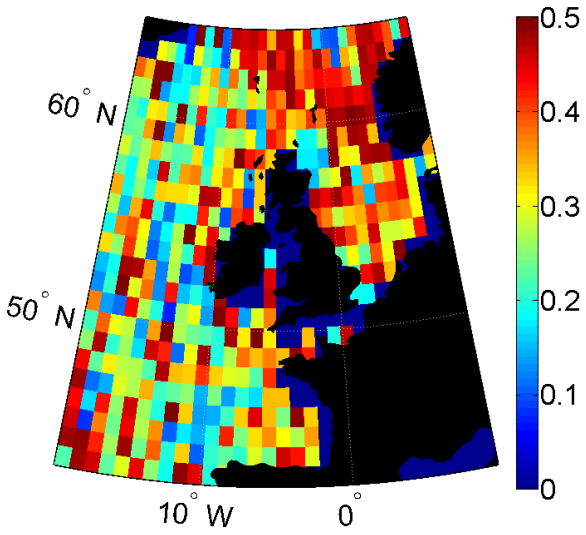


a) Real part of the loading vector for the first PC on *Calanus finmarchicus*.

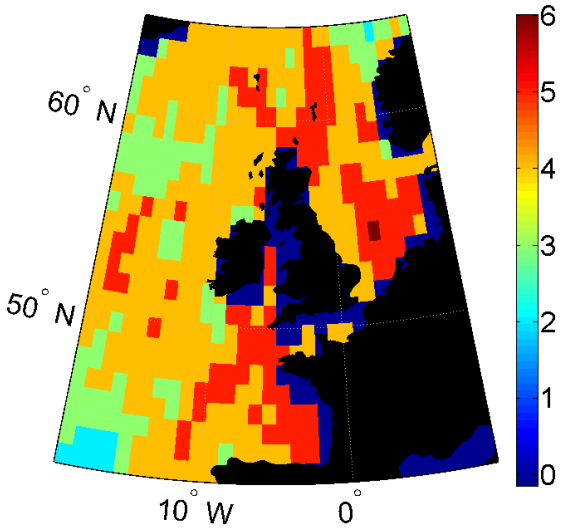


b) Real part of the loading vector for the first principal component on *Calanus helgolandicus*.

Figure 5.20: Plots of the real parts of the weights on two indicator species for the first principal component.

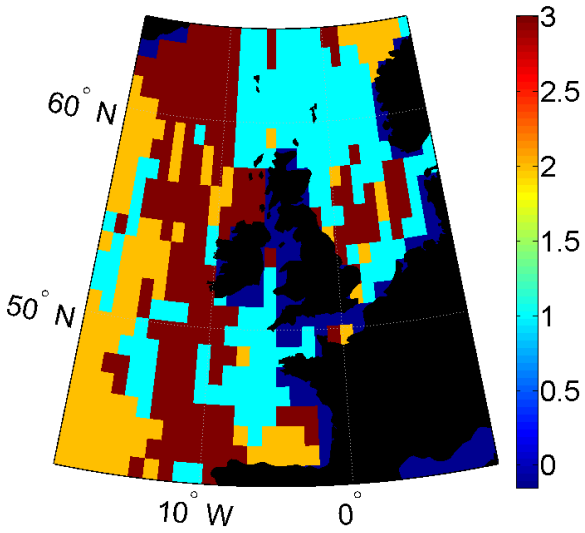


a) Sparsity parameter on Phytoplankton species and across all time for the first principal component.



b) Number of principal components for Phytoplankton species and across all time.

Figure 5.21: Sparsity parameter and number of principal components across Phytoplankton species for the entire time course from 1958 till 2009.



Regions based on Phytoplankton species over all time.

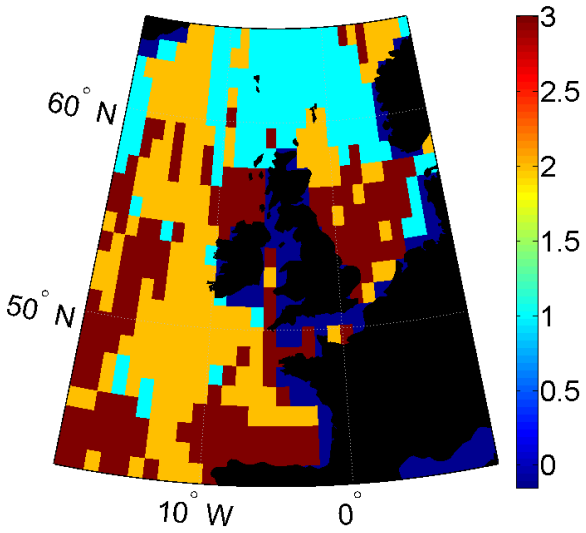
Figure 5.22: Plots of regions based on the first component on the phytoplankton over all time. Cluster one is blue, cluster two is orange and cluster three is dark red.

indicates that the shallower waters are dominated by temperate and coastal species. Cluster three occupies the central region between clusters one and two. The most strongly weighted species here are two species that prefer cosmopolitan environments [78]: *Skeletonema costatum* and *Asterionellopsis glacialis*. Cluster two primarily covers the open ocean to the very west. There are a mixture of different types of species that have large weights in this region, including: *Bacteriastrum spp*, a temperate species [78]; *Navicula spp.*; *Cylindrotheca closterium*, a cosmopolitan species [78], and *Pseudo-nitzschia seriata*, a cold water species known to produce harmful toxins [78]. When clustering on the first principal component the regions are slightly different, see figure 5.23. The first PC strongly resembles the behaviour of the AMO, with the correlation being significant over half of the locations even accounting for the false discovery rate. When comparing the correlation against the centre of each cluster, this relationship is strongest in cluster one, which occupies the north east of the North East Atlantic. The AMO explains 66.71% of the variation in the centre of cluster one with the median subtracted and the Pearson's

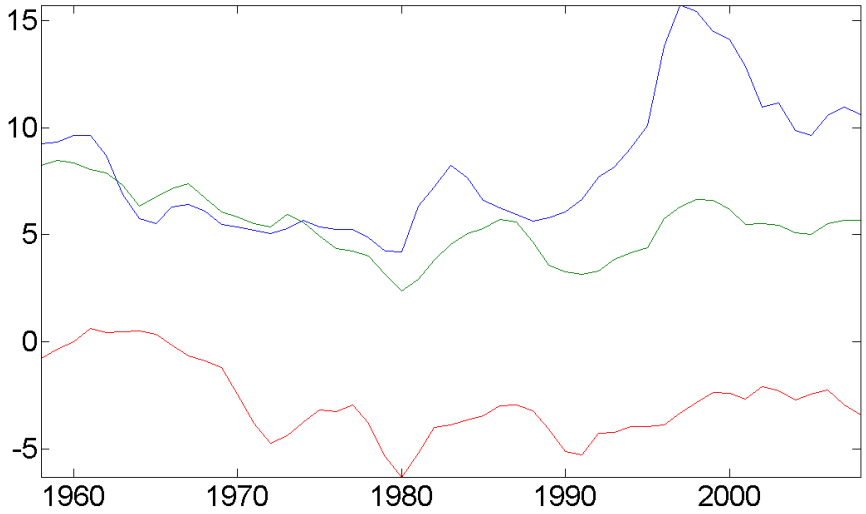
correlation coefficient between the two is 0.7834. This is judged to be significant, with a p-value of less than 0.0001. When decomposing the sea surface temperature using spatial PCA the AMO is shown to have the most positive weights in a similar region to cluster one, which may explain why it is more influential in this cluster. The correlation between the AMO and the centres of the other two clusters is still significant but the relationship is less strong, with a Pearson's correlation coefficient of 0.4164 in cluster two and 0.4329 in cluster three. On the second time course, the regionalisation is divided along east and west, with clusters two and three covering the eastern part and cluster 1 in the west. The time courses also correlate with the AMO somewhat, with this correlation being strongest in cluster two. Figure 5.24 shows the Pearson's correlation coefficient between the first principal component on the phytoplankton and the NHT warming trend and the AMO respectively. Though there is a slight relationship with the NHT, there is a much stronger positive correlation with the AMO. Even controlling for the false discovery rate [35] this correlation between the AMO and the first component is significant at around 275 locations. This means that for most of the North East Atlantic the AMO is the most important driver on the joint behaviour of the Diatom subgroup, which agrees with pre-exiting knowledge of their physiology [56]. The first component has some correlation with the NHT trend but the Pearson's correlation coefficient is smaller than for the AMO. The correlation between the AMO and the second principal component on the phytoplankton communities is positive across some regions but is slightly less significant than on the first principal component. The relationship is significant across 175 locations when the false discovery rate is controlled for.

5.5.4 Spatial Structure in Relation to Bathymetry

Figure 5.25 shows the depth of the sea across space. This is compared to the sparsity parameter, which represents the group size. For the phytoplankton group this has a reasonably significant positive correlation with the sparsity parameter on the first

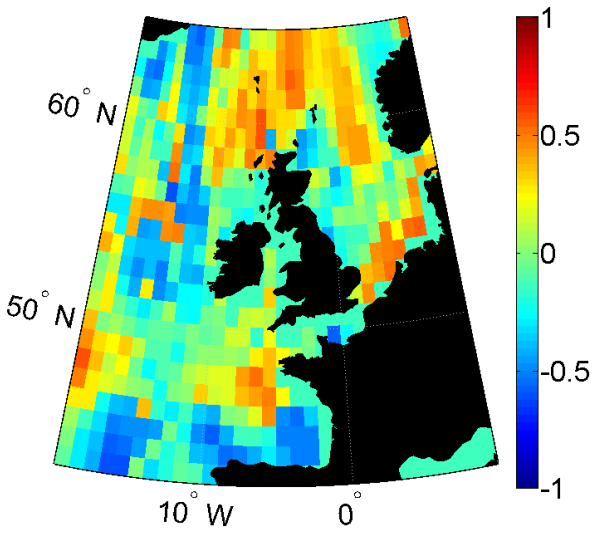


a) Regions based on the time courses for Phytoplankton species. Cluster one is blue, cluster two is orange and cluster three is dark red.

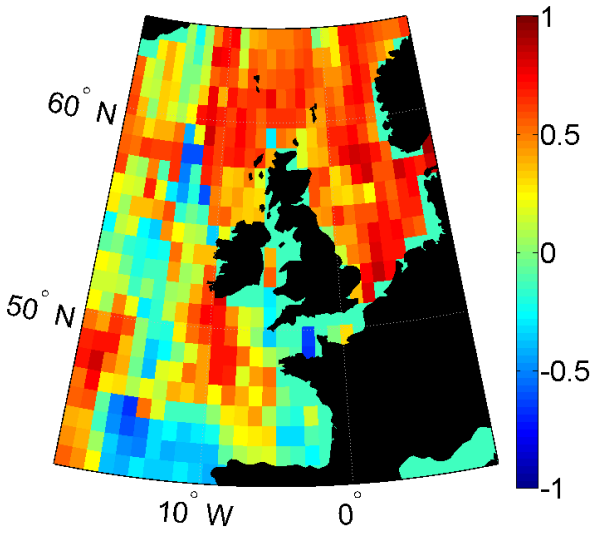


b) Time courses for each cluster.

Figure 5.23: Plots of regions based on the time courses for the first component over all time.



a) Regression against the NHT warming trend.



b) Regression against the AMO.

Figure 5.24: Plots of Pearson's correlation coefficient between the first principal component on the phytoplankton species at each location and the climate indices.

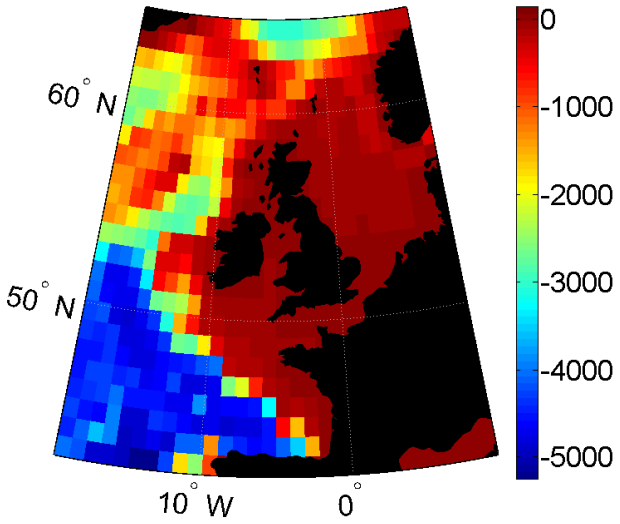


Figure 5.25: Spatial plots of the level below the sea surface, where a red pixel represents shallow waters and a blue pixel represents deeper waters.

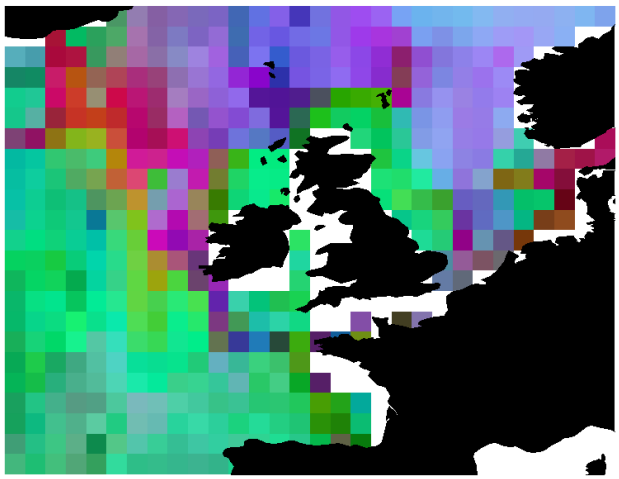
principal component (p-value of 0.0505), which means that where waters are shallower the first PC is less sparse. Likewise there is a positive correlation between the number of principal components on the phytoplankton species and the bathymetry, indicating that in shallower waters more components are required to adequately explain the variability. The Pearson's correlation coefficient for the AMO and the first principal component for the Diatoms correlates with the Bathymetry with a Pearson's correlation coefficient of 0.3868, which indicates that there is some relationship between depth and the influence of the AMO on plankton, possibly related to vertical mixing [68]. This indicates that the AMO influences the Diatoms more strongly in shallower waters.

5.5.5 Mixing Regions as Described by Colour Plots

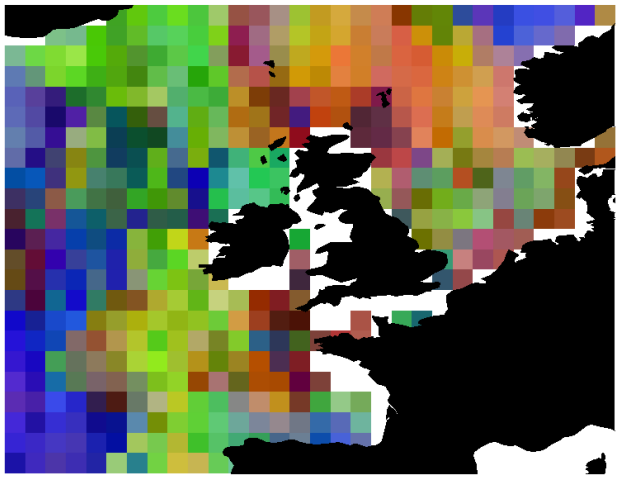
For ecoregions defined by species assemblages the divisions between regions might be blurred, i.e. the meeting point of two biogeographical regions may be a mixing region [161] and as has been shown in previous studies mixing regions have an

effect on the composition of plankton communities [68]. Clusters are first found using normal K-means on the real part (since PCA has been computed in the Fourier domain each loading will have a real and an imaginary part) of the loading vector and the centre of each cluster is found. Rather than assigning each location to a single cluster, proportional membership of each cluster at each location is found by taking the inner product for the real values of the normalised loading vector at that location with the normalised centre of each cluster. This inner product can take values between zero and one, with a value of one indicating the loading vector at that location is identical to the centre of that particular cluster. Since this inner product is found for each cluster, this produces a three dimensional vector at each location which can be represented in MatLab using the RGB (red, green, blue) colour scale. In this colour map the vector [1 0 0] gives a red pixel, [0 1 0] a green pixel and [0 0 1] a blue pixel. Therefore a pixel that is completely one of these colours indicates that the loading vector at that location is exactly the same as the centre of one of the clusters. If the loading vector at a fixed location has some agreement with more than one cluster the pixel will be a mixture of colours.

Figure 5.26 shows the RGB plot for the first loading vector on the zooplankton and the first loading vector on the phytoplankton. For the zooplankton the northern North Sea is mostly blue. There is a mixing region that is purple along the ocean shelf north of Ireland. The north west is covered by the red cluster, whilst the rest of the open sea is in green. It can be seen that the transition from the blue to the red cluster is continuous. The regions defined on the phytoplankton follow the bathymetry fairly closely. The North Sea is predominantly red, although there is some green in the very southern part. The ocean shelf is covered by a green cluster and the open sea is blue. At the edge of each cluster there is some mixing of colours, suggesting that there is a mixture of species groups at the transition between regions.



a) Clusters on the first loading vector for the zooplankton.



b) Clusters on the first loading vector for the phytoplankton.

Figure 5.26: RGB plots of the clusters on the real values of the first loading vector.

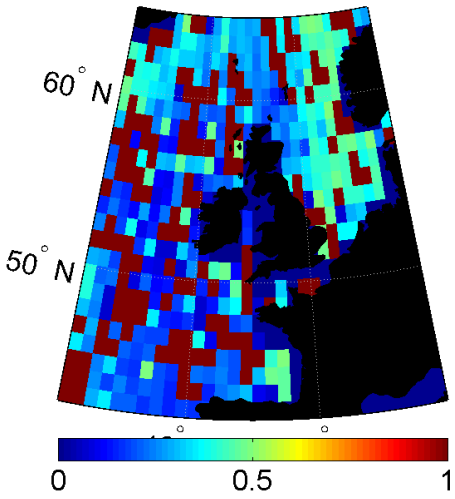
5.6 Simulation Studies

One remaining question that may arise is whether the spatial structure is genuinely an artefact of the dataset or if it is a result of the smoothing. This can be verified using a simulation study [168]. Data is randomised by selecting the time courses for each variable at each location to be a randomly selected weight from the output of the zooplankton analysis multiplied by a randomly selected principal component. Gaussian random noise with a standard deviation of 0.01 is added to the simulated data) The data is then smoothed as before using Kernel smoothing methods. Figure 5.27 shows the results of the analysis run on data simulated in this way. It can be seen that the sparsity parameter has little spatial structure, save for having a slightly higher value in the southern North Sea). There are also a number of areas where the resulting loading vector is not sparse, leading to a sparsity parameter taking the value 1. The clustering on the loading vector and the first principal component indicates no spatial structure, with the cluster that a location belongs to being seemingly random.

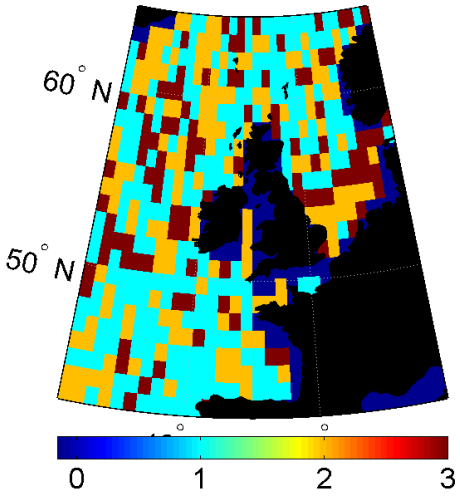
In order to find the footprint of the Kernel smoothing function a vector with zero values except for at a single pixel is interpolated on to the spatial grid. In this case the vector is chosen to be one over longitude $14.5°$W to $15.5°$W and between $54.5°$N and $55.5°$N and zero elsewhere. The spatial pattern at a single time point estimated from smoothing this data is shown in figure 5.28. This shows that pixels more than five degrees apart are more or less independent under this smoothing function. This again supports the suggestion that the spatial patterns are true features of the dataset rather than artefacts of the smoothing.

5.7 Discussion

In this chapter the structure of the CPR data has been investigated over a larger spatial region and further biologically interpretable results have been obtained. Both



a) The sparsity parameter for the simulated data)



b) Clusters on the real values of the first loading vector for the simulated data)

Figure 5.27: Spatial patterns on randomised data.

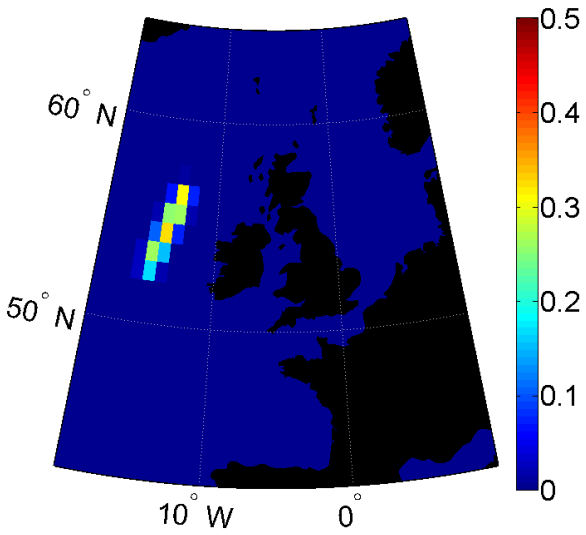


Figure 5.28: Plot of the footprint of the Kernel smoothing function.

spatial and species PCA are used to better understand the spatio-temporal variation within the data and the influence of climate variables is explored. This shows that the influence of climate varies both across space and species, with the warming trend being important for most species of zooplankton and the AMO being influential on the Diatoms. It can also be observed that the results are more interpretable when these two groups of species are considered separately. Whilst the spatial PCA is useful for gaining an understanding of individual species variation, species PCA provides an useful overview of joint variability across species groups. The spatially coherent patterns in both diversity and regions determined by species groups can be shown to be a result of the data rather than an artefact of the smoothing through comparison with simulation studies. Physical factors that appear to impact spatial structure include temperature, causing a north-south division; the presence of mixing regions, such as the bay of Biscay and the northern part of the continental shelf, and the bathymetry. Since these patterns are found from the ecological data without prior knowledge, it can be concluded that they are highly significant.

Chapter 6

Modelling Changes in Biogeographical Regionalisation

6.1 Overview

In previous chapters a statistical learning approach has been taken to find structure across space, time and species in the CPR data without prior knowledge of the ecology or the physical oceanography. When these results are compared with existing knowledge they are shown to be interpretable, in that they can be accounted for using pre-existing knowledge of the ecosystem. In this chapter an application of these techniques to a biological question, whether changes in the biogeographical regions of the North Atlantic as defined by the plankton have occurred over the past few decades, is explored. Various other studies have found a northwards shift in zooplankton species [30, 22, 132]. The strength of the analysis in this study is it allows one to gain an overview of the structure across all species. Ecoregions can be found by taking whole communities in to account. Since no previous studies have carried analysis across such a large number of taxa it is possible to gain new insights about the community structure related to biogeographical shifts.

Regionalisation is defined by K-means clustering on either the time course or the

weight vector, which are found using sparse species PCA. By dividing the dataset in to two periods, before 1985 and after 1985, it is possible to run the analysis again on each portion of the dataset. This in turn allows us to identify whether the ecoregions have changed.

6.2 Methods Used in this Chapter

This chapter follows directly from the previous one, using the same dataset which has been interpolated using Kernel Smoothing (section 2.1). The same methodology is used to find dominant species assemblages and their joint functional behaviour across space, namely sparse principal component analysis (section 2.3), except that this analysis is now carried out on the data restricted to before 1985 and the data restricted to after 1985 separately in order to assess how the structure has changed. As described in chapter 2, sparse principal component analysis will find joint behaviour of dominant or keystone species. This means that changes in the loading vectors will represent changes in the dominant species and changes in the time courses will represent changes in the joint behaviour of the dominant species. Biogeographical regions are defined on either the loadings, which recall represent ecoregions defined by dominant species assemblages, or on the time courses, in which case they represent regions defined by joint functional behaviour, using K-means clustering (as described in section 2.4). The main results of this chapter therefore can be used to assess how the regions might have changed over time. This can be interpreted as whether the regions of the North East Atlantic as defined by the dominant species of plankton have changed before and after 1985, i.e. has there been a 'regime shift' in the dominant plankton assemblages across space.

6.3 Modelling Changes across all Species

6.3.1 Time Courses and Sparsity Parameters Across all Species

In the first instance the 'regime shift' is explored using data for all species together, although later the zooplankton and the Diatoms will be considered separately. Before investigating changes in the ecoregions, one can look at the number of components, sparsity parameter and time course before and after 1985. The 'regime shift' might be observed as a non-stationarity in the dataset (i.e. a change in mean or variance of the signal) and this non-stationarity might exist in time or space. There might be changes in the behaviour of communities across time or a change in the spatial distribution.

Figure 6.1 shows the average principal components for the data across the entire time period and the data restricted to before and after 1985 for all species. The time courses before and after 1985 tend to follow closely the time course for the entire period. This indicates that there is no regime shift in the temporal behaviour on average across species. The first PC seems to on average be an oscillation, which resembles the AMO.

Figure 6.2 shows the sparsity parameter for the first principal component on the data before 1985 and the data after 1985. In the pre-1985 regime it tends to be slightly higher in the north west, whilst after 1985 it is slightly higher in the south. It is thought that on average species are moving north [30, 132], with warm water species increasing in abundance further north and cold water species decreasing. Whether this will lead to an overall increase in total numbers of organisms will depend on whether the rate of increase of warm water species exceeds the rate of decline of cold water species. One reason for the sparsity parameter in the north decreasing after 1985 is that cold water species may be disappearing from this region as the average temperature increases or moving yet further north in to Arctic waters [30]. This would imply that the rate of decline of cold water species is higher

than the rate of increase of warm water species in the northern North Sea, although some caution must be taken in interpreting these results as both phytoplankton and zooplankton have been included in this analysis.

The number of components (see figure 6.3) seems to be lower post-1985. This suggests that there are fewer distinct functional groups after this period. This may also be a result of the movement of cold water species away from the region, supporting the hypothesis that warm water species may not be increasing in abundance at the same rate as cold water ones are disappearing [30]. Another interpretation is that species are behaving in more similar ways after 1985, meaning fewer components are required to explain most of the variation.

6.3.2 Changes in Regionalisation Across all Species

Changes in the ecoregions of the North East Atlantic are explored using the output of the sparse PCA. The loading vectors are then used to define regions before and after 1985. The interpretation of the loading vectors is that they represent dominant species groupings and the sparsity constraint means that only those that contribute most to the total variation are included, i.e. only the most dominant species. Clustering on these finds spatial patterns defined by the dominant species groups before and after 1985 based on the dominant species groups. These patterns are not completely smooth in space, as can be expected since there will be some mixing at the boundary regions. Figure 6.4 shows the clustering on the real values of the first loading vector before and after 1985. There is a clear relationship with the bathymetry, with the shallower waters being covered by cluster two in both regimes, although this area is smaller after 1985. Cluster one covers most of the open sea. Cluster three before 1985 covers some very small regions, whilst afterwards it covers the area where cluster one and one meet. There are several species common to cluster one both before and after 1985, with *Paralia sulcata*, *Skeletonema costatum* and *Bacteriastrum spp.* having positive weights both before and after 1985 on

average in cluster one. There are also certain species that have strong weights in cluster 1 on average before but not after 1985: *Fraglaria spp.* and *Guinardia striata*. Likewise there are species that are strongly weighted after but not before: *Navicula spp* and *Cylindrotheca closterium*, which is common in coastal waters [78]. The two regimes are comparable because the time courses follow the average time course for the whole dataset. Therefore it is possible to conclude that there has been some changes in the types of species occurring in these shallower waters. In cluster one, the open sea, there is a high degree of overlap in which species have strong weights both before and after. *Odontella sinensis*, *Thalassiothrix longissima*, *Ditylum brightwelli* and *Eucampia zodiacus* all have strong positive weights. *Eucampia zodiacus* is a cosmopolitan species, which is found in many regions save for polar waters [78], which explains its presence in the open sea cluster. *Asterionellopsis glacialis* has a strong weight before but not after. This leads to the conclusion that there has been relatively little change in the types of species in the open sea. The regionalisation based on the first loading vector has changed relatively little pre and post 1985, which may be due to having failed to distinguish spatial patterns due to combining species that are recorded differently [160].

Figure 6.5 shows the clusters on the real value of the second loading vector across all species. Cluster one covers the shallower waters in both the pre-1985 and the post 1985 data. There is not much change in those species that have large weights on average in this region from before to after 1985. Those that have large weights in cluster one include: *Paralia sulcata*, *Skelentonema costatum*, *Ditylum brightwellii*, *Eucampia zodiacus* and *Fragilaria spp.*. Clusters two and three cover the open ocean. A few zooplankton species have non-zero weights both before and after 1985, namely *Evadne spp.*, *Podon spp.* and *Echinoderm larvae*, but the weights are mostly dominated by phytoplankton species. *Talassiothrix longissima* and *Bacteriastrum spp.* have large weights both before and after 1985 in the open sea. A few species have large weights after 1985 but not before, indicating some

changes in the composition of species over the 'regime shift'. These include *Ditylum brightwellii* and *Guinardia delicatula*. As with the first component there is some change in species before and after 1985 but not in the spatial patterns, indicating again the need to separate different species types. The regionalisation based on the first principal component is less structured. There is a some suggestion of a north south divide in the regionalisation, which may be attributable to the different trends in the phytoplankton between the north and the south [99], with little change before and after 1985. The regionalisation based on the second component is even less structured, although there is a slight east-west gradient.

Across both components the regionalisation based around the species remains similar in both regimes, being governed by the bathymetry. There is greatest change in species in the shallower waters, suggesting a greater influence of climate variables in these areas. When both zooplankton and phytoplankton are considered together the phytoplankton have the strongest weights, suggesting that it may be useful to consider the two groups separately. The higher weighting of phytoplankton species is likely due to the fact that abundance rather than biomass is considered and so the two groups are not comparable [160, 99]. There is little change in the regionalisation based on the temporal behaviour and since the time courses before and after follow the time courses for the whole dataset the two regimes are directly comparable.

6.4 Changes in the Regionalisation for the Zooplankton

6.4.1 Non-Stationarity in the Time Courses for the Zooplankton

Since there was little change in the regionalisation when all species were considered together, it is of interest to determine whether the same holds true when different groups are considered separately. In earlier chapters it appeared that when all species were considered together the output was dominated by phytoplankton

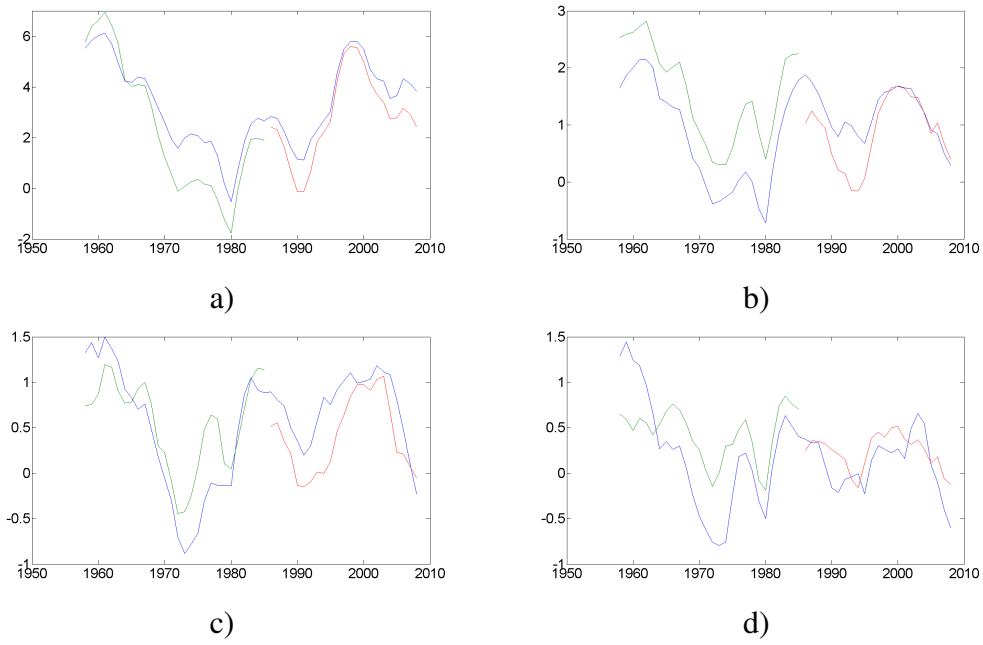
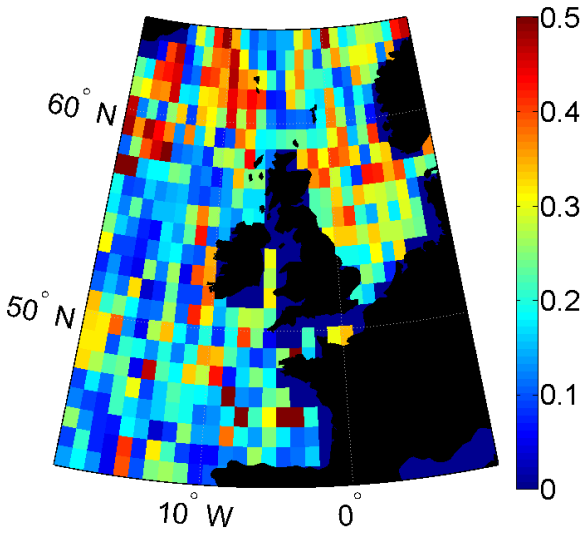
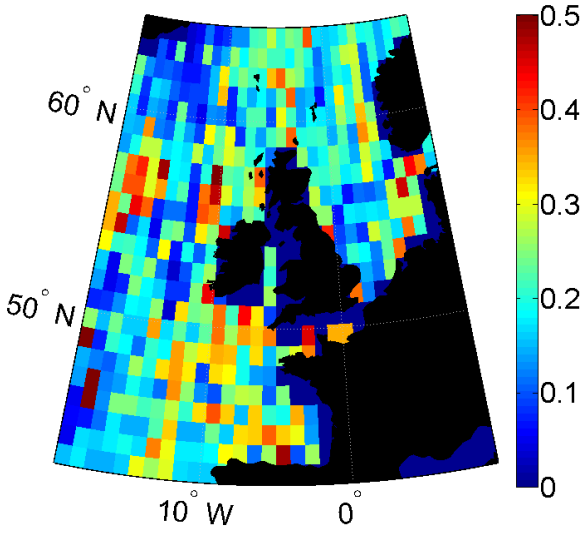


Figure 6.1: Principal components all species for the entire time course from 1958 till 2009. Time course for the full dataset is shown in blue, before 1985 in green and after 1985 in red. a) Averaged first principal component across all time and for each half of the data before and after 1985 for all species. b) Averaged second principal component across all time and for each half of the data before and after 1985 for all species. c) Averaged third principal component across all time and for each half of the data before and after 1985 for all species. d) Averaged forth principal component across all time and for each half of the data before and after 1985 for all species.

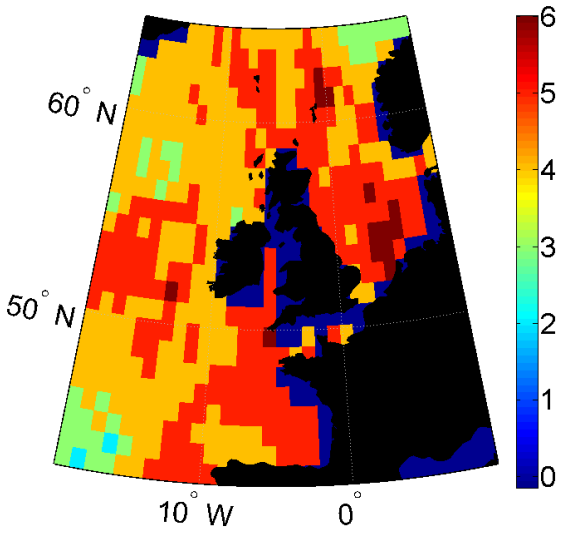


a) Sparsity parameter on all species before 1985.

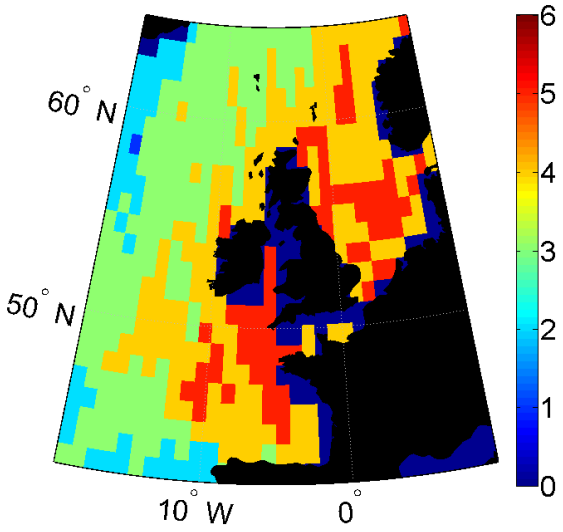


b) Sparsity parameter on all species after 1985.

Figure 6.2: Sparsity parameter across all species. There is a decrease in the sparsity parameter in the North West and a slight decrease in the sparsity parameter in the North Sea.

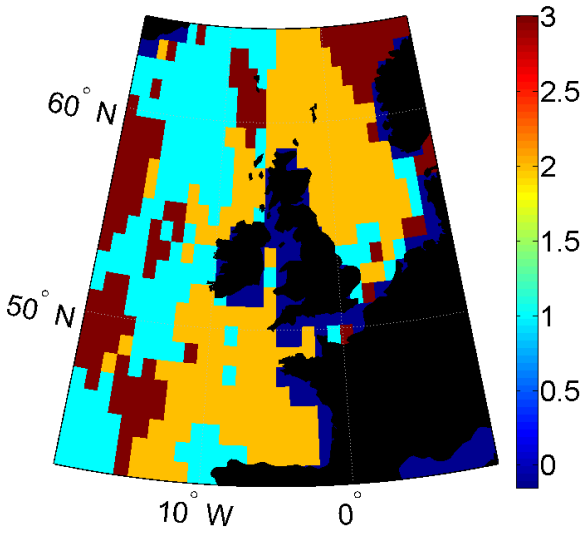


a) Number of principal components on all species before 1985.

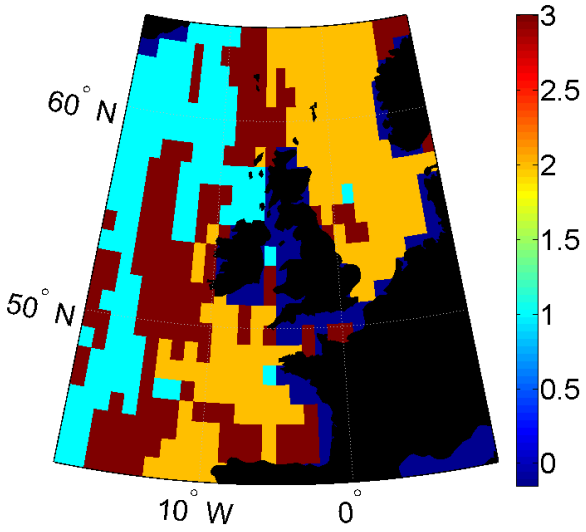


b) Number of principal components on all species after 1985.

Figure 6.3: Number of PCs on across all species.

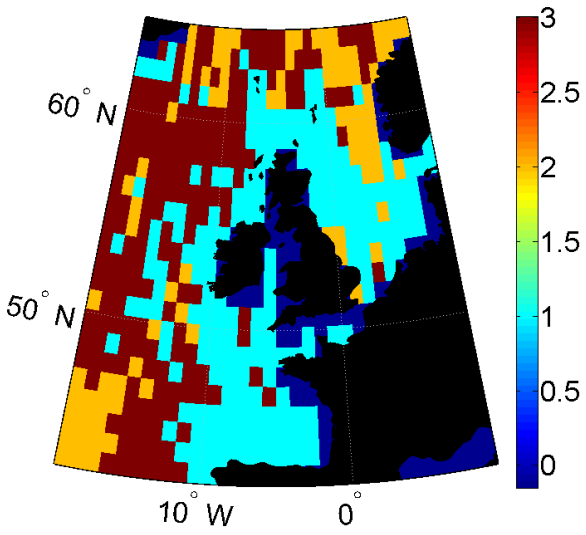


a) Regionalisation based on all species before 1985.

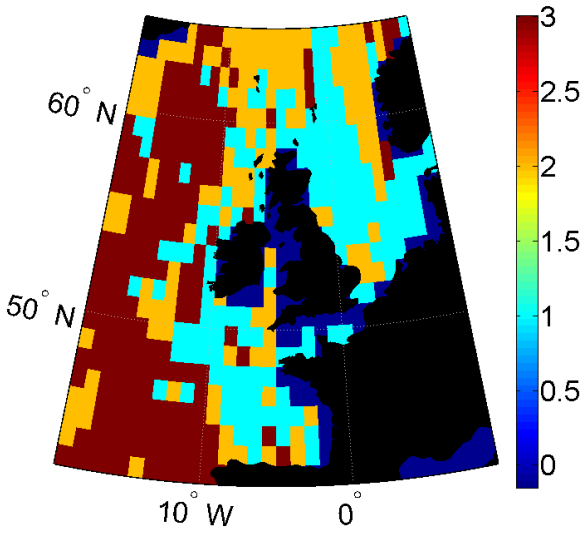


b) Regionalisation based on all species after 1985.

Figure 6.4: Plots of regions based on the first loading vector before and after 1985.



a) Regionalisation based on all species before 1985.



b) Regionalisation based on all species after 1985.

Figure 6.5: Plots of regions based on the second loading vector before and after 1985.

species, which further implies that it might be necessary to consider zooplankton and phytoplankton separately. Figure 6.6 shows the time courses for each principal component averaged over all locations for the entire time series in blue, the principal components on the data restricted just to before 1985 in green and after 1985 in red. For the first PC the average time series is oscillatory, with some evidence of a decline in time, and has a minimum in the early 1970's. In chapter 5 it was determined that the first PC on the zooplankton data correlated with the NHT across most locations. The average of the first principal component restricted just to the first half of the data follows the full time series quite closely. After 1985 there is a shift in the regime and the average first PC instead represents a decline in time. The behaviour of the first PC after 1985 is similar to the second PC on the whole time course, indicating that this 'regime shift' manifests as a change in the ordering of the components, i.e. that a trend that was of only secondary importance previously has now become the most important trend. The interpretation of this is that there is a non-stationarity in the time series for the zooplankton species. If this is considered in terms of a 'regime shift' it might be hypothesised that after 1985 there is a significant change in the dominant behaviour of the species, perhaps driven by some climate driver. From an ecological viewpoint this is important because it shows that there has been a significant change in the temporal behaviour of the zooplankton species after 1985. An explanation of this is that after 1985 some physical variable that is driving zooplankton abundance has become more prominent. One suggestion is that prior to 1985 natural climate variability might play a larger role in determining the abundance, whilst afterwards the influence of sea surface temperature warming becomes more important than natural variability in determining the abundance of zooplankton. If this is the case then this change indicates a dramatic shift in the ecosystem. The strength of the decline after 1985, for instance, in abundance might be reflective of the northwards migration of cold water species out of the North Atlantic and in to more Arctic regions [30, 22]. Subsequent components

exhibit a similar pattern, with the component on the first half of the data following closely the component on the full time series and then the component on the second half behaving differently. PC 2 on the full time series is oscillatory prior to the mid-1980's and after this shows a steep decline. On the data restricted to just after 1985 there is a switch in the ordering of the components and this decline becomes the most important trend. There is some debate over whether changes in the behaviour of the plankton should be viewed as 'regime shifts' or trends [146]. The change in ordering of the PCs suggests it is a shift in the importance of functional groups. The change in the ordering of the principal components from the pre-1985 data to the post-1985 data can be viewed as a non-stationarity in the time series, which is indicative of a 'regime shift' rather than a trend.

Figure 6.7 shows the sparsity parameter for the first principal component over space for the zooplankton data before 1985 and after 1985. Before 1985 the sparsity parameter is larger in shallower regions, particularly in the North Sea, whilst after 1985 the sparsity parameter is less structured in space. This could be because the dominant trend in the first half of the data follows a different 'regime' to the second half, with shallow water species being more important prior to 1985. It is also thought that certain cold water species are moving away from the North Sea [30, 22, 74, 132], which may explain the decrease in group size in this region. It has also been suggested that the presence of an ecological niche for many species in the North Sea could cause dramatic shifts at critical temperatures [23, 75, 23, 21], which might explain the decrease in the group size. There is also a decrease in the number of principal components required to explain most of the variation in most regions in the post-1985 regime (see figure 6.8). This indicates a decrease in the number of separate functional groups, again perhaps attributable to the decline in cold water copepods at certain thermal thresholds [21]. Where the number of components remains high post-1985 is in the fertile region in the bay of Biscay [57]. Before 1985 the numbers of components show some spatial structure, typi-

cally larger numbers of components are required to explain most of the variability in shallower and warmer waters. After 1985 the number of components is higher in the Bay of Biscay than most of the rest of the region in this time period. In general only a few components are required in the North West of the North East Atlantic after 1985. The numbers of components after 1985 have in general decreased, indicating perhaps there are fewer distinct functional groups. An interpretation of this result is that after 1985 the species 'behave together' more frequently, meaning that they are responding to the same climate trend so fewer components are required to explain most of the variation. This implies that in the latter half of the dataset there is an overriding effect that is driving the behaviour of most of the zooplankton, which as has been discussed might be the warming trend. The warming trend is present across the entirety of the North East Atlantic [21, 97] and so will have an influence on zooplankton assemblages across the region. It may also be speculated that there is a critical temperature threshold at which ecosystem changes might occur. For some species the North Sea is at the edge of an ecological niche and so small perturbations in temperature might lead to drastic changes in behaviour [75]. Prior to this there is no dominant driver and so the species are behaving in more distinct ways, perhaps responding instead to natural variability which may have a more heterogeneous effect. If there is a dominating trend after 1985 then this provides further evidence of a shift in the behaviour of the zooplankton.

6.4.2 Changes in Regionalisation for the Zooplankton

The next step is to consider changes in the regionalisation based on the species groupings for the zooplankton data. Figure 6.9 shows regions defined by clustering on the absolute values of the loading vectors for the first component before and after 1985. Cluster one is depicted by light blue pixels, cluster two is orange and cluster three is dark red. On the pre-1985 dataset the clusters divide the North East Atlantic in to three regions: clusters one and two covering the open ocean and cluster three

mostly covering shallower waters and the very north of the region. *Acartia spp.*, *Centropages typicus*, *Podon spp.* and *Evadne spp.* are all strongly weighted species in cluster one before 1985. *Acartia spp.* and *Centropages typicus* are both species of Copepod, the former is known to prefer temperate waters whilst the latter lives in coastal waters [78, 68]. After 1985 cluster one covers almost all of the open ocean region. *Acartia spp.* is still strongly weighted in the oceanic region but *C. typicus* has a much smaller weight. *Oithona spp.* is more strongly weighted in the oceanic cluster in the post-1985 regime. *Oithona spp.* is a cyclopoid copepod species, which prefers brackish waters and having late blooms [68, 78]. Changes in sea surface temperature in the North Atlantic may have benefited this species, as warmer conditions now occur earlier in the year [26]. The species *Evadne spp.* is strongly weighted in this cluster both before and after 1985. A few species have strong weights in cluster one after but not before 1985, including: *Oithona spp.*, *Copepod nauplii* and *Pseudocalanus (Adult atlantic)*. The latter two of these are small copepods [68, 78]. This indicates that the changes in these species have become more significant in the open sea after 1985.

Cluster two in the pre-1985 regime covers the central North Atlantic. Highly weighted species in this region pre-1985 are *Pseudocalanus (Adult atlantic)*, *Podon spp.*, *Copepod nauplii* and *Echinoderm larvae*. After 1985 this cluster is displaced northwards and now covers a larger potion of the southern North Sea. *Pseudocalanus (Adult atlantic)*, *Podon spp.* and *Copepod nauplii* all still have strong weights after 1985 but *Echinoderm larvae* is less strongly represented in this region. In addition *Temora longicornis*, *Centropages typicus* and *Evadne spp.* have strong weights after 1985 but not prior in cluster two.

Both before and after 1985 *Centropages typicus*, *Oithona spp.*, *Copepod nauplii* and *Pseudocalanus (Adult atlantic)* have weights with a large magnitude in cluster three. Cluster three covers the northernmost part of the region in both regimes but has shifted further north after 1985, which is most likely attributable to the north-

wards movement of species [30, 22, 75]. Both *C. finmarchicus* and *Echinoderm larvae* have weights with a large magnitude post-1985. This indicates that these species are changing significantly in this region after 1985. It can be shown by a spatio-temporal analysis of the species that *Echinoderm larvae* are increasing in the north of the North East Atlantic, whilst *C. finmarchicus* by contrast are decreasing in this region. This is reflected in the signs of the real part of the loading vector, which is on average positive for *Echinoderm larvae* in this region and negative for *C. finmarchicus*. The decline of *C. finmarchicus* can be explained by rising sea surface temperatures [74, 132]. It has been previously shown in our analysis that the northern region is where the greatest changes are occurring in abundance of *Calanus finmarchicus* and that it is declining as temperatures rise. After 1985 there is a general northwards movement of the clusters, with the oceanic region being covered by a single cluster post-1985. This indicates more homogeneity in the open ocean in terms of species after 1985. Since the first component before and after 1985 follow different time courses it is clear that there is a regime shift in the data, with different functional groups becoming more important post 1985. Clusters on the absolute values of the second loading vector are less structured in space. There is some north-south divide, with clusters two and three mostly covering the north of the region and cluster one being mostly in the south.

Figure 6.10 shows clusters defined on the first principal component for the data before and after 1985 and the centres of these clusters. Before 1985 cluster one covers the north east, cluster two covers shallower waters around the oceanic shelf and cluster three covers the open ocean. The temporal trend in the north east before 1985 is oscillatory and slightly increasing. Cluster two is covered by a steady trend and cluster three a declining trend. After 1985 cluster one covers most of the shallower waters covered by cluster two pre-1985. Instead of remaining constant the trend now peaks, then declines in this cluster. In cluster two, which overlaps with cluster three pre-1985, the trend continues to decline. Finally in cluster three,

which covers the bits to the north of the region, the temporal trend is increasing. Clusters on the second principal component before and after 1985 have very little spatial structure. The difference between the regions defined on the temporal trend may be due to the fact that not all regions are warming at the same rate (see figure 3.1) and the fact that the zooplankton's response to the warming trend is heterogeneous across space [108]. Different regions are dominated by different species groups, all of which have differential responses to climate. Matching clusters on the time courses before and after 1985 is trickier than matching clusters on the loading vectors because inner products can not be used to determine whether it represents the same trend. As such this clustering may be less interpretable than that carried out on the weight vector.

6.5 Changes in the Regionalisation for the Phytoplankton

6.5.1 Time Courses and Sparsity Parameters for the Diatoms

The same analysis is now carried out restricted to the Diatom species only. Figure 6.11 shows the time course for the first four PCs on the phytoplankton for the full dataset in blue, restricted to before 1985 in green and after 1985 in red. The most notable feature is that unlike for the zooplankton there is no evidence of a regime shift in the PCs, as both the pre and post 1985 segment follow the time course for the full dataset closely. The first principal component resembles the Atlantic Multidecadal Oscillation [141, 43], as it is an oscillation with a minimum about 1980. In chapter 5 the Pearson's correlation coefficient between the first PC on the phytoplankton and the AMO was show to be statistically significant over a large number of locations. The analysis here shows that this holds both before and after 1985. Subsequent PCs appear to be declining in time, suggesting climate warming may be a driver but is a less important influence than the AMO.

Figure 6.12 shows the sparsity parameter on the first component before and after

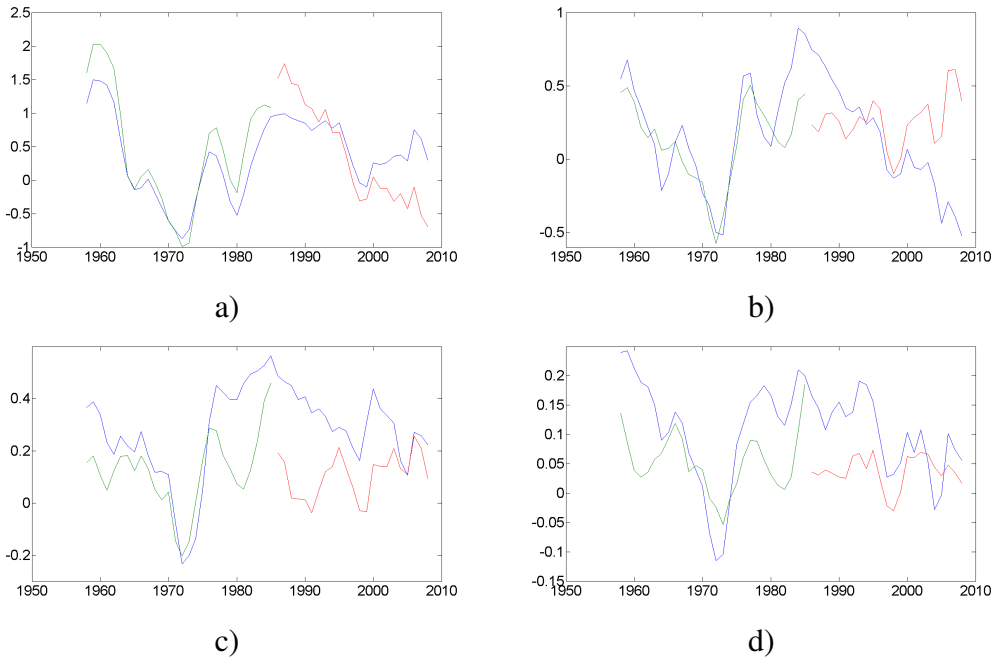
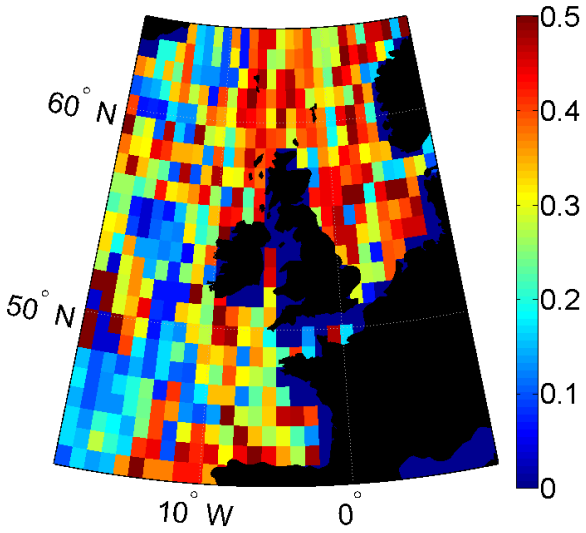
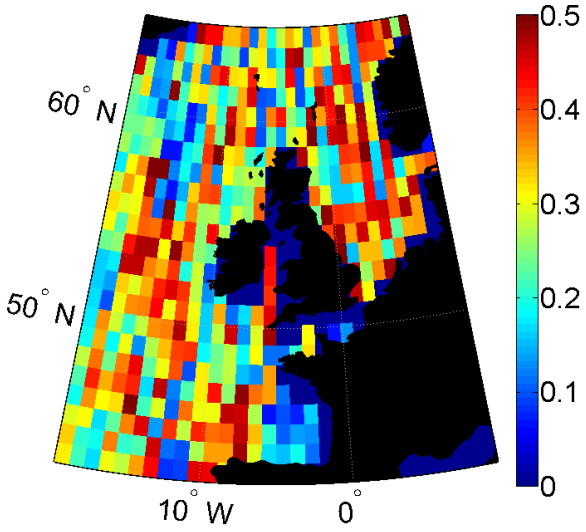


Figure 6.6: Principal components zooplankton species for the entire time course from 1958 till 2009. The time course for the entire period is shown in blue, before 1985 in green and after in red. a) Averaged first principal component across all time and for each half of the data before and after 1985 for zooplankton species. b) Averaged second principal component across all time and for each half of the data before and after 1985 for zooplankton species. c) Averaged third principal component across all time and for each half of the data before and after 1985 for zooplankton species. d) Averaged forth principal component across all time and for each half of the data before and after 1985 for zooplankton species.

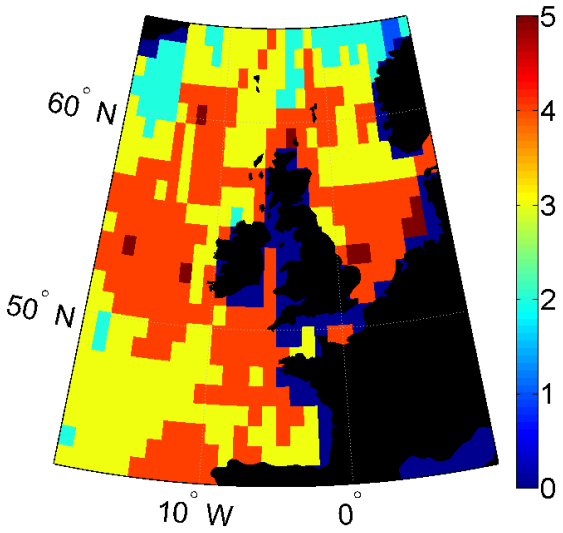


a) Sparsity parameter on zooplankton before 1985.

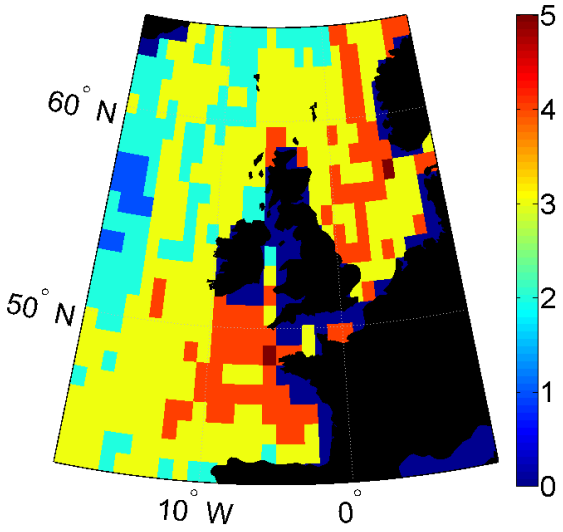


b) Sparsity parameter on zooplankton after 1985.

Figure 6.7: Sparsity parameter across zooplankton species.

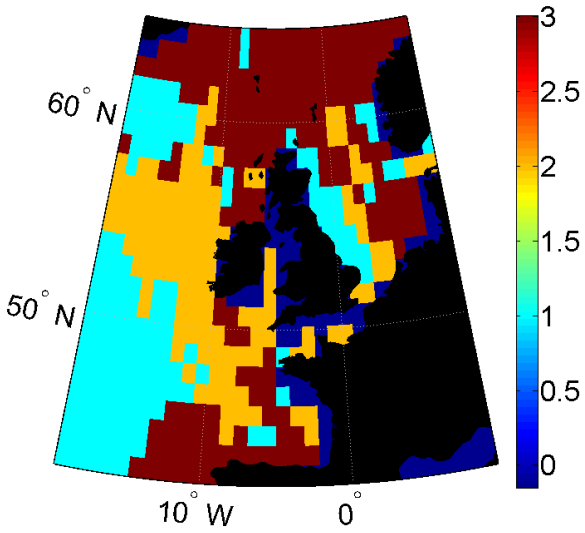


a) Number of principal components on zooplankton species before 1985.

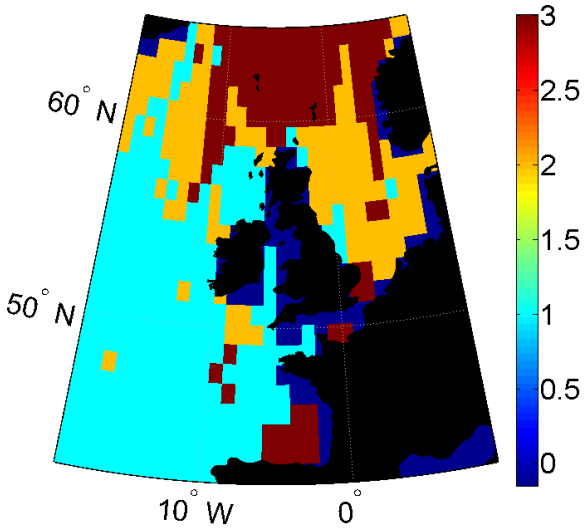


b) Number of principal components on zooplankton species after 1985.

Figure 6.8: Number of PCs on across zooplankton species.

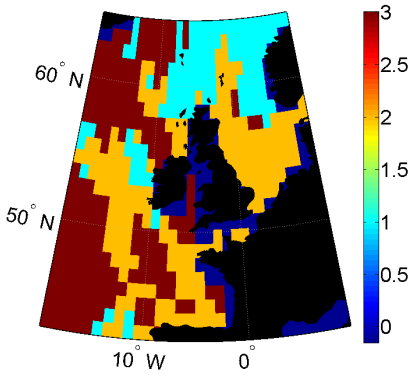


a) Regionalisation based on zooplankton species before 1985.

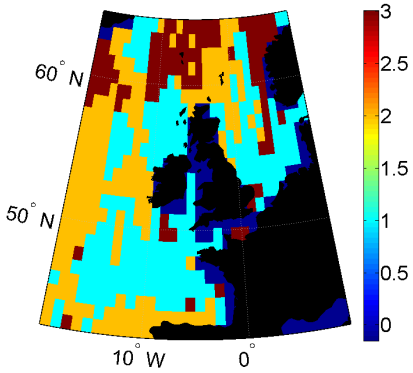


b) Regionalisation based on zooplankton species after 1985.

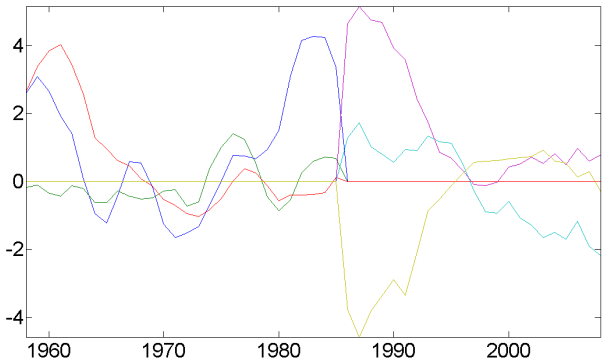
Figure 6.9: Plots of regions before and after 1985.



a) Regionalisation based on the time courses for zooplankton species before 1985.



b) Regionalisation based on the time courses for zooplankton species after 1985.



c) Time courses for each cluster before and after 1985.

Figure 6.10: Plots of regions based on the time courses for the first component before and after 1985.

1985. There is little spatial structure but there is on average an increase after 1985, predominantly in shallow waters as the AMO enters its positive phase [141]. The number of principal components (see 6.13) is lower after 1985 than before. There is spatial structure both before and after, with more PCs being required to explain the variation in shallower waters. Pre-1985 the highest group sizes are found in mixing regions, such as the north of Scotland, and in fertile areas, such as the bay of Biscay and the southern North Sea. After 1985 it appears to be the bathymetry that governs the number of principal components. The decrease in the number of PCs in the post-1985 regime indicates a decrease in the number of distinct functional groups, which in turn may indicate a decrease in diversity. Since the group size on the first PC has increased after 1985, it can be deduced that the species are behaving in more similar ways in the positive phase of the AMO, whilst there are more distinct functional groups during its low phase.

6.5.2 Changes in Regionalisation for the Diatom Species

The regionalisation based on the first loading vector on the Diatom species appears to change after 1985. Figure 6.14 shows clusters on the real value of the first loading vector before and after 1985 for the phytoplankton communities. As the matching of the clusters was less well defined than for the zooplankton species, different colour schemes are used pre and post 1985 to indicate that it is not clear whether regions match. The clusters after 1985 are more spatially structured than those before, with the region being clearly divided into oceanic and shallower waters. Since the behaviour of the first principal component reflects the AMO, then changes in the regionalisation must be a response to the AMO. Between the start of the dataset and 1985 the AMO is declining and afterwards it is increasing. The positive phase of the AMO is thought to influence wind intensities [90] and it has already been demonstrated that its influence on the diatoms may be stronger in shallower waters (figure 5.24). Recalling from chapters 3 and 5 that although the AMO is not a

strong driver of sea surface temperature in the North East Atlantic but nevertheless has an indirect effect on the abundance of diatoms, this suggests that there is some other mechanism that governs the relationship between diatoms and the AMO in shallow waters. One potential explanation is that in the absence of sufficient nutrients diatom species can enter a dormant state and sink to the bottom of the sea [68]. Water column mixing is required to return them to the plankton. Since this only occurs in shallower waters this might explain why mixing has a greater impact on diatom abundance in shallow waters. However the mechanism is not well understood and further exploration would be required to verify this hypothesis. Together this suggests that the spatial structure after 1985 is governed by the differential spatial influence of the AMO on wind and currents [56, 141, 90]. The regionalisation based on the second loading vector is less well defined, although there is some differentiation between the deeper and shallower waters before 1985.

6.6 Discussion

In this chapter it was shown how the methodology developed can be adapted to answering questions of biological importance, in addition to being used as an exploratory tool. The principal result in this section is the existence of non-stationarity in the dataset, either in the temporal behaviour of the species assemblages or in the ecoregions defined by groups of species. This provides evidence in support of the hypothesis that a 'regime shift' has occurred in the CPR data around the mid-1980's. For the zooplankton this manifests in a change in the dominant temporal behaviour and a shift in the spatial patterns, with a northwards movement of species groupings occurring. For the diatom species although there is no shift in the dominant behaviour in time, there is increased structuring of the spatial regionalisation, which seems to follow the pattern of the bathymetry. Since these species appear to be responding to a natural oscillation, namely the AMO, this restructuring may also

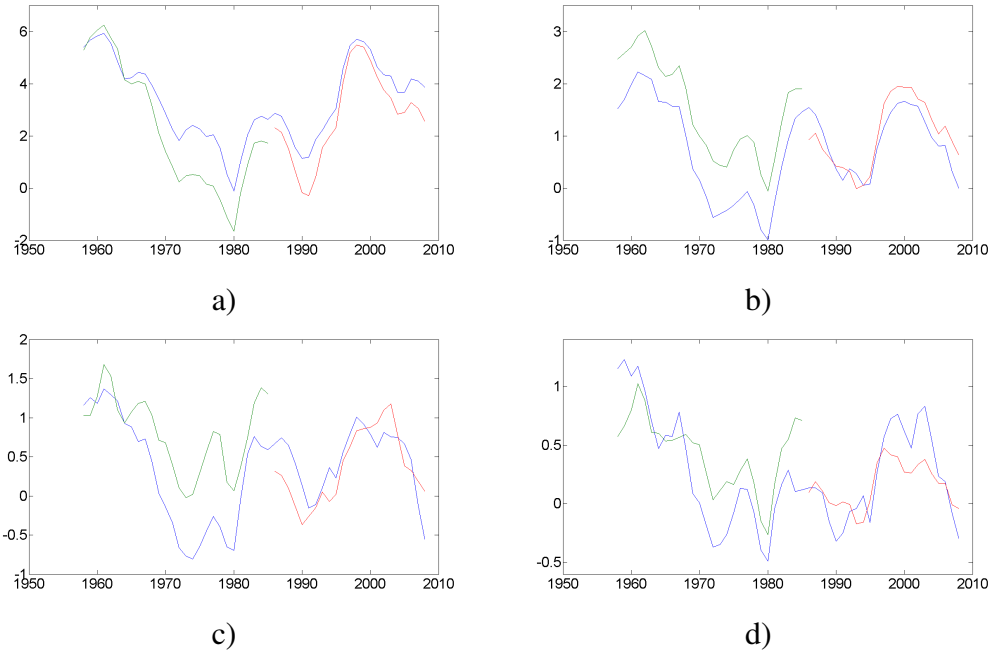
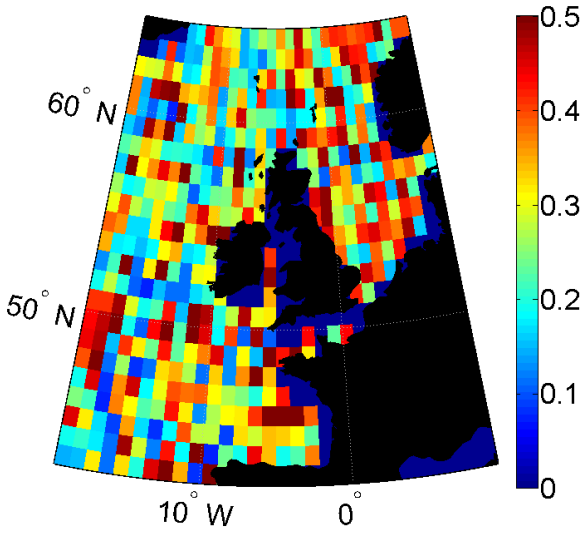
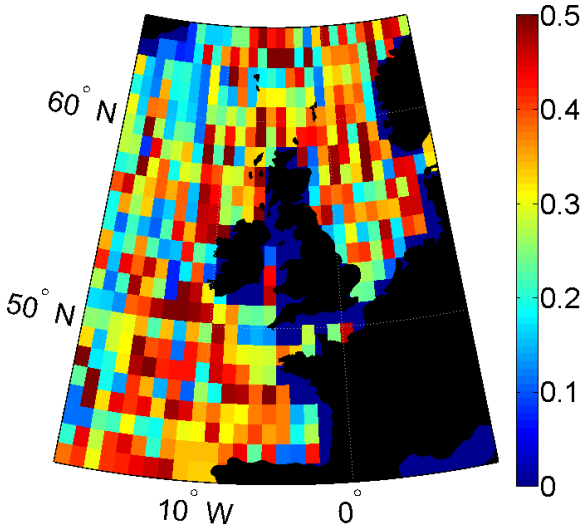


Figure 6.11: Principal components Phytoplankton species for the entire time course from 1958 till 2009. The time courses for the full dataset are shown in blue, before 1985 in green and after 1985 in red. a) Averaged first principal component across all time and for each half of the data before and after 1985 for Phytoplankton species. b) Averaged second principal component across all time and for each half of the data before and after 1985 for Phytoplankton species. c) Averaged third principal component across all time and for each half of the data before and after 1985 for Phytoplankton species. d) Averaged forth principal component across all time and for each half of the data before and after 1985 for Phytoplankton species.

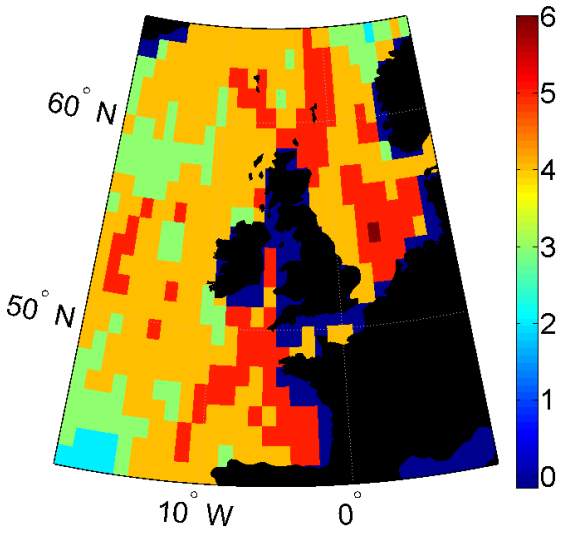


a) Sparsity parameter on Phytoplankton before 1985.

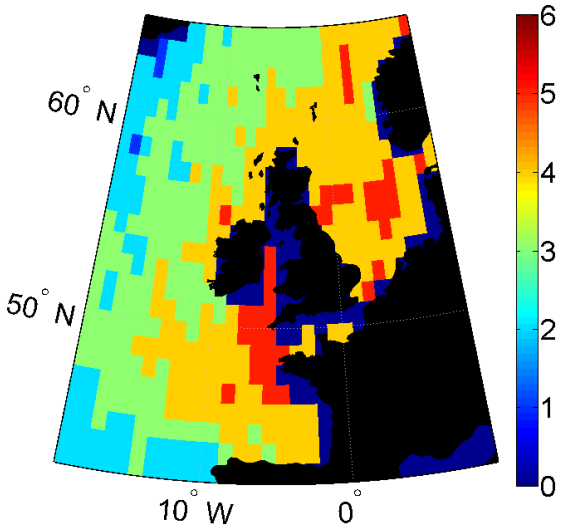


b) Sparsity parameter on Phytoplankton after 1985.

Figure 6.12: Sparsity parameter across Phytoplankton species.

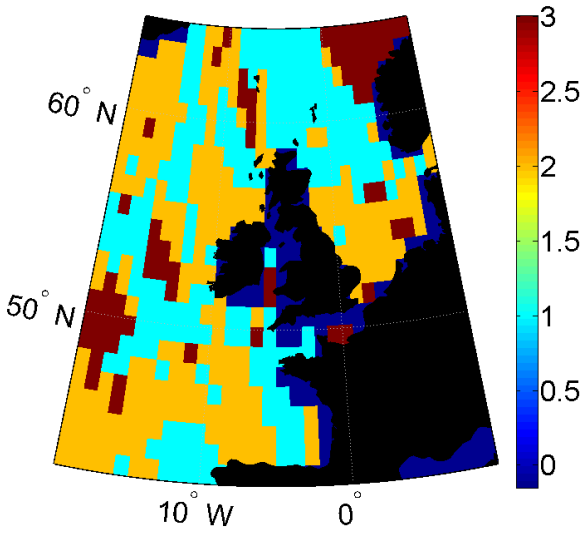


a) Number of principal components on Phytoplankton species before 1985.

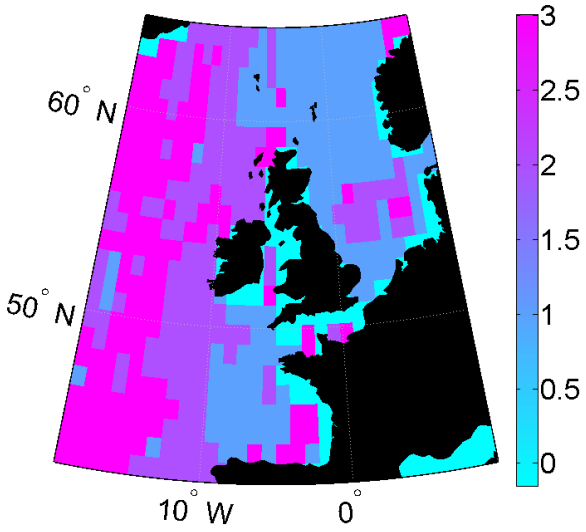


b) Number of principal components on Phytoplankton species after 1985.

Figure 6.13: Number of PCs on across Phytoplankton species.



a) Regionalisation based on Phytoplankton species before 1985.



b) Regionalisation based on Phytoplankton species after 1985.

Figure 6.14: Plots of regions before and after 1985.

be a natural part of the diatom's long term cycle. The 'regime shift' is therefore characterised in different ways for different species groups. For the zooplankton the evidence here seems to support the hypothesis that they have moved northwards over the past few decades. For the phytoplankton, as defined by the diatoms, the shift appears to occur in the spatial structure alone and might be a response to natural climate variability, although since the mechanism is not yet well understood further empirical studies are required to better understand the reasons for the spatial behaviour of the diatoms.

This suggests that changes in the biogeographical regionalisation of the North East Atlantic are driven both by the effect of rising temperatures on certain species and by natural oscillations in climate indices. In both cases these changes will doubtless have consequences for other organisms at different scales. Understanding variability in the plankton is crucial for developing an understanding of the marine ecosystem as a whole. For example the spatial reorganisation of plankton species could force changes in the spatial distribution of other organisms, as they are forced to move to where their food source is most abundant [30]. From this analysis it is apparent that the ecosystem has undergone drastic changes over the past few decades, this will have wide reaching consequences.

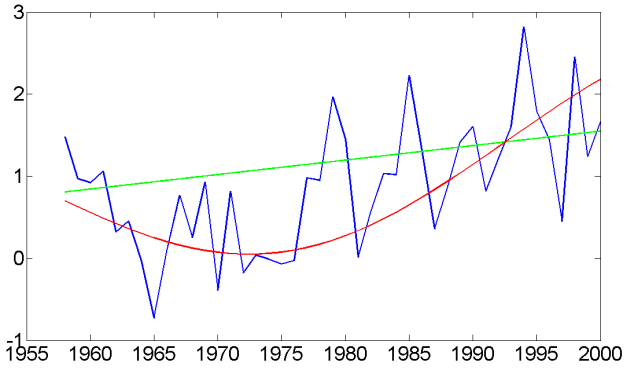
Chapter 7

Modelling Vulnerability to Climate Variables

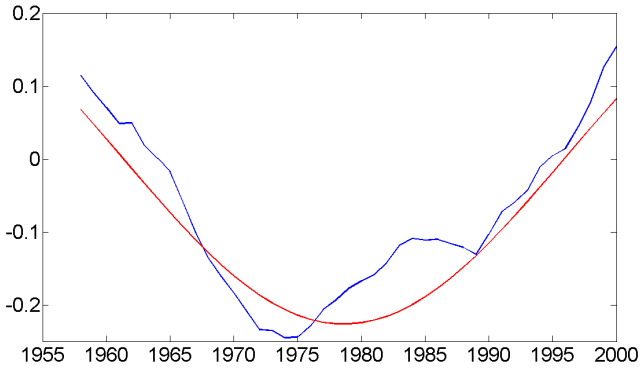
In this chapter we explore the vulnerability across space of different variables to changes in the climate covariates. Species abundances are modelled as linear responses to climate covariates (see equation 7.3), where the species abundances are taken to be responses to common trends across time and these common trends are modelled as linear combinations of climate signals. The linear regression model can be used to make predictions about changes in species abundance in response to changes in the climate variables and vulnerability of a species is defined to be a large change in species abundance in response to a relatively small change in the climate variable, although some care must be taken not to extrapolate to far outside the existing data. In a similar way the joint responses of the zooplankton and diatom assemblages can be explored in order to assess the vulnerability of groups of species across space. It has previously been established that not all regions response to climate change and natural climate oscillations in the same way [108], both in the response of the sea surface temperature signal [43] (see chapter 3) and in the responses of the plankton (see chapter 5). This suggests that some regions will be more vulnerable to climate effects than others, either for an individual species or

for the joint behaviour of a functional group of species, and we define these to be climate 'hotspots'. In the first section three indicator species are studied: *Calanus finmarchicus*, a typical cold water copepod [48, 74]; *Calanus helgolandicus*, a typical warm water copepod [48], and *Echinoderm larvae*, the offspring of benthic organisms and so know to be influential in benthic pelagic coupling [88]. In the second section the focus is shifted to exploring the responses of the joint behaviour of functional groups, concentrating on the zooplankton and Diatom groups. In the final part a multiscale downscaling approach, where the data is first analysed over a large region and then analysed at smaller scales with the average subtracted, is taken to explore whether the influence of different climate indices vary across different scales.

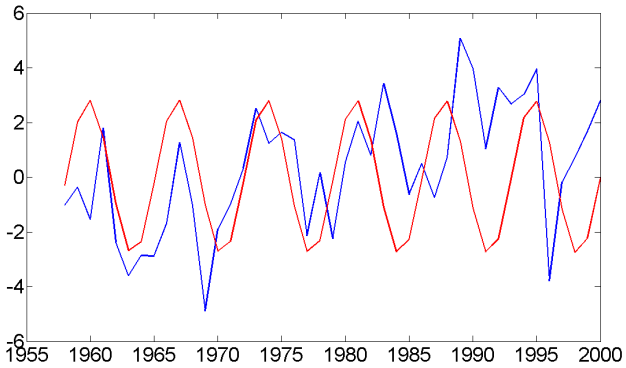
In our model the abundance each of species at each location is modelled as a linear combination of the principal components and these in turn are modelled as responses to climate covariates. This model can be used to make predictions as to how the species abundance will change with changes in the climate variables. The covariates used are the NAO, the AMO and the NHT signals. Since out of sample predictions, i.e. predictions of the behaviour of the plankton outside the timescale of the available data, can not be made using the existing climate data, an approximation of the climate indices must be found. The AMO and the NAO can be approximated by fitting sinusoids using the method described by Rice [133]. The NHT can be approximated by fitting a combination of a linear trend and an oscillation. Figure 7.1 shows these signals plus the fitted models for each signal. The modelled NHT and AMO follow closely the true signals. The modelled NAO fits better before 1985 than it does after, which may mean that PCs which are strongly associated with the NAO might be less accurately modelled in this time period.



a) The NHT signal (blue), the fitted trend line (green) and the modelled NHT (red).



b) The AMO signal (blue) and the modelled AMO (red).



c) The NAO signal (blue) and the modelled NAO (red).

Figure 7.1: Plots of the climate variables and the fitted climate variables. The true climate variables are plotted in blue, the fitted variables in red and for the NHT the trend line is plotted in green. The units of the NHT and the AMO are in degrees Celsius. The NAO is taken from a dataset with standardised units.

7.1 Methods Used in this Chapter

In this chapter the vulnerability to different climate indices is explored using the output of sparse PCA, which is described in detail in section 2.2. Vulnerability to climate is explored both for individual species and for groups of species. Recall from section 2.5 that each principal component might be seen as a joint response of a species assemblage and that this might have a linear relationship with climate trends. The parameters in equation 7.1 (see section 2.6) can be estimated and these estimates can be used to make predictions about how the common component might change under different values of the climate indices.

$$z_i(t) = \beta_{i,0} + \beta_{1,i}c_1(t) + \beta_{2,i}c_2(t) + ... + \beta_{N,i}c_N(t) + \epsilon_i(t), \tag{7.1}$$

If the magnitude of $\beta_{i,j}$ standardised by the magnitude of $c_j(.)$ takes a high value at a particular location then it can be said that principal component i is sensitive to the climate variable $c_j(.)$ at this location, since this implies that a large change in $c_j(.)$ will lead to a significant change in the principal component. Species that have large weights on that particular component might also be said to be sensitive to that climate variable. Linear regression analysis relies on a number of assumptions. The relationship between the response and the predictors should be linear and the residuals should be normally distributed. These assumptions are expected to approximately hold for the CPR data at least within a limited range of the covariates. The assumption of normality of the residuals may be violated if there is a predictor of species abundance that has not been included in the model, as there may be some structure that is not explained by the covariates that have been included. Typically it would be expected that there is no collinearity between the predictors. This is unlikely to hold for the climate variables. The AMO and the NAO are not linearly independent for instance, as the former can reinforce the latter during its high phase. An alternative might be to use the principal components of the predictors to

ensure no collinearity. However this might not be as readily interpretable as using the climate covariates as predictors and so a compromise is made. The fit of the model can be assessed by computing standard errors, which are the square root of the sum of square residuals divided by the degrees of freedom, which are equal to the number of observations minus two. The standard errors measure the amount of variation in the dependent variable, i.e. the plankton variable, that is not explained by the predictors and so should be relatively small if the model predicts well.

The vulnerability of each species to the climate variables can also be explored using this model. If each species if a linear combination of the common components then the regression model can be used to predict changes in the abundance of a particular species, which recall can be written as

$$\begin{aligned} Y^{(p)}(t;l) &= \sum_{i=1}^{\hat{p}(l)} a_i^{(p)}(l)(\beta_{i,0} \\ &+\beta_{i,1}c_1(t;l) + ... + \beta_{i,n}c_n(t;l) + \mu_i(t;l)) + \epsilon_p(t;l) \end{aligned} \tag{7.2}$$

.

Section 2.6 describes how the principal components might be used in linear regression analysis to make predictions for individual taxa. The limitation of the linear regression model is that predictions should not be made too far outside the range of the existing data, since the assumption of a linear relationship may not continue to hold. The advantage of using the principal components instead of each species directly is that it saves computational effort by only estimating the regression parameters for the smaller set of components rather than every variable.

7.2 Sensitivity of Different Species to Climate Change

7.2.1 The Spatial Sensitivity of *Calanus finmarchicus*

In order to assess the validity of the model abundances were first approximated from the model using the fitted values of the various covariates and then compared with the true abundances. Using the regression model and the modelled climate signals shown in figure 7.1, estimates of the abundances in 2008 and 1958, which are both years at which data on the true abundances exist, of *Calanus finmarchicus* can be obtained. It is important that it be shown that the modelled data and the fitted values of the covariate accurately represent the true data before out of sample predictions are made, in order to assess the validity of those predictions. Figure 7.2 shows the estimated abundances of *Calanus finmarchicus* at two time points along with the true abundances recorded in the CPR dataset at the same time points. The abundance in the northern North Sea in 1958 is underestimated by the model, whilst in 2008 it overestimates the abundance in the north west but models the northern North Sea well. Figure 7.3 shows the standard errors of the regression model as given by equation 2.19. The prediction error is largest in the northern North Sea but overall most of the variability in abundance is captured by the regression model. The histogram of the residuals shows that they are approximately normally distributed. From this it can be concluded that there is some error in the model but it does produce reasonable estimates of the spatial structure. This model implicitly assumes that the response of the species, at least for small perturbations, to climate will be linear and that most of the variability in the species abundance can be explained by the climate variables that have been included in the model. Where there are small errors it may be due to a violation of one of these assumptions, in particular the model does not account for local conditions which may have an influence on the abundance in certain regions.

After verifying the model it is possible to use it to make predictions. If it is as-

sumed that the AMO and the NAO will continue in the same oscillatory behaviour over the next ten years, then projected values of these covariates in ten years time can be taken from the fitted oscillations. Since the fitted NAO replicates the true values of the NAO less well than the fitted AMO does the true AMO then there will be some errors in the predictions for those locations where the NAO has a strong influence. Figure 7.4 shows the predicted change in abundance of *Calanus finmarchicus* under an increase of 1 degree in NHT over 10 years. Since the abundances have been logarithmically transformed this change is presented on a log-scale. Supposing the change in abundance is relatively small then the change in log-abundance is approximately equal to the proportion of change. Supposing the abundance in 2018 is equal to the abundance in 2008, denoted by $Y(2008)$, plus some change δ then

$$\log(Y(2008)) - \log(Y(2008)+\delta) = \log\left(\frac{Y(2008)}{Y(2008)+\delta}\right) = -\log\left(1+\frac{\delta}{Y(2008)}\right) \tag{7.3}$$

.

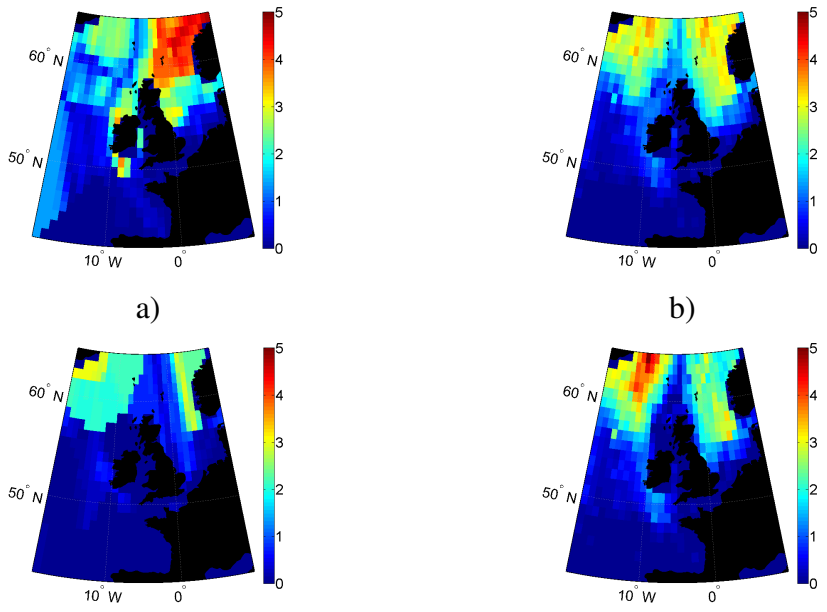
This is approximately equal to $-\frac{\delta}{Y(2008)}$. This means that a change of 0.2 in logged abundance suggests approximately a 20% predicted decrease in abundance from 2008 till 2018. The values of the AMO and NAO are taken from the model in figure 7.1 in year 2018 and the NHT is taken as the value at 2008 from the model plus a degree. This model estimates that the greatest region of change will be in the North Sea, particularly the north. Under this model abundance of *Calanus finmarchicus* is expected to decline in this region, as the values are greater under the model in 2008 than in the projected model. The weights of *Calanus finmarchicus* are negative on the first principal component on the zooplankton data in the northern North Sea (see figure 5.20) and the first PC was thought to be positively responding to the NHT in this region, which explains from a statistical perspective why the model predicts a decline in this species under an increase in temperature in these

regions. From an ecological perspective previous studies have remarked on the migration of *C. finmarchicus* towards more Arctic regions [74, 132], which may explain why it is disappearing from the northern North Sea, which is a region in which it was previously abundant.

The model can also be used to estimate the change in temperature required to produce small changes in abundance. It should not be used to predict large changes in abundance because the assumption of the linear relationship may not continue to hold at large values. The inverse model, in which equation 7.3 is rearranged to give the NHT in terms of abundance, the AMO and the NAO, can be used to find the approximate temperature change associated with a 10% decrease in the logged abundance of *Calanus finmarchicus*. In the northern North Sea this is approximately a 1° increase. In the southern and open sea part of the North East Atlantic the abundance of *Calanus finmarchicus* is low and so the model is less appropriate in these regions and so the temperature is set to zero in these regions.

7.2.2 The Spatial Sensitivity of *Calanus helgolandicus*

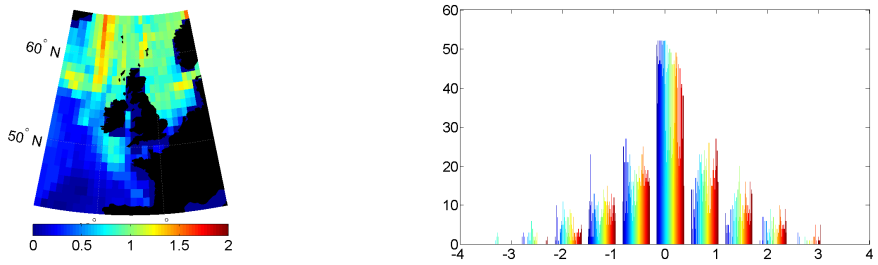
As for *Calanus finmarchicus* the model is assessed for validity initially by estimating abundances of *Calanus helgolandicus* at 1958 and 2008 using the modelled covariates and comparing with the true abundances. Figure 7.5 shows the true abundances of *Calanus helgolandicus* at two time points and the estimates from the model. The spatial pattern is closer to the true abundances than those predicted from *Calanus finmarchicus*, although it slightly underestimates how drastically the abundance has increased in the northern North Sea. In 1958 both the true and modelled abundance is highest off the south coast of the United Kingdom and off the coasts of France and Spain. In 2008 abundance has increase dramatically in the northern North Sea, where previously this species had had little presence. In the very south of its habitat, abundance seems to have decreased in 2008. Previous studies have suggested that *Calanus finmarchicus* is alternating in abundance with



a) b)

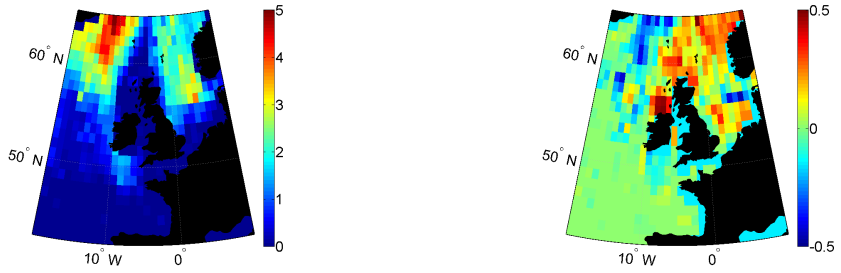
c) d)

Figure 7.2: Plots of the true log-abundances of Calanus finmarchicus and the log-abundances estimated from the modelled climate signals. a) Abundance of Calanus finmarchicus in 1958. b) Abundance of Calanus finmarchicus estimated from the regression model and using the modelled climate signals in 1958. c) Abundance of Calanus finmarchicus in 2008. d) Abundance of Calanus finmarchicus estimated from the regression model and using the modelled climate signals in 2008.



a) b)

Figure 7.3: a) Plot of the standard errors of the regression model given by equation 7.3 for Calanus finmarchicus. b) A histogram of the residuals from the regression model across all locations, with different locations being denoted by different coloured bars.



a) b)

Figure 7.4: Plots of the predicted change in logged Calanus finmarchicus abundance over 10 years under the model. a) Estimated abundance of Calanus finmarchicus under a 1 degree increase in NHT over 10 years. b) Abundance of Calanus finmarchicus estimated from the regression model in 2008 minus the estimated abundance under a 1 degree increase in NHT over 10 years.

Calanus helgolandicus, the latter increasing as the former declines [48]. Morphologically the two species are similar apart from the fact that *C. helgolandicus* prefers higher temperatures [117]. This explains why it is able to take hold in regions previously dominated by *C. finmarchicus*. Figure 7.6 shows the standard errors across space for the regression model for *Calanus helgolandicus*, which shows that the prediction error is smaller than for *Calanus finmarchicus* and is largest along the oceanic shelf, and a histogram of the residuals, which are approximately Gaussian.

Projected values are then calculated for *Calanus helgolandicus* using the model in order to determine which regions in space at which it is most sensitive to changes in climate. Figure 7.7 shows that under the model a one degree increase in NHT will result in an increase of *Calanus helgolandicus* in the northern North Sea and off the coast of Ireland, with little or no change elsewhere. Since *Calanus helgolandicus* is a warm water Copepod [37] it can be expected that rising temperatures will result in an increase in its abundance. The model shows that the northern North Sea is particular region of interest for this species, again exhibiting opposite behaviour to *C. finmarchicus*. The northern North Sea might be a suitable habitat for *Calanus helgolandicus* in terms of the available nutrients and food sources. It is morphologically similar to *C. finmarchicus* [48], apart from preferring warmer temperatures,

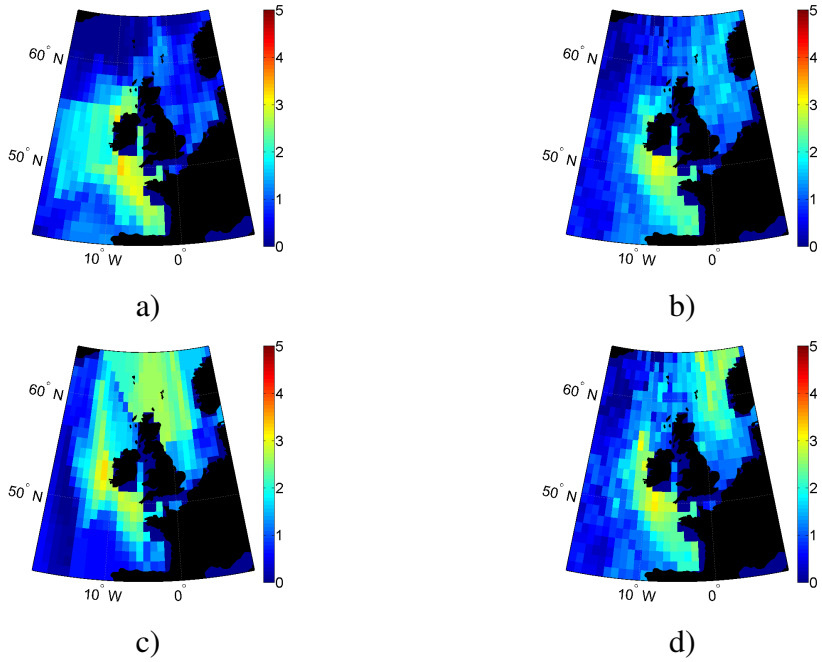
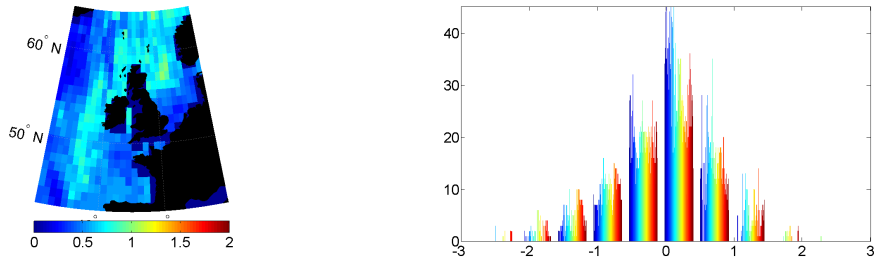


Figure 7.5: Plots of the true log-abundances of Calanus helgolandicus and the log-abundances estimated from the modelled climate signals. a) Abundance of Calanus helgolandicus in 1958. b) Abundance of Calanus helgolandicus estimated from the regression model and using the modelled climate signals in 1958. c) Abundance of Calanus helgolandicus in 2008. d) Abundance of Calanus helgolandicus estimated from the regression model and using the modelled climate signals in 2008.

and as the abundance of *C. finmarchicus* is high in the northern North Sea before the increase in temperature this suggest that the habitat is indeed suitable for species of *Calanus*. This explains why *C. helgolandicus* is most able to take hold here as temperatures increase. The model shows a much smaller decrease in abundance in the very south of the region, which might be attributable to temperatures rising too much there for *Calanus helgolandicus*, although the change here is less dramatic.

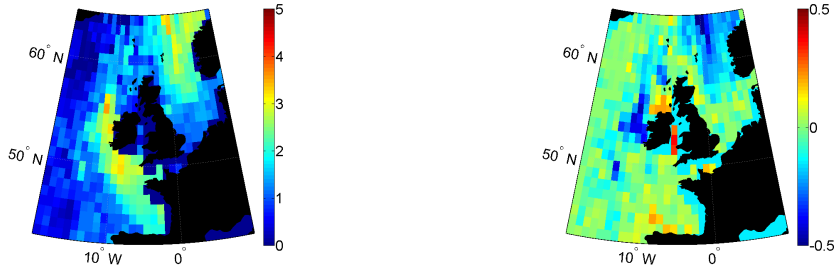
7.2.3 The Spatial Sensitivity of *Echinoderm Larvae*

Figure 7.8 shows the true and modelled abundances of *Echinoderm larvae* in 1958 and 2008. The modelled abundances are slightly smoother in space than the true abundances. Before 1958 this species is mainly present in the southern North Sea,



a) b)

Figure 7.6: a) Plot of the standard errors of the regression model given by equation 7.3 for Calanus helgolandicus. b) A histogram of the residuals from the regression model across all locations, with different locations being denoted by different coloured bars.



a) b)

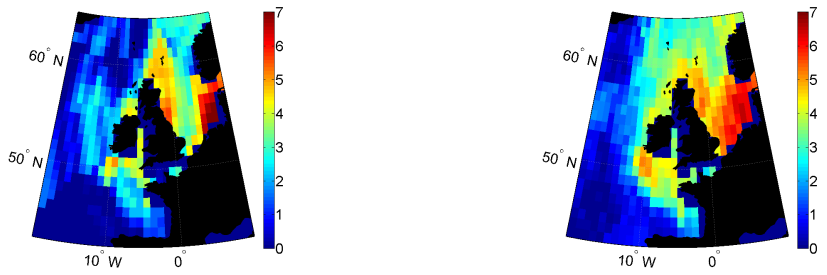
Figure 7.7: Plots of the predicted change in logged Calanus helgolandicus abundance over 10 years under the model. a) Estimated abundance of Calanus helgolandicus under a 1 degree increase in NHT over 10 years. b) Abundance of Calanus helgolandicus estimated from the regression model in 2008 minus the estimated abundance under a 1 degree increase in NHT over 10 years.

whilst afterwards there has been a dramatic increase in its abundance particularly in shallower waters and the northern North Sea. The change in shallower waters may relate to a change in the adults, which are benthic species. One potential cause of this increase might be due to the warming in these shallower waters, from which benthic organisms are able to benefit from more than in deeper waters. In shallower waters light is able to penetrate to the sea floor, which is beneficial to plant life on the sea floor [68], and this is essential for the survival of many benthic organisms. Consequently this may lead to an increase in the larvae of these organisms. Figure 7.9 shows the standard errors of the regression model for Echinoderm larvae across space and a histogram of the residuals. The standard errors are larger than those for *Calanus finmarchicus* and *Calanus helgolandicus*, suggesting the model performs less well for Echinoderm larvae. This might indicate that there is some other covariate that drives abundance of this species that has not been included in the model. The residuals are reasonably normally distributed.

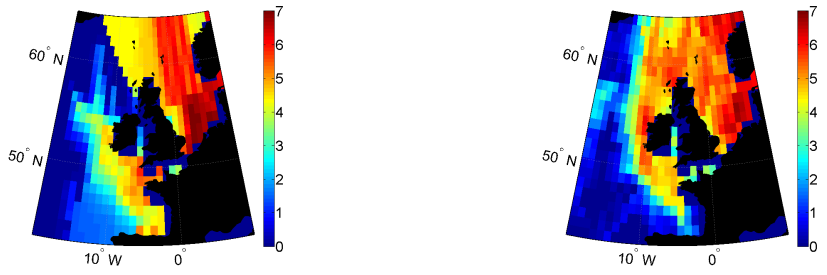
Figure 7.10 shows that under the model a 1 degree increase in the NHT will lead to an increase in the abundance of *Echinoderm larvae* in the northern North Sea. The *Echinoderm larvae* are a pertinent example of how changes in plankton can impact the entire ecosystem, as increased numbers of larvae may well indicate increased numbers of the adults because of benthic-pelagic coupling [88]. From the positive relationship between *Echinoderm larvae* and NHT it can be inferred that organisms such as starfish and sea urchins may be moving north as temperatures rise.

7.3 Sensitivity of Joint Responses over Space to Climate Change

In this section projected community responses to climate variation across space are explored. The linear regression model is used to estimate the relationship between the principal components, representations of the joint responses across communi-

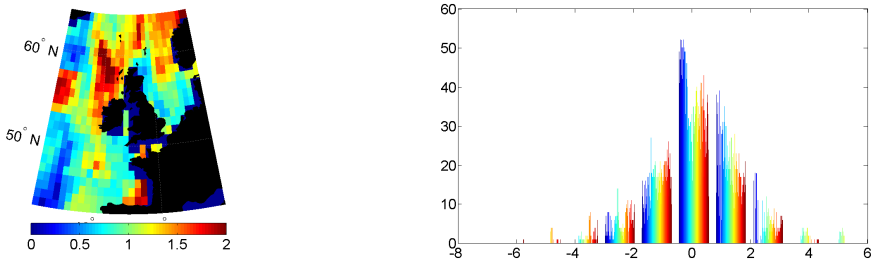


a) b)



c) d)

Figure 7.8: Plots of the true log-abundances of Echinoderm larvae and the log-abundances estimated from the modelled climate signals. a) Abundance of Echinoderm larvae in 1958. b) Abundance of Echinoderm larvae estimated from the regression model and using the modelled climate signals in 1958. c) Abundance of Echinoderm larvae in 2008. d) Abundance of Echinoderm larvae estimated from the regression model and using the modelled climate signals in 2008.



a) b)

Figure 7.9: a) Plot of the standard errors of the regression model given by equation 7.3 for Echinoderm larvae. b) A histogram of the residuals from the regression model across all locations, with different locations being denoted by different coloured bars.

a) b)

Figure 7.10: Plots of the predicted change in logged Echinoderm larvae abundance over 10 years under the model. a) Estimated abundance of Echinoderm larvae under a 1 degree increase in NHT over 10 years. b) Estimated abundance of *Echinoderm larvae* from the regression model in 2008 minus abundance of *Echinoderm larvae* under a 1 degree increase in NHT over 10 years.

ties, and the different climate variables. By varying the value of a given covariate in the model changes in the principal components in responses to changes in climate can be estimated. These changes can be thought to correspond to changes in the joint behaviour of functional groups of species as a response to different climate variables. The biological problem relates to how regions where the community response is most sensitive to climate variation can be found, and this is achieved by determining in which regions the principal component changes the most under a change in the value of the covariate. In this case the principal components are viewed as being equivalent to joint responses of assemblages of dominant species, since rarer species have been forced by the sparsity constraint to have zero weight. For both the Diatom and the zooplankton groups the model is used on the first PC, which represents the joint behaviour of the most dominant assemblage across each location. The model is then used with the modelled covariates (see figure 7.1) to find estimates of the PCs at each time point in order to see how well the model predicts the true spatial pattern of the component. The values of the covariates are then varied and the difference between the PCs determined by those new covariates and the true PCs is examined in order to see how changing each covariate separately influences the species community.

7.4 Joint Response of the Zooplankton Group

As was done for the individual species the model can be verified by using it to estimate the values of the principal component at time points for which data exists and then comparing these with the true values of the components at those time points. Figure 7.11 shows the true and the modelled first principal component in 2008 for the zooplankton group. The model gives a reasonably good estimate of the first component. The values are higher in the north, where the principal component has a positive correlation with the NHT warming trend and lower further south, where the correlation is negative. Standard errors of the regression model across space are shown in figure 7.12.

Figure 7.13 show how the first PC varies when the covariates are increased. The principal component is estimated by fixing the values of two of the covariates and increasing the third by 50% of its value. The figure shows the estimated PC from these covariates minus the true values of the PC in 2008 across space. The plots therefore show the estimated change in the principal component under a 50% increase in one covariate whilst the other two remain fixed. The model for the change under a 50% rise in the NHT can be written as:

$$z_1(2008) - \hat{\beta}_{0,1} - \hat{\beta}_{1,1}\text{AMO}(2008) - \hat{\beta}_{2,1}\text{NAO}(2008) - 1.5 \times \hat{\beta}_{3,1}\text{NHT}(2008). \quad (7.4)$$

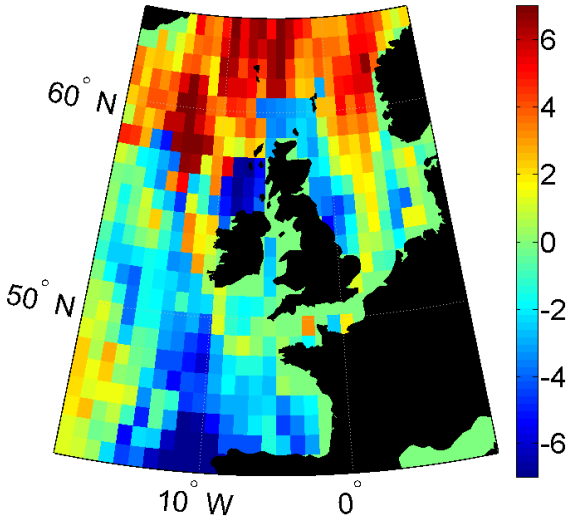
Where $z_1(2008)$ is the value of the first component in 2008 and $\hat{\beta}_{j,1}$ are the estimated parameters from the linear regression model. The model will not exactly estimate the principal component, due to the effect of the error term in the regression model, which corresponds to the effect of noise. It is assumed that the noise will be normally distributed with mean zero. Increasing the AMO whilst fixing the NHT and the NAO results in a slight increase in the north North Sea, particularly the very north east of the region. The change is relatively small elsewhere. Increasing

the NAO results in a very slight increase in the north east of the North Sea but this is a very small in magnitude. The biggest change comes from varying the NHT, which has the strongest correlation with the first PC across most spatial locations. There is an increase in the north, particularly the north North Sea, with the increased NHT. Helaout [75] discusses the concept of ecological niches in contributing to the spatial sensitivity of the zooplankton and other studies have commented on critical values of temperature that can lead to 'regime shifts' in zooplankton behaviour [21]. For regions that lie at the edge of these thermal boundaries only a small change is needed to force a large change in group behaviour of zooplankton species. This can be seen in the northern North Sea in the zooplankton data. The reason for the sensitivity of this region may be attributable to the high abundance of cold water species in the early part of the dataset, which are gradually moving towards more Arctic regions, coupled with the fact that these waters are beginning to reach temperatures at which more warm water species can survive. Further south the increase in average temperature has had less dramatic an effect, although in the south there is some decrease as the NHT increases. The absolute values of the change with the increased NHT are highest in the north east North Sea, suggesting that the zooplankton are particularly sensitive to changes in temperature in this region, as seen from the individual species model on *Calanus finmarchicus* and *Calanus helgolandicus*.

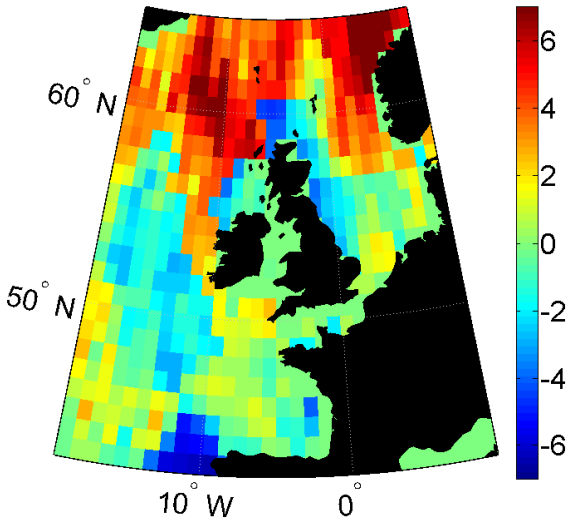
7.5 Joint Response of the Diatoms

The same analysis is carried out on the Diatom species. It has already been established that these species are less sensitive to changes in temperature but that the spatial structure is influenced by changes in natural oscillations. Figure 7.14 shows the real and the modelled first PC for the Diatom community and figure 7.15 shows the standard errors of the regression model across space. The PC is positive over

most of the North East Atlantic but particularly in the north of the region at this time point. The first PC has a positive correlation in most places with the AMO and this is particularly strong in the north of the North East Atlantic and in the North Sea. In 2008 the AMO is in its high phase, resulting in a positive principal component in most parts of the North East Atlantic. As can be observed the model estimates the values of the first PC reasonably well, although it occasionally overestimates the magnitude. Figure 7.16 shows the modelled change in the principal component when each of the covariates is increased by 50%. It is clear that the greatest change occurs by increasing the value of the AMO, which is to be expected as it has already been established that this climate index has the strongest relationship with the diatoms of all the climate variables under consideration. This leads to an increase in the PC particularly in the North Sea and the north east of the region. Elsewhere the change in response to the AMO is less dramatic and even slightly negative in the very south, though the magnitude of the change is smaller in these regions. The response to the AMO may be linked to the Bathymetry as increasing the AMO has more of a positive effect on the first principal component in shallower waters. This effect is something that has been discussed earlier, where it was hypothesised that the AMO has a greater influence on the Diatom community in shallower waters, and it was suggested the influence of the AMO in shallower water may be linked to its supposed influence on wind speeds and currents [141, 90]. As was discussed in chapter 5 this implies that the influence of the AMO on water column mixing, which is influential in driving Diatom abundance [56], is stronger in these shallower waters. The changing the NAO whilst keeping the other two covariates the same has relatively little effect on the PC, suggesting that the dominant trend in the Diatoms is not sensitive to the NAO. There is some change when the NHT is increased. It has a slight positive effect in some areas and a negative effect in others, particularly coastal regions. The change however is smaller than that seen under varying the AMO. Increased temperature may have an impact on Diatom abundance, although



a) The first PC across space at 2008.



b) The modelled first principal component in space at 2008.

Figure 7.11: Plots of the true first principal component on the zooplankton species in 2008 and the modelled values.

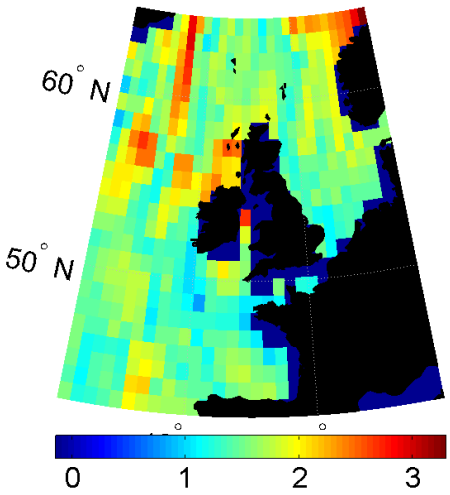
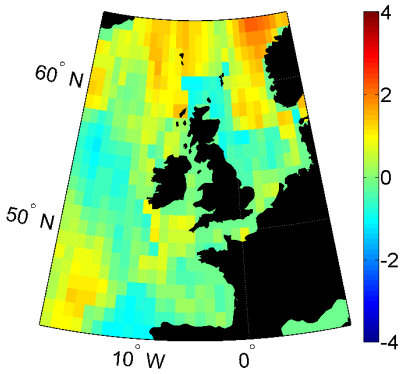


Figure 7.12: Plots of the standard errors of the regression model for the first principal component of the zooplankton communities.

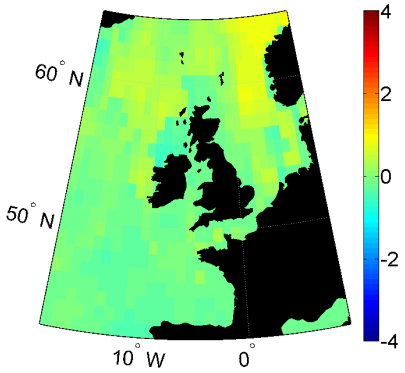
clearly this is secondary to the AMO.

7.6 Multiscale Downscaling

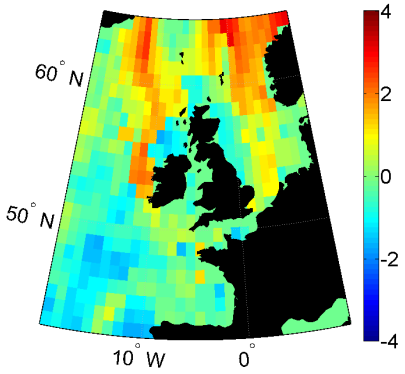
Up to this point the influence of different drivers on variability has either been studied at a large scale or at an individual level. The NHT and the AMO influence the average behaviour across the whole North East Atlantic but it may be the case that more small scale effects only have an influence at a more local level. Furthermore as the average warming trend might obscure the influence of natural oscillations [141], so might the influence of these major drivers obscure more local effects. As it has been already shown that there is spatial heterogeneity in responses to climate, the major drivers on plankton abundance may vary across different scales. For example although the impact of the average warming trend is the most important driver of zooplankton abundance over the whole North East Atlantic, at a local level the impact of nutrients or other localised effects might be important. To study this a multiscale downscaling approach can be used by finding provinces defined solely



a) The estimated first PC under a 50% rise in the AMO.

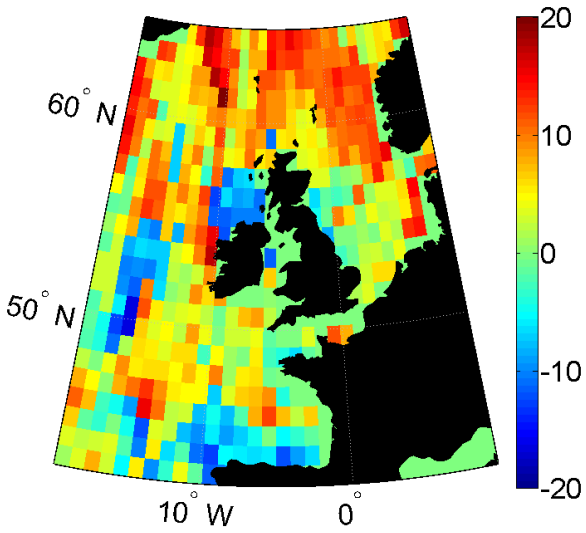


b) The estimated first PC under a 50% rise in the NAO.

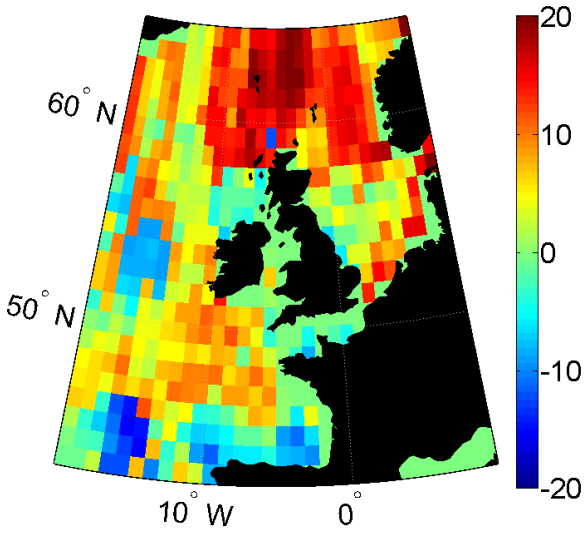


c) The estimated first PC under a 50% rise in the NHT.

Figure 7.13: Plots in space showing the difference between the first principal component modelled under varying the covariates and the real first component.



a) The first PC across space at 2008.



b) The modelled first principal component in space at 2008.

Figure 7.14: Plots of the true first principal component on the phytoplankton species in 2008 and the modelled values.

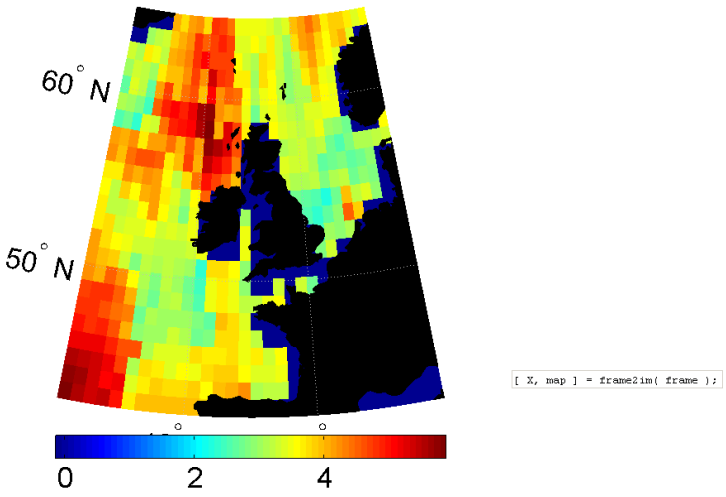
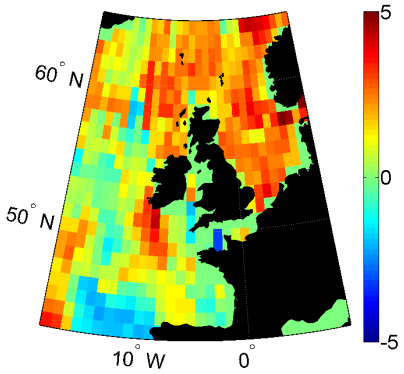
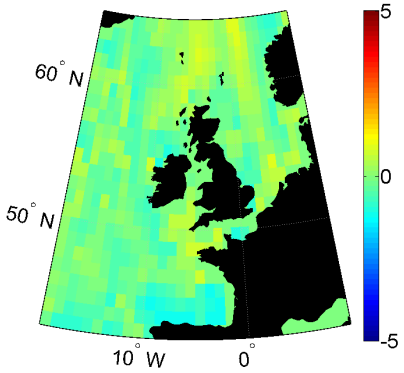


Figure 7.15: Plots of the standard errors of the regression model for the first principal component of the diatom communities.

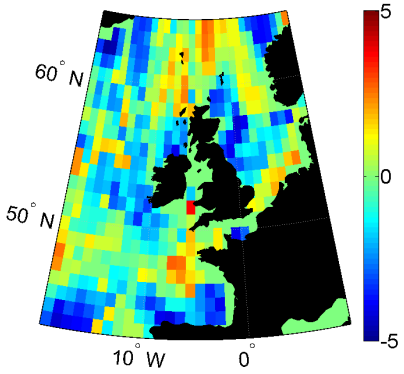
on the CPR data [145]. In order to explore the different trends at different scales the data is first averaged across all pixels, with abundances being weighted by the size of each pixel, and the analysis is carried out over the whole North East Atlantic. Since pixels are 1 degree latitude by 1 degree longitude they will have different areas at different latitudes, due to the curvature of the earth. The length of 1 degree in longitude is proportional to the cosine of the latitude. This means the weights are given by $(\cos(\text{lat}_1) - frac\cos(\text{lat}_1) - \cos(\text{lat}_2)2)^2$, where lat_1 is the lowest degree of latitude and lat_2 is the most northerly degree of latitude that bounds the pixel. In order to analyse the data at more local scales the weighted average for each species across time, normalised by the mean of the weights, is subtracted from the time course for each species at each location. To find provinces the analysis is carried out on the data with the averages removed and k-means clustering is used on the weight vectors. Once these provinces have been defined the analysis is repeated on the averages over each province in order to find local trends that are not obscured by large scale features of the data.



a) The estimated first PC under a 50% rise in the AMO.



b) The estimated first PC under a 50% rise in the NAO.



c) The estimated first PC under a 50% rise in the NHT.

Figure 7.16: Plots in space showing the difference between the first principal component modelled under varying the covariates and the real first component.

7.6.1 Zooplankton Data at Different Scales

Region	Component	NHT	AMO	NAO
North East Atlantic	PC 1	-0.5	/	/
Cluster 1	PC 1	-0.5433	/	/
Cluster 1	PC 2	/	/	-0.5467
Cluster 2	PC 1	0.5572	/	/
Cluster 3	PC 1	/	/	0.4623
Cluster 3	PC 2	/	0.6491	/

Table 7.1: Table of the Pearson's correlation coefficients between the principal components of the zooplankton data over each cluster and different climate indices.

Recall in chapter 5 the zooplankton and the diatom species were analysed over the average abundances for the whole North East Atlantic. On the data averaged across the whole North East Atlantic the first PC on the zooplankton shows a general decreasing trend. This is negatively correlated with the NHT trend with a time lag of approximately 9 years and a p-value of less than 0.05. The Pearson's correlation coefficient is close to -0.5. The second component did not correlate strongly with any of the climate variables under consideration and so is likely to be driven by some unknown covariate. The third PC resembles the AMO (see chapter 5). This implies that at a large scale long term trends and low frequency oscillations drive most of the abundance of zooplankton. Local effects are averaged out and so do not influence the results, meaning that heterogeneity [108] across space is cancelled out in taking averages.

Figure 7.17 shows clusters in space on the real values of the weight vector for the zooplankton data, where the average for the whole North East Atlantic has been subtracted at each location. The regions are distinct from those found earlier in the study because the average behaviour is no longer dominating the spatial structure and can be seen as representing regions defined by species assemblages which have significant local variability not fully accounted for by the average trend. The clusters here might be viewed as regionalisation driven by local behaviour rather than

the spatial influence of large scale climate drivers. Whereas earlier when clustering on the first loading vector for the data without the average removed the clusters showed a north-south division, the spatial pattern now seems driven more by the bathymetry. The deeper waters around the open ocean are divided in to two clusters, clusters one and three. Cluster one covers the southern half of this area and cluster three the north west. Cluster two occupies shallower waters, such as the North Sea and around the ocean shelf in the south. On average it is thought that the temperature was the main driving force being the spatial pattern of the zooplankton, with cold water species dominating the north and warm water species further south [12]. There was also a trend towards increasing numbers of warm water species in the north [5, 12]. Once the average effect has been removed bathymetry and currents may play more of a role in determining the spatial distribution and this effect was obscured by the larger effect of temperature. The influence high frequency oscillations, such as the NAO, on plankton may be determined by physical features, such as the ocean shelf, rather than bathymetry [59]. The pattern of the influence of the NAO on the SST is not determined by latitude but by the position of its two pressure centres (see figure 3.2 and [76]). This also explains why the spatial patterns might change once the average effect has been removed.

Figure 7.18 shows the principal components for each cluster. Summaries of the relationships between the principal components of the zooplankton data for each region and the different climate indices are shown in table 7.1. In clusters one and two the first component correlates strongly with the NHT warming trend. In cluster one, which covers the southern part of the open sea, the first component has a Pearson's correlation coefficient with the NHT of -0.5433 (see table 7.1). In cluster two, which covers the shallower waters including the North Sea the Pearson's correlation coefficient between the first component and the NHT is positive (see table 7.1). Unlike the averaged data, however, there is no time lag between the two signals. Given there is some correlation with the NHT trend in these regions even after the average

for the whole North East Atlantic is subtracted the warming trend is likely most important in these regions than it is for the average over the entire area. The second PC in region one has a negative correlation with the NAO (see table 7.1). In cluster three, the north west region, the warming trend is less important. The first PC correlates with the NAO positively and the second has a strong positive correlation with the AMO (see table 7.1). Once the average trend has been removed it becomes clear that NAO has an influence on the plankton abundance as well, although this might have been obscured by the dominant trend, which is thought to be primarily driven by the NHT. The relationship between the NAO and zooplankton has been observed in *C. finmarchicus* and *C. helgolandicus* [59], although other studies have shown that this is less significant than the overall warming trend [74]. This implies that it may be a more localised effect on the zooplankton.

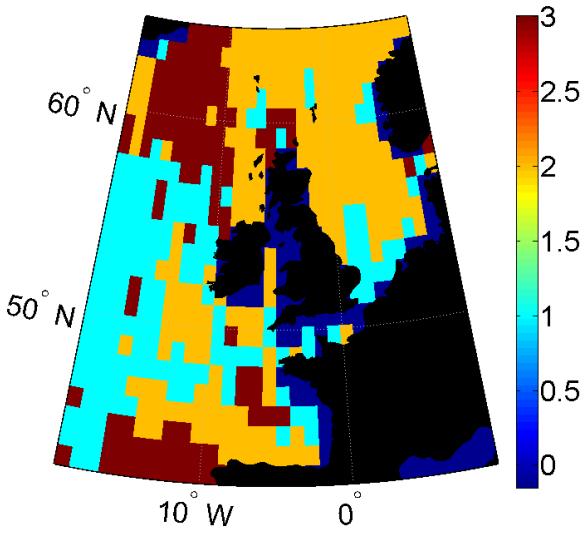
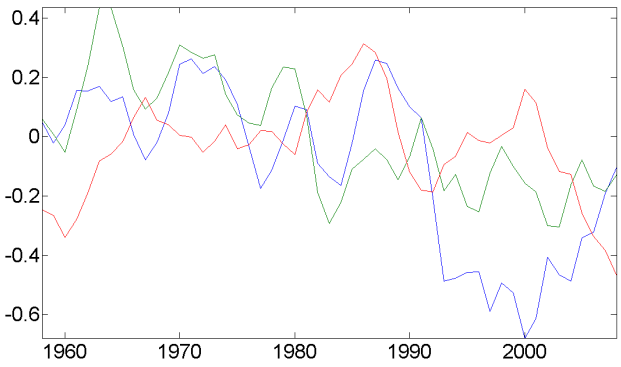
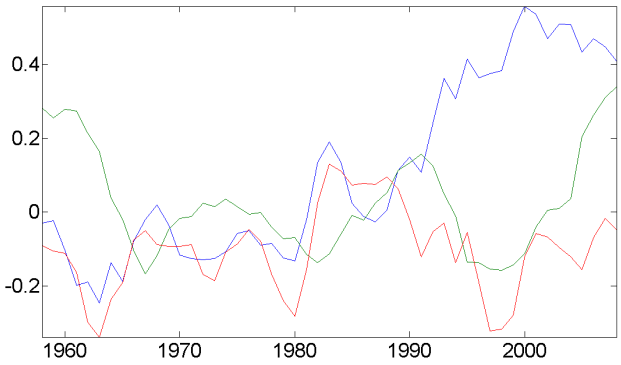


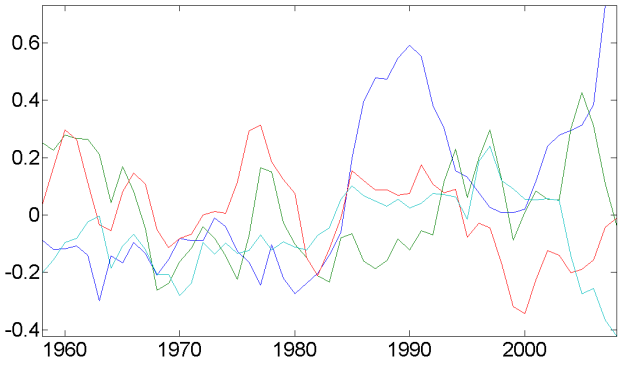
Figure 7.17: Plots of the clusters on the first loading vector for the zooplankton data with the spatial average for the whole North East Atlantic removed.



a) The principal components for the zooplankton with the spatial average removed averaged across the first cluster.



b) The principal components for the zooplankton with the spatial average removed averaged across the second cluster.



c) The principal components for the zooplankton with the spatial average removed averaged across the third cluster.

Figure 7.18: Plots of the principal components for each region for the zooplankton with the spatial average for the whole North East Atlantic removed across the each of the clusters. The first component is shown in dark blue, the second in green, the third in red and the fourth in light blue.

Region	Component	NHT	AMO	NAO
North East Atlantic	PC 1	/	0.6563	/
Cluster 1	PC 1	/	0.5917	/
Cluster 1	PC 2	/	/	0.4194
Cluster 2	PC 1	/	-0.7487	/
Cluster 3	PC 1	/	/	-0.5073
Cluster 3	PC 2	0.3857	/	/

Table 7.2: Table of the Pearson's correlation coefficients between the principal components of the diatom data over each cluster and different climate indices.

7.6.2 Diatom Data at Different Scales

On the Diatom species the first principal component over the whole region has a strong correlation with the AMO [56], with a Pearson's correlation coefficient of 0.6563. The principal components for the Diatom species on the data averaged over the North East Atlantic shows oscillatory behaviour in time and do not correlate strongly with the NHT warming trend. Before the average is removed the clusters on the Diatom data follow closely the pattern of the Bathymetry, especially post-1985 (see chapter 5). Once the average for the whole North East Atlantic has been subtracted at each location and clusters on the loading vectors from the data with this average subtracted are found, the spatial pattern becomes less structured. This may be because the spatial structure that was driven by the AMO has been removed. It still follows the Bathymetry to an extent, as seen from figure 7.19, since clusters one and three seem to primarily cover deeper waters and cluster two primarily covers shallow waters. Cluster three covers the region in the south where the ocean shelf lies and there is a lot of mixing of species. However it also covers the very shallow waters of the southern North Sea. Around the northern coast of Scotland, which is a know mixing region due to the presence of the oceanic shelf and the boundary between North Sea and open ocean waters [161], there is a mixture of regions belonging to different clusters, meaning the regionalisation is not spatially smooth here.

Another effect of removing the average vector is that the number of principal components needed to explain most of the variation increases. Figure 7.20 shows the principal components and loading vector for the Diatom data with the average for the whole North East Atlantic subtracted at each location them averaged over each cluster. Table 7.2 contains summaries of the Pearson's correlation coefficient between the principal components of the diatom data across each region and the different climate indices. In the first cluster, which mostly covers the open sea, the first PC correlates positively with the AMO (see table 7.2). The second component correlates positively with the NAO (see table 7.2). In cluster two, which covers the shallower waters, the first PC correlates with the AMO with a large negative Pearson's correlation coefficient (see table 7.2). As with the NHT and the zooplankton data there remains a correlation with the principal components on the Diatoms and the AMO in some regions even after the average signal has been subtracted. This indicates that these regions respond more strongly to the AMO than the average signal for the whole North East Atlantic. In the third cluster, which covered some of the mixing regions and the southern North Sea, the NAO correlates negatively with the first PC (see table 7.2). The second component in this region has a weaker correlation with the NHT warming trend with a positive Pearson's correlation coefficient (see table 7.2), suggesting that temperature might have a small impact on Diatom abundance in this region. As with the zooplankton the NAO only correlated with the phytoplankton components once average effects had been removed. Schlesinger [141] discusses how the warming trend can obscure the effect of natural oscillations when looking at the SST signal. The same appears to be true of the plankton, in that higher frequency oscillations can be obscured by the average trends, since PCA finds the trends that explain the majority of the variance.

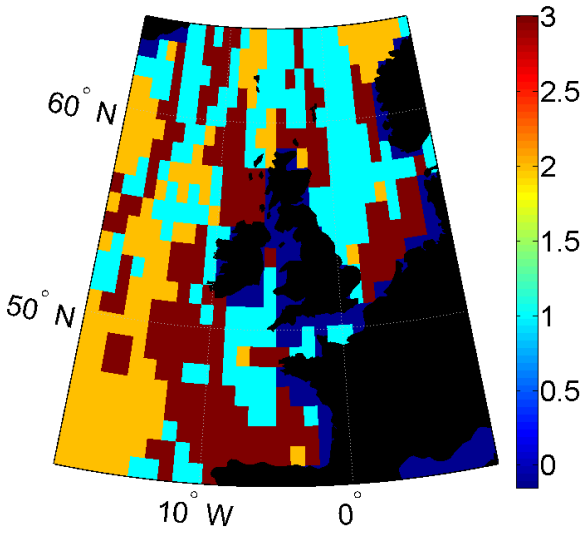
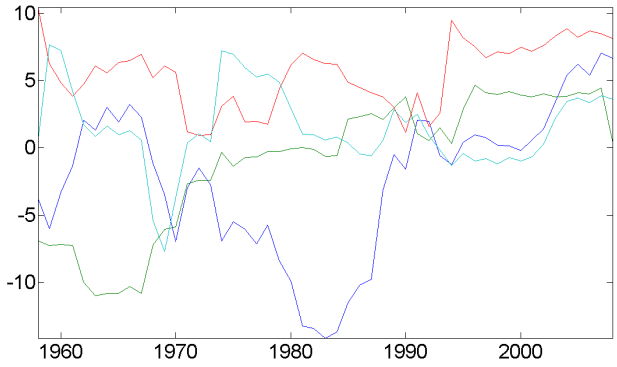


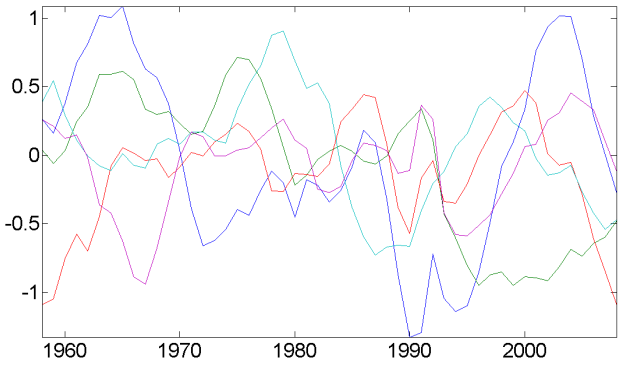
Figure 7.19: Plots of the clusters on the first loading vector for the Phytoplankton data with the spatial average for the whole North East Atlantic removed.

7.7 Discussion

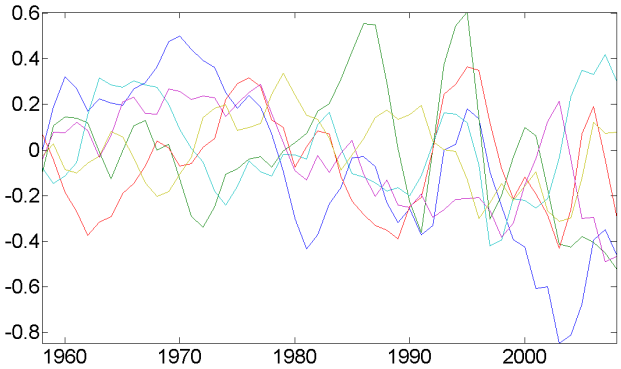
In this chapter the question of whether the sensitivity or the plankton to climate variables varies in space and what changes in the distribution of the plankton are likely to occur with changes in climate variables. The results suggest that the northern North Sea is a particular region of vulnerability for the zooplankton, with warm water species being increasing and cold water species declining in abundance. This change seems to be primarily driven by the NHT. For the Diatom species the greatest change occurs by varying the AMO and this change is greatest in shallower waters, with the presence of 'hotspots' being linked to the bathymetry for the Diatom species. Some caution must be taken in making out of sample predictions, as these make the assumption that the linear relationship will continue. Clearly it is untrue that the linear relationship will hold indefinitely, for example it is likely that there is a critical temperature at which the abundance of the warm water copepod *Calanus helgolandicus* will begin to decline, and so only relatively small changes in the climate variables are explored in this study. The slight decline of this species



a) The principal components for the Phytoplankton with the spatial average removed averaged across the first cluster.



b) The principal components for the Phytoplankton with the spatial average removed averaged across the second cluster.



c) The principal components for the Phytoplankton with the spatial average removed averaged across the third cluster.

Figure 7.20: Plots of the principal components for the Phytoplankton with the spatial average for the whole North East Atlantic removed across the each of the clusters. The first component is shown in dark blue, the second in green, the third in red, the fourth in light blue, the fifth in purple and the sixth in yellow.

in the very south of the region under an increased NHT supports the hypothesis that there is a maximum optimal temperature for this species. Predictions are, however, useful for members of industry and for policy makers who wish to understand better what steps to take in light of projected changes in the ecology.

At different scales the sensitivity of the plankton to environmental factors changes. Once the large scale trends have been removed, for instance, the Diatom species show more complex local structure. It is also apparent that although the NAO was not found to have a significant correlation with the plankton data in the earlier part, once the influence of other drivers is removed it does have a significant correlation in certain provinces. This suggests that the influence of smaller scale oscillations can be obscured by longer term trends. This chapter demonstrates how sparse PCA might be used to assess the vulnerability of different species to different climate drivers both across space and at different scales. Although some care must be taken to keep in mind the assumptions of the model, this chapter shows that the methods described can be used to make projections about the behaviour of plankton communities.

Chapter 8

Conclusions

8.1 Discussion of the Main Results

In this study a novel approach was used to explore the CPR dataset across different dimensions and two applications of this methodology are presented. This approach adapts existing statistical methodology in order to study the structure of the CPR dataset over space, time and species. This approach to the analysis of the CPR data was novel as most previous studies have restricted to focusing on one or two indicator species [25] and even those that take a more multivariate approach have studied only a relatively small subgroup of species [24]. The two research questions are addressed are whether a 'regime shift' in the biogeographical regionalisation of the North Atlantic has occurred and whether particular regions can be said to be more vulnerable to climate change. Only by using suitable statistical methods can these questions be addressed at a community level. The results of this work show that complex multivariate structure in both space and time across the Continuous Plankton Recorder can be modelled using sparse PCA and that this model allows new insights in to the ecosystem to be made. This methodology has been applied to two biological questions but it is clear that it could be extended as a tool for addressing various different research problems with the CPR data or even using

other ecological datasets. In this chapter an overview of the main findings of the study is presented and suggestions for further extensions of this work are made.

8.1.1 Species Structure

In this analysis the impact of climate drivers was explored both across individual species and across communities. One potential consequence of the 'regime shift' is a change in the species composition of the North East Atlantic [75, 66, 143, 5] and this will have consequences for the entire habitat.

Analysis of individual indicator species using spatial PCA can be used to gain some understanding of how climate can effect the ecosystem of the North Atlantic [25, 31] and how the influence of different drivers might vary in space. In this study it has been shown that species of copepod, such as *Calanus finmarchicus* and *Calanus helgolandicus*, are primarily driven by fluctuations in temperature. These two species are morphologically similar but respond in different ways to temperature. By using sparse PCA to determine the vulnerability of different species to climate the northern North Sea is shown to be a particular region of change for these two species, with a predicted decrease in *C. finmarchicus* and a predicted increase in *C. helgolandicus* if temperatures continue to rise in this region (see figures 7.4 and 7.7). These results support the hypothesis that the species composition has changed and is likely to continue to change with rising sea surface temperatures, since certain species becoming more able to establish themselves with rising temperatures in waters which would have previously been too cold. Meanwhile it has been shown in this analysis that cold water species, such as *C. finmarchicus*, are likely to retreat even further north. The study of the vulnerability across space shows temperature changes may lead to switching in the dominant species from cold water species to warm water ones [48] and this can be seen by comparing *C. finmarchicus* and *C. helgolandicus*.

Spatial PCA shows that phytoplankton biomass is on average increasing as tem-

peratures rise. This has been reported in a number of other studies [54]. The phytoplankton form the basis of the marine food web and so changes in their abundance and composition will have knock on effects on other organisms [68]. The AMO has a secondary influence on phytoplankton biomass. Some care must be taken in interpreting these results however, since the phytoplankton colour index is merely a representation of total phytoplankton biomass [13] and it is known that not all species of phytoplankton respond to climate in the same way [99].

Plankton are also important to the ecosystem due to their influence other marine organisms through the coupling of pelagic and benthic systems [88]. The abundance of *Echinoderm larvae*, which are the offspring of the group of marine organisms that includes starfish, is also predicted to be increasing in the northern North Sea. If the rate of survival to adulthood remains constant then this will have an impact on the adults and thus effect the composition of the benthic system. This again demonstrates the importance of understanding the behaviour of the plankton in order to make projections about how possible changes in climate may affect the marine ecosystem.

One advantage of this multivariate analysis over analysing individual species is that the sensitivity to different drivers can be tested across multiple species without the computational intensity of performing a regression analysis for each individual species at each location. This is due to the fact that most of the variation is explained by five or fewer components. This part of the research is novel in that an extensive analysis of the CPR data across species, space and time has never been carried out before. The techniques used in this study allow one to make use of the extensive data available and to summarise complex structure in an interpretable way.

In carrying out the community analysis various challenges were presented. The first was how to determine the number of species to retain for each component. A mixture model was used to estimate this number and then the method was carried on out pre-processed data in order to assess the validity. The WinCPR dataset was

used to select the mixture model for the loading vector by showing that a Laplace mixture captured the distribution well (see figure 4.1). Further modelling challenges were presented when analysing the raw dataset. It became apparent that there might be difficulties with combining different types of species in a single analysis. Since zooplankton and phytoplankton have different biomasses and are counted in different ways the weights on these species might not be representative (see figure 6.4). The loading vectors placed a much higher weight on the phytoplankton species than the zooplankton, suggesting that the results might not be informative about the behaviour of the zooplankton and that it is inadvisable to compare abundances of phytoplankton and zooplankton without taking in to account their relative magnitudes. Therefore the analysis was also carried out on just the zooplankton and just the Diatoms in addition to the whole community.

8.1.2 Temporal Structure

In this study PCA was used to find the dominant trends in both sea surface temperature and plankton assemblages across time. For the sea surface temperature data PCA proves to be a useful tool in separating the spatio-temporal trends. In particular two of the dominant modes of variability in the sea surface temperature data were shown to correspond to known climate indices. After detrending the first and third largest sources of variation can be matched to the AMO and the NAO respectively both in their spatial and temporal representations (see figure 3.2). The second component cannot be identified with a known climate oscillation, although the north-south dipole in its spatial influence resembles the spatial pattern of the EAP [43, 165]. Several possible explanations for what may be driving the second component were proposed but none of these was verifiable.

The study then focused on the dominant trends across both indicator species and species communities for the CPR dataset. Using the results of the sparse PCA on the species assemblages the influence of both the recent warming trend and nat-

ural climate oscillations was explored. The initial analysis using the pre-processed WinCPR dataset demonstrated the importance of taking in to account misalignment when working across different species, in particular when comparing the dominant trends with climate signals (see figure 4.11). This implies that there are time lags between species. These may arise for a number of reasons: species with differing physiology might respond to changes more or less quickly or there may be lags between species which interact with one another. The latter might be the case for species that predate on others. If the change in the climate variable affects the predated species then the predator's abundance might be impacted. As there is causal link between abundance of predator and the abundance of prey it can be expected that there will be a time delay between the change in abundance of the predated species and the change in abundance of the predator. This suggests that time lags might be useful for understanding species relationships.

Once the time lags had been accounted for it was found that there were strong correlations between the joint responses of the plankton assemblages and various climate drivers across a large number of spatial locations. Whilst the greatest driver of abundance for the zooplankton is the average warming trend [127], for the Diatoms it is the AMO [56] (see figures 5.19 and 5.24). There is a time lag between the average zooplankton response over the whole North Atlantic and the NHT, suggesting perhaps some delay in response times. The AMO also plays a role in driving zooplankton abundance in addition to the warming trend and thus should not be ignored. There is however a high degree of variability in the importance of these drivers in space [108].

For many species assemblages temperature is the most important driver of abundance but for those groups of species that are less directly metabolically affected by changes in temperature other factors, such as the mixed layer depth, play a more important role. Diatoms are known to be sensitive to mixed layer depth [56], which in turn is driven by wind intensity and currents which are thought might be linked to

the Atlantic Multidecadal oscillation. The warming trend might obscure the presence of natural oscillations and so it can often by useful to linearly detrend the sea surface temperature before exploring the influence of these oscillations [141, 90].

An extension of this analysis, which looked at how the abundance of different species of plankton might change under possible changes in the climate drivers, was also presented. These results indicated that a further increase in the warming trend would influence the zooplankton group in particular (see figure 7.13).The model is useful for making short range predictions, which may effect the decisions of policy makers, but since the plankton ecosystem is highly complex continued study is also necessary. With the increase in the number of global surveys it can be hoped that more detailed data over more spatial regions might be available in the future in order to study the long term impact of climate change.

8.1.3 Spatial Structure

Part of this study explored spatial heterogeneity. Spatial patterns were found using a statistical learning approach and without prior knowledge. These results were shown to be interpretable in light of the ecology and oceanography, which supported the validity of the methodology as an exploratory tool. One example of the spatial heterogeneity is in the average warming trend. Unlike the average over the Northern Hemisphere in the subpolar gyre a cooling effect is instead observed, which may be attributable to changes in the flow of currents in this region [118]. The southern North Sea is instead warming at a slightly quicker rate. Consequently any approach to climate modelling that does not take in to account the spatial heterogeneity might lead to misleading results. It has been shown that the influence of climate drivers on both sea surface temperature and plankton is heterogeneous in space [108] and across species. The presence of different physical features, such as changes in the bathymetry and ocean gyres, can determine how influential different drivers are in space [90, 43, 123]. The plankton forms a complex ecosystem

on which most marine life is dependent [68] and so understanding this heterogeneity is vital to both environmental policy makers and member of marine industries [136]. From this it can be shown that there is non-stationarity in the regionalisation of the North Atlantic as defined by plankton assemblages and that certain regions are indeed more vulnerable to different climate variables. This analysis could also be extended to other spatial regions by including additional climate variables, since on a global scale different climate variables are more influential in different regions of the world.

In investigating the influence of climate on individual species across space the concept of ecological niches is important. An ecological niche is the region and environment to which a particular species is well suited [75, 74]. For certain cold water species, such as the Copepod *Calanus finmarchicus*, the North Sea might lie at the edge of an ecological niche making them particularly vulnerable to changes in climate in this region [75, 74]. It is thought that these rising temperatures will impact the spatial distribution of species such as *C. finmarchicus* and that given its importance to the ocean habitat that this will have a major effect on the entire marine system [132] and the spatial distribution of many important marine organisms [57, 27]. For many other species temperature is also a the primary driver in their spatial distribution [117, 21], meaning these sort of changes can be expected on a large scale should the warming trend continue. In our study we observe that there are regions in which certain species are more vulnerable to changes in climate.

The spatial structure is also explored across species communities. The strength of the community analysis is that, in addition to gaining insights in to how different drivers effect abundance of plankton across space, it allows one to visualise the complex spatio-temporal structure across species groups. This in turn allows the identification of ecoregions defined on species groups rather than just individual species and to understand how these ecoregions might change, which provides entirely new insights in to the CPR dataset. The output of the sparse PCA also

produces summaries of the assemblage structure across space, which are given in terms of the sparsity parameter and the number of components. The estimates of the sparsity parameter, which is a representation of the number of members of each assemblage, for the WinCPR data from this model are spatially interpretable. It can be shown that across all species using the WinCPR data the sparsity parameter and the number of principal components, which is a representation of the number of distinct assemblages, are highest in the north west of the North Sea. This is the region where the oceanic waters mix with the waters from the North Sea, implying that there will be a mixture of different species types in this region.

The number of principal components, which could be seen as a representation of the number of functional groups, also shows some structure in latitude. The number of groups is particularly high in the southern North Sea and the bay of Biscay, suggesting higher zooplankton diversity in these regions. For the Diatoms the major driver of the spatial distribution is the bathymetry. It may be that they are sensitive to changes in currents driven by the bathymetry. The sparsity parameter for the zooplankton species is higher in shallower waters, particularly in the bay of Biscay.

Cluster analysis on both the output of the sparse PCA on the WinCPR dataset and on the raw data produces spatially coherent and biologically meaningful regions. Clustering on the species representation for the WinCPR data, for example, shows a north-south division in the North Sea, which may be due to differences in temperature and bathymetry. Similarly interpretable spatial patterns are found using the raw data, with the regions on the zooplankton species being driven by temperature and the regions on the diatoms relating to the bathymetry. The spatial pattern of the regions based on the species for the zooplankton for the raw shows a north-south divide (figure 6.9), for example. One potential challenge is how one can be certain that these spatial patterns are not simply due to the interpolation methods used. A simulation study can be used to verify that these spatial patterns

are features of the data rather than an effect of the smoothing (see figures 5.27 and 5.28).

This analysis was then used in order to find whether there could be said to be a 'regime shift' in the ecoregions of the North East Atlantic defined by dominant species of plankton. The community analysis also demonstrates how spatial patterns of plankton are changing over time. The north North Sea is a particular region where change is occurring, with warm water species increasing in abundance as temperatures rise and cold water species declining (see figure 5.20). Recent studies predict a northwards movement of species of zooplankton to be occurring [132, 127, 166, 75] and changes in the regime of the North Atlantic to be happening as a result [22]. These sort of changes could have a profound effect on the marine ecosystem, for example for fish larvae that prey upon the zooplankton [27]. Another effect of the changing regime in the North Atlantic is the detection of species that have not been observed in this region previously in post-industrial era data [131]. Our results show that clusters on the zooplankton species have indeed shifted northwards post 1985 (see figure 6.9). For the Diatoms temperature is a less important driver and so this northwards movement is not observed. There is however increasing structure seen in the spatial pattern, which seems to follow the bathymetry, which may be attributable to the positive phase of the AMO (figure 6.14). When investigating the spatial vulnerability of species the northern North Sea was revealed to be a 'hotspot' for the zooplankton. In this analysis joint responses of the plankton were assumed to be linearly dependent on climate variables, with the coefficients being estimated by regression. The linear regression model has some limitations in its assumptions, meaning it can not be used to predict changes over long time periods. In particular it may be inadvisable to assume that the linear relationship will hold for large perturbations. This means that continued monitoring, particularly in vulnerable regions, of plankton ecosystems is required in order best understand how the environment might change over time.

In the final part of the analysis variability at smaller spatial scales was briefly explored. Although at a large scale the warming trend and the AMO are the primary drivers of zooplankton and Diatom abundance respectively, the multiscale approach demonstrates how other influences such as the NAO might also have an effect. In the north west North Atlantic and to a lesser extent other parts of the North Atlantic the principal components for the zooplankton found once the average trend has been removed correlate with the NAO. The North Atlantic Oscillation is known to be a driver of plankton abundance [122] but it may be that at a large scale its effects are obscured by the dominating influence of the NHT trend on abundance. For the Diatom species the structure becomes more complex after the average signal is removed, suggesting more variability at a local level. There can be found some correlation between the resulting principal components and the NHT and NAO trends, which was not apparent before the average signal was subtracted. There is the opportunity for further investigation to be carried out using these tools, in particular looking at physical drivers that might have an influence on more localised behaviour.

8.2 Possibilities for Further Study

In the first part of the analysis the spatio-temporal structure of the sea surface temperature was explored. Since the second trend in the sea surface temperature data explained a large proportion of the variation it would be of great interest to establish what may be driving it and so this may warrant further investigation. Further studies of the spatial heterogeneity of climate indices might look at other spatial regions and how the influences of different climate indices can change or even at the influence of different climate indices at different scales, in order to determine which smaller scale effects might influence local climate. Regions with more complex local climate might be better understood by investigating more localised effects.

Long term changes in climate warrant further monitoring. It is known that changes in climate in certain regions may have a detrimental effect on the environment as a whole [166] but it is not certain whether current patterns will continue or whether these will change. For example the subpolar gyre may not continue to cool. By continuing to monitor these effects the long term consequences may be better understood.

The second part of the plankton analysis focused on the raw CPR dataset, which unlike the WinCPR dataset analysed earlier had not been interpolated on to a regular spatial grid. The challenges associated with modelling such a complex dataset were discussed in detail. Amongst these challenges is how best to approach the irregularity of the spatial sampling. One approach is the inverse distance interpolation method used to produce the WinCPR dataset. Another is Kernel smoothing, which allows one to control the amount the data is smoothed through a bandwidth parameter. In this study when studying the plankton data Kernel smoothing was first used to both estimate missing values and to reduce the effects of noise. In this case a fixed bandwidth was used across space but an alternative approach might have been to vary the bandwidth according to the amount of sampling in a region and to thus make use of the fact that certain regions were better sampled. An extension of this may be to investigate how fine a scale it is possible to study the data at before the spatial irregularity of the sampling becomes problematic. Too small a bandwidth might not remove the effect of the sampling transects from the resulting dataset. However with the wealth of data available it would be of interest to study the influence of small scale features, such as localised nutrients and pollution, on the plankton, for which a fine resolution would be useful. A spatially varying bandwidth would preserve the detail available along the transects, whilst accounting for the fact that some regions have almost no samples taken from them.

This study also highlighted the importance of taking in to account time delays between species when studying temporal behaviour across communities. This in-

teresting result highlights a number of possible avenues for further study. Further investigation might focus in more depth upon differences in time delays between species, for example. Time delays between species might be used in prediction in order to investigate how changes between one species might influence another. On the monthly data phase shifts could be used to investigate whether seasonal blooms have moved forward with rising average temperatures. This is of particular interest when studying predator-prey relationships, as changes in seasonal blooms might lead to misalignment between the annual cycles of predators and prey.

The relationship between plankton and climate was studied extensively in this analysis. One particularly interesting result of this was the strength of the relationship between the diatom communities and the AMO. It has been speculated that this may be due to the influence of mixed layer depth of the diatoms [56] but the mechanism is not well understood and as such warrants further investigation.

Further extension of the community analysis could be used to investigate species interactions. This research did not investigate whether species co-occurrence was due to them responding to similar climate trends or was due to species interactions. Other areas of ecology have developed methods for approaching species interactions, which could be applied to the CPR dataset. Ovaskainen [124] explores species interactions in multivariate ecological datasets. When two or more species occur together in a region it may be that they are interacting or that they are responding to the same climate trend. Our study does not presently differentiate between these two scenarios. Ovaskainen's approach is to use multivariate logistic regression on presence or absence data, where the value for each species is 1 if it has been observed and 0 is there are no observations so that the response is binary, for that species in order to determine whether a particular species is under or over-represented in a region according to what is predicted from the covariates. A similar approach might be applicable to the CPR dataset.

Finally whilst the concept of different drivers at different spatial scales has been

touched upon, the effects of local influences such as nutrient influxes were not investigated. This study focuses upon the influences of long term temperature fluctuations and the influence of natural climate oscillations but the spatial resolution of the data would allow for the study of more local influences as well. The effect of smoothing was discussed and one possible avenue of exploration is how fine a resolution it would be possible to model whilst still accommodating for the irregularity of the sampling. Data on nutrient fluxes and currents could be incorporated in to the model that has already been developed in order to better analyse more local features and to understand better the spatial heterogeneity of the plankton.

8.3 Conclusions

In conclusion the community analysis is a useful tool in both confirming previously stated hypotheses about changes in the spatio-temporal structure and for providing new insight in to the behaviour of the ecosystem, including a community level analysis. The plankton form a complex system, in which variability across spatial, temporal and species dimensions must all be taken in to consideration. This work has involved developing techniques for producing summaries of the data across all of these dimensions, which has not been done before. From this strong evidence is brought forward about the nature of the changing ecosystem in the North Atlantic, changes which could have severe consequences for all marine life. Finally although this analysis focused upon one dataset in particular, the tools and techniques could be applied to other types of ecological data in order to gain a deeper understanding over complex ecological structure. The issue of climate change is likely to be of great importance to policy makers over coming decades, which may lead to more surveys like the CPR being required to understand the complex relationship between climate and the environment. Long term changes can be expected to be seen in many different ecosystems, particularly if the warming trend contin-

ues, but these changes can be non-intuitive due to the multitude of variables and possible covariates as well as the heterogeneity of different responses. Together this makes the development and application of a range of statistical tools essential to understanding these systems as a whole.

References

[1] J.A. Adams. The primary ecological sub-divisions of the north sea: some aspects of their plankton communities. *Developments in fisheries research in Scotland. Fishing News (Books) Ltd, London*, 1987.

[2] J. Aldrich. R. A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, AUG 1997.

[3] R. Allen. To investigate the response of planktonic ecosystems to climatic regimes in the North sea and Irish sea using the continuous plankton recorder database. *Ocean data symposium Dublin Castle (Ireland), 15-18 Oct 1997. Marine Institute (Ireland), Marine Data Centre*, pages 90–91, 1998.

[4] M.C. Austen, J.B. Buchanan, H.G. Hunt, A.B. Josefson, and M.A. Kendall. Comparison of long-term trends in benthic and pelagic communities of the North Sea. *Journal of the Marine Biological Association of the United Kingdom*, 71:179–190, 1991.

[5] V. Bainbridge. Warm-Water Species in the Plankton off Newfoundland During the Winter months. *Nature*, 191:1216–1217, 1961.

[6] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *The Journal of Machine Learning Research archive*, 9:485–516, 2008.

[7] R. Barnard, S. Batten, G. Beaugrand, C. Buckland, D.V.P. Conway, M. Edwards, J. Finlayson, L.W. Gregory, N.C. Halliday, A.W.G. John, D.G. Johns, A.D. Johnson, T.D. Jonas, J.A. Lindley, J. Nyman, P. Pritchard, P.C. Reid, A.J. Richardson, R.E. Saxby, J. Sidey, M.A. Smith, D.P. Stevens, C.M. Taylor, P.R.G. Tranter, A.W. Walne, M. Wootton, C.O.M. Wotton, and J.C. Wright. Continuous Plankton Records: Plankton Atlas of the North Atlantic Ocean (1958-1999). II Biogeographical Charts. *Marine Ecology Progress Series*, pages 11–75, 2004.

[8] A. G. Barnston and R. E. Livezey. Classification, Seasonality and Persistance of Low-Frequency Atmospheric Circulation Patterns. *Monthly Weather Review*, 115:1083–1126, 1986.

[9] B.M. Bary. Temperature, Salinity and Plankton in the Eastern North Atlantic and Coastal Waters of Britain, 1957. II. The relationships between species and water bodies. *Journal of the Fisheries Research Board of Canada*, 20:1031–1065, 1963.

[10] S. Batten. Changes in the oceanic northeast Pacific plankton populations; what may happen in a warmer ocean. 2007.

[11] S. D. Batten, R. Clark, J. Flinkman, G. Hays, E. John, A. W. G. John, T. Jona, J. A. Lindley, D. P. Stevens, and A. Walne. CPR Sampling: the Technical Background, Materials and Methods, Consistency and Comparability. *Progress in Oceanography*, 58:193–215, 2003.

[12] S. D. Batten and A. W. Walne. Variability in Northwards Extension of Warm Water Copepods in the NE Pacific. *Journal of Plankton Research*, 33(11):1643–1653, 2011.

[13] S.D. Batten, A.W. Walne, M. Edwards, and S.B. Groom. Phytoplankton biomass from continuous plankton recorder data: an assessment of the phytoplankton colour index. *Journal of Plankton Research*, 25:697–702, 2003.

[14] D. J. Beare, S. D. Batten, M. Edwards, E. McKenzie, P. C. Reid, and D. G. Reid. Summarising Spatial and Temporal Information in the CPR Data. *Progress in Oceanography*, 58:217–233, 2003.

[15] G. Beaugrand. SAHFOS and the CPR survey. Use of CPR derived plankton indicators to monitor response of pelagic ecosystems to climate change. *Newsletter. GLOBEC International*, 8(39), 2002.

[16] G. Beaugrand. Long-term changes in copepod abundance and diversity in the north-east Atlantic in relation to fluctuations in the hydroclimatic environment. *Fisheries Oceanography*, 12(4-5):270–283, SEP 2003. 2nd GLOBEC Open Science Meeting, Qingdao, Peoples R China, OCT 15-18, 2002.

[17] G. Beaugrand. The North Sea Regime Shift: Evidence, mechanisms, causes and consequences. *Progress in Oceanography*, 60:245–262, 2004.

[18] G. Beaugrand. Decadal changes in climate and ecosystems in the North Atlantic Ocean and adjacent seas. *Deep-Sea Research Part II-Topical Studies In Oceanography*, 56(8-10):656–673, APR 2009.

[19] G. Beaugrand. Unanticipated biological changes and global warming. *Marine Ecology-Progress Series*, 445:293–301, 2012.

[20] G. Beaugrand and M. Edwards. Differences in performance amongst four indices used to evaluate diversity in planktonic ecosystems. *Oceanologica Acta*, 24(4), 2001.

[21] G. Beaugrand, M. Edwards, K. Brander, C. Luczak, and F. Ibanez. Causes and projections of abrupt climate-driven ecosystem shifts in the North Atlantic. *Ecology Letters*, 11(11):1157–1168, 2008.

[22] G. Beaugrand, M. Edwards, K. Brander, C. Luczak, and F. Ibanez. Causes and projections of abrupt climate-driven ecosystem shifts in the North Atlantic. *Ecology Letters*, 11:1157–1168, 2008.

[23] G. Beaugrand and P. Helaouet. Simple procedures to assess and compare the ecological niche of species. *Marine Ecology-Progress Series*, 363:29–37, 2008.

[24] G. Beaugrand, F. Ibaez, and P. C. Reid. Spatial, seasonal and long-term fluctuations of plankton in relation to hydroclimatic features in the English Channel, Celtic Sea and Bay of Biscay. *Marine Ecology Progress Series*, 200:93–102, 2000.

[25] G. Beaugrand, A. F. Ibaezb, and J. A. Lindleya. An Overview of Statistical Methods Applied to CPR Data. *Progress in Oceanography*, 58:235–262, 2003.

[26] G. Beaugrand, F. Ibanez, and J. A. Lindley. Geographical distribution and seasonal and diel changes in the diversity of calanoid copepods in the North Atlantic and North Sea. *Marine Ecology-Progress Series*, 219:189–203, 2001.

[27] G. Beaugrand and R. R. Kirby. Climate, plankton and cod. *Global Change Biology*, 16(4):1268–1280, 2010.

[28] G. Beaugrand and R. R. Kirby. Spatial changes in the sensitivity of Atlantic cod to climate-driven effects in the plankton. *Climate Research*, 41(1):15–19, 2010.

[29] G. Beaugrand and R.R. Kirby. Spatial changes in the sensitivity of Atlantic cod to climate-driven effects in the plankton. *Climate research*, 41:15–19, 2010.

[30] G. Beaugrand, C. Luczak, and M. Edwards. Rapid Biogeographical Plankton Shifts in the North Atlantic Ocean. *Global Change Biology*, 15(7):1790, 2009.

[31] G. Beaugrand and P. C. Reid. Long-term Changes in Phytoplankton, Zooplankton and Salmon Related to Climate. *Global Change Biology*, 9:801–817, 2003.

[32] G. Beaugrand, P. C. Reid, F. Ibanez, J. A. Lindley, and M. Edwards. Reorganization of North Atlantic marine copepod biodiversity and climate. *Science*, 296(5573):1692–1694, MAY 31 2002.

[33] G. Beaugrand, P.C. Reid, and F. Ibanez. Major reorganisation of North Atlantic pelagic ecosystems linked to climate change. *GLOBEC International Newsletter*, 8:30–33, 2002.

[34] L. Bel, D. Allard, J. M. Laurent, R. Cheddadi, and A. Bar-Hen. CART algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics & Data Aanalysis*, 53(8):3082–3093, JUN 15 2009.

[35] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Soceity. Series B.*, 57(1):289–300, 1995.

[36] Y. Benjamini and D. Yekutieli. The control of the false discover rate in multiple testing under ddependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[37] D. Bonnet. An overview of *Calanus helgolandicus* ecology in European waters. *Progress in Oceanography*, 65:1–53, 2005.

[38] B. B. B. Booth, N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin. Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature*, 484(7393):228–32, 2012.

[39] Z. Botta-Dukat, E. Kovacs-Lang, T. Redei, M. Kertesz, and J. Garadnai. Statistical and biological consequences of preferential sampling in phytosociology: Theoretical considerations and a case study. *Folia Geobotanica*, 42(2):141–152, 2007.

[40] N. Bouguila, K. Almakadmeh, and S. Boutemedjet. A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection. *Expert Systems With Applications*, 39(7):6641–6656, JUN 1 2012.

[41] D. G. Boyce, M. R. Lewis, and B. Worm. Global phytoplankton decline over the past century. *Nature*, 466:591–596, 2010.

[42] K. Brander and M. Edwards. Indicator 12: Northerly movement of marine species. *EEA-JRC-WHO, Copenhagen*, pages 1–5, 2008.

[43] H. Cannaby and Y. S. Husrevoglu. The influence of low-frequency variability and long term trends in North Atlantic sea surface temperature on Irish waters. *International Council for Exploration of the Sea*, pages 1480–1489, 2009.

[44] C. Carvalho, J. Chang, J. Lucas, J. R. Nevins, Q. Wang, and M. West. High Dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103, 2008.

[45] E. Cavatorta. Unobserved Common Factors in Military Expenditure Interactions Across Mena Countries. *Defence and Peace Economics*, 21(4):301–316, 2010.

[46] P. Chylek, C. Folland, L. Frankcombe, H. Dijkstra, G. Lesins, and M. Dubey. Greenland ice core evidence for spatial and temporal variability of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters*, 39, MAY 3 2012.

[47] J. R. Clark, S. J. Daines, T. M. Lenton, A. J. Watson, and H. T. P. Williams. Individual-based modelling of adaptation in marine microbial populations using genetically defined physiological parameters. *Ecological Modelling*, 222(23-24):3823–3837, DEC 10 2011.

[48] P.A. Clark. Exploration of the processes that contribute to observed alternating abundance of *Calanus finmarchicus* and *Calanus helgolandicus* in the North Sea. *University of Plymouth, Plymouth*, page 99, 2002.

[49] A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal Solutions for Sparse Principal Component Analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

[50] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, 49:434, 2007.

[51] P. J. Diggle, R. M. Da Mota Leite, and T. Su. Geostatistical Analysis under Preferential Sampling. *Applied Statistics*, 2010(59):1–20, 2006.

[52] J. M. Drake and B. D. Griffen. Early warning signals of extinction in deteriorating environments. *Nature*, 467:456–459, 2010.

[53] M. Edwards, G. Beaugrand, G. C. Hays, J. A. Koslow, and A. J. Richardson. Multi-decadal oceanic ecological datasets and their application in marine policy and management. *Trends in Ecology and Evolution*, 25:602–610, 2010.

[54] M. Edwards, P. Reid, and B. Planque. Long-term and regional variability of phytoplankton biomass in the Northeast Atlantic (1960-1995). *ICES Journal of Marine Science*, 58(1):39–49, FEB 2001.

[55] D. B. Enfield, A. M. Mestas-Nunez, and P. J. Trimble. The Atlantic Multidecadal Oscillation and its Relation to Rainfall and River Flows in the Continental U.S. *Geophysical Reasearch Letters*, 28(10):2077–2080, 2001.

[56] P. G. Falkowski and M. J. Oliver. Mix and match: how climate selects phytoplankton. *Nature Reviews Microbiology 5*, pages 813–819, 2007.

[57] E. Fernndez, J. Cabal, J. L. Acua, A. Bode, A. Botas, and C. Garca-Soto. Plankton distribution across a slope current-induced front in the southern Bay of Biscay. *Journal of Plankton Research*, 15(6):619–641, 1993.

[58] S. Frhwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

[59] J. M. Fromentin and Planque B. Calanus and Environment in the Eastern North Atlantic: Role of the North Atlantic Oscillation on *calanus finmarchicus* and *c. helgolandicus*. *Marine Ecology Progress Series*, 134:11–118, 1996.

[60] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms and Applications*. Society for Industrial and Applied Mathematics, 2007.

[61] F. P. Garcia Marquez and I. Pena Garcia-Pardo. Principal Component Analysis Applied to Filtered Signals for Maintenance Management. *Quality and Reliability Engineering International*, 26(6):523–527, OCT 2010.

[62] C. Garcia-Soto and R. D. Pingree. Atlantic Multidecadal Oscillation (AMO) and sea surface temperature in the Bay of Biscay and adjacent regions. *Journal of the Marine Biological Association of the United Kingdom*, 92(2):213–234, MAR 2012.

[63] K. J. Gaston and T. M. Blackburn. *Pattern and Progress in Macroecology*. Blackwell Science, 2000.

[64] C. Gebuehr, K. H. Wiltshire, N. Aberle, J. E. E. van Beusekom, and G. Gerdts. Influence of nutrients, temperature, light and salinity on the occurrence of *Paralia sulcata* at Helgoland Roads, North Sea. *Aquatic Biology*, 7(3):185–197, 2009.

[65] D. Gerten and R. Adrian. Effects of climate warming, North Atlantic Oscillation, and El Nino-Southern Oscillation on thermal conditions and plankton dynamics in northern hemispheric lakes. *The Scientific World Journal*, 2:586–606, 2002.

[66] L. Gertzen, E and B. Leung. Predicting the spread of invasive species in an uncertain world: accommodating multiple vectors and gaps in temporal and spatial data for Bythotrephes longimanus. *Biological Invasions*, 13(11, SI):2433–2444, 2011.

[67] J. P. Grist, S. A. Josey, R. Marsh, S. A. Good, A. C. Coward, B. A. de Cuevas, S. G. Alderson, A. L. New, and G. Madec. The roles of surface heat flux and ocean heat transport convergence in determining Atlantic Ocean temperature variability. *Ocean Dynamics*, 60(4):771–790, AUG 2010.

[68] A. Hardy. *The Open Sea: Its Natural History. Part I: The World of Plankton*. The Fontana New Naturalist, 1971.

[69] A. C. Hardy. The Discovery Expedition. A new method of plankton research. *Nature*, 118:630–632, 1926.

[70] A. C. Hardy. The Continuous Plankton Recorder: a new method of survey. *Rapports et Proces-Verbaux des Reunions. Conseil International pour l'Exploration de la Mer*, 95, 1935.

[71] A. C. Hardy. Ecological investigations with the Continuous Plankton Recorder: Object, plan and methods. *Hull Bulletins of Marine Ecology*, 1:1–57, 1939.

[72] V. Harris, E. Edwards, and S. C. Olhede. Multidecadal Atlantic Climate Variability and its Impact on Marine Pelagic Communities. *Journal of Marine Systems*, 2012 (In Revision).

[73] E. J. H. Head and P. Pepin. Spatial and inter-decadal variability in plankton abundance and composition in the Northwest Atlantic (1958-2006). *Journal of Plankton Research*, 32(12):1633–1648, DEC 2010.

[74] P. Helaouat, G. Beaugrand, and P. C. Reid. Macrophysiology of *Calanus finmarchicus* in the North Atlantic Ocean. *Progress In Oceanography*, 91(3):217 – 228, 2011.

[75] P. Helaout and G. Beaugrand. Physiology, Ecological Niches and Species Distribution. *Ecosystems*, 12:1235–1245, 2009. 10.1007/s10021-009-9261-5.

[76] J. W. Hurrell. Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science*, 269(5224):676–679, 1995.

[77] J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck. The North Atlantic Oscillation: Climatic Significance and Environmental Impact. *Geophysical Monograph, the American Geophysical Union*, 134, 2003.

[78] info@marinespecies.org. World Register of Marine Species. http://www.marinespecies.org/.

[79] Giovannoni S. J. and Vergin K. L. Seasonality in Ocean Microbial Communities. *Science*, 335(6069):671–676, 2012.

[80] M. G. Jafari and M. D. Plumbley. Fast Dictionary Learning for Sparse Representations of Speech Signals. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):1025–1031, SEP 2011.

[81] A. K. Jain, Murty M. N., and Flynn P. J. Data Clustering: A Review. *ACM Computing Surveys*, 31(3), 1999.

[82] I. M. Johnstone and B. W. Silverman. Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences. *The Annals of Statistics*, 32:1594–1649, 2004.

[83] I. M. Johnstone and B. W. Silverman. Empirical Bayes Selection of Wavelet Thresholds. *The Annals of Statistics*, 33:1700–1752, 2005.

[84] I. Joint, R. Wollast, L. Chou, S. Batten, M. Elskens, E. Edwards, A. Hirst, P. Burkill, S. Groom, S. Gibb, A. Miller, D. Hydes, F. Dehairs, A. Antia, R. Barlow, A. Rees, A. Pomroy, U. Brockmann, D. Cummings, R. Lampitt, M. Loijens, F. Mantoura, P. Miller, T. Raabe, X. Alvarez-Salgado, C. Stelfox, and J. Woolfenden. Pelagic production at the Celtic Sea shelf break. *Deep-Sea Research Part II-Topical Studies in Oceanography*, 48(14-15):3049–3081, 2001.

[85] I. T. Joliffe. *Principal Component Analysis*. Springer, 2004.

[86] I. T. Joliffe, N. T. Trendafilov, and M. Uddin. A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[87] P. D. Jones, T. Jonsson, and D. Wheeler. Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and south-west Iceland. *International Journal of Climatology*, 17(13):1433–1450, 1997.

[88] R. R. Kirby, G. Beaugrand, J. A. Lindley, A. J. Richardson, M. Edwards, and P. C. Reid. Climate effects and benthic-pelagic coupling in the North Sea. *Marine Ecology-Progress Series*, 330:31–38, 2007.

[89] J. R. Knight, C. K. Folland, and A. A. Scaife1. Climate Impacts of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters*, 33, 2006.

[90] M. Knudsen, M. Seidenkrantz, B. Jacobsen, and A. Kuijpers. Tracking the Atlantic Multidecadal Oscillation Through the Last 8,000 Years. *Nature Communications*, 2(178), 2011.

[91] T. G. Kolday and B. W. Baderz. Tensor Decompositions and Applications. *SIAM Review*, 2008.

[92] Hinder S. L., Manning J. E., and Gravenor M. B. Long-term changes in abundance and distribution of microzooplankton in the NE Atlantic and North Sea. *Journal of Plankton Research*, 34(1):83–91, 2012.

[93] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254–4278, 2009.

[94] H. R. Langehaug, I. Medhaug, T. Eldevik, and O. H. Ottera. Arctic/Atlantic Exchanges via the Subpolar Gyre. *Journal of Climate*, 25(7):2421–2439, APR 2012.

[95] P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier, 2003.

[96] T. M. Letcher, editor. *Climate Change: Observed Impacts on Planet Earth*. Elsevier, 2009.

[97] T. M. Letcher, editor. *Climate Change: Observed Impacts on Planet Earth*. Elsevier, 2009.

[98] S. C. Leterme, R. D. Pincree, M. D. Skogen, L. Seuront, P. C. Reid, and M. J. Attrill. Decadal fluctuations in North Atlantic water inflow in the North Sea between 1958-2003: impacts on temperature and phytoplankton populations. *Oceanologica*, 50(1):59–72, 2008.

[99] S. C. Leterme, L. Seuront, and M. Edwards. Differential contribution of diatoms and dinoflagellates to phytoplankton in the NE Atlantic Ocean and the North Sea. *Marine Ecology-Progress Series*, 312:57–65, 2006.

[100] T. B. Letessier, M. J. Cox, and A. S. Brierley. Drivers of *Euphausiid* species abundance and numerical abundance in the Atlantic Ocean. *Marine Biology*, 156(12):2539–2553, NOV 2009.

[101] Q. Li and N. Lin. The Bayesian Elastic Net. *Bayesian Analysis*, 5(1):151–170, 2010.

[102] X. Liu and M. C. K. Yang. Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis*, 53:1361–1376, 2009.

[103] A. Longhurst. *Ecological Geography of the Sea*. Elsevier, 2007.

[104] H. Luo, A. Bracco, and E. D. Lorenzo. The interannual variability of the surface eddy kinetic energy in the Labrador Sea. *Progress In Oceanography*, 91(3):295 – 311, 2011.

[105] M. Maar, E. F. Moller, J. Larsen, K. S. Madsen, Z. Wan, J. She, L. Jonasson, and T. Neumann. Ecosystem modelling across a salinity gradient from the North Sea to the Baltic Sea. *Ecological Modelling*, 222(10):1696–1711, MAY 24 2011.

[106] E. Martinez, D. Antoine, F. D'Ortenzio, and C. de B. Montegut. Phytoplankton spring and fall blooms in the North Atlantic in the 1980s and 2000s. *Journal of Geophysical Research-Oceans*, 116, NOV 19 2011.

[107] R. M. May. Bifurcations and dynamic complexity in simple ecological models. *American Naturalist*, 110(974):573–599, 1976.

[108] N. McGinty, A. M. Power, and M. P. Johnson. Variation among northeast Atlantic regions in the responses of zooplankton to climate change: Not all areas follow the same path. *Journal of Experimental Marine Biology and Ecology*, 400(1-2, SI):120–131, APR 30 2011.

[109] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley-Blackwell, 2000.

[110] P. D. McNicholas and S. Subedi. Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, 142(5):1114–1127, MAY 2012.

[111] A. McQuatters-Gollop, D. E. Raitsos, M. Edwards, and M. J. Attrill. Spatial patterns of diatom and dinoflagellate seasonal cycles in the NE Atlantic Ocean. *Marine Ecology-Progress Series*, 339:301–306, 2007.

[112] A. McQuatters-Gollop, P. C. Reid, M. Edwards, P. H. Burkill, C. Castellani, S. Batten, W. Gieskes, D. Beare, R. R. Bidigare, E. Head, R. Johnson, M. Kahru, A. Koslow, and A. Pena. Is there a decline in marine phytoplankton? *Nature*, 472:6–7, 2011.

[113] D. Michalcova, S. Lvoncik, M. Chytry, and O. Hajek. Bias in vegetation databases? A comparison of stratified-random and preferential sampling. *Journal of Vegetation Science*, 22(2):281–291, APR 2011.

[114] S. Minobe1, A. Kuwano-Yoshida, N Komori, S. Xie, and R. J. Small. Influence of the Gulf Stream on the Troposphere. *Nature*, 452:206–209, 2008.

[115] C. Moffat, R. Emmerson, A. Weiss, C. Symon, and L. Dicks. Quality status report. *OSPAR Commission. London.*, page 176, 2010.

[116] B. Moghaddam, Y. Weiss, and S. Avidan. Spectral Bounds for Sparse PCA: Exact and Greedy Algorithms. *Advances in Neural Information Processing Systems*, 18, 2006.

[117] E. F. Moller, M. Maar, S. H. Jonasdottir, T. G. Nielsen, and K. Tonnesson. The effect of changes in temperature and food on the development of *Calanus finmarchicus* and *Calanus helgolandicus* populations. *Limnology and Oceanography*, 57(1):211–220, JAN 2012.

[118] J. C. Montero-Serrano, N. Frank, C. Colin, C. Wienberg, and M. Eisele. The climate influence on the mid-depth Northeast Atlantic gyres viewed by cold-water corals. *Geophysical Reasearch Letters*, 38, 2011.

[119] R. Msadek and C. Frankignoul. Atlantic multidecadal oceanic variability and its influence on the atmosphere in a climate model. *Climate Dynamics*, 33:45–62, 2009. 10.1007/s00382-008-0452-0.

[120] F. Munoz. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. *Ecological Modelling*, 220(20):2683–2689, OCT 24 2009.

[121] National Oceanographic and Atmospheric Administration. List of Climate Indices. http://www.esrl.noaa.gov/psd/data/climateindices/list/.

[122] G. Ottersen, B. Planque, A. Belgrano, E. Post, P. C. Reid, and N. C. Stenseth. Ecological Effects of the North Atlantic Oscillation. *Oecologia*, 2001.

[123] L. Otto, J.T.F. Zimmerman, G.K. Furnes, M. Mork, R. Saetre, and G. Becker. Review of the physical oceanography of the North Sea. *Netherlands Journal of Sea Research*, 26(2-4):161 – 238, 1990.

[124] O. Ovaskainen, J. Hottola, and J. Siitonen. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91(9):2514–2521, 2010.

[125] R. T. Paine. The Pisaster-Tegula Interaction: Prey Patches, Predator Food Preference, and Intertidal Community Structure. *Ecological Society of America*, 50(6):950–961, 1969.

[126] M. Park, W. andLatif. Ocean Dynamics and the Nature of AirSea Interactions over the North Atlantic at Decadal Time Scales. *Journal of Climate*, 18(7), 2005.

[127] C. J. M. Philippart, R. Anadon, R. Danovaro, J. W. Dippner, S. J. Drinkwater, K. F.and Hawkins, T. Oguz, G. O'Sullivan, and P. C. Reid. Impacts of climate change on European marine ecosystems: Observations, expectations and indicators. *Journal of Experimental Marine Biology and Ecology*, 400(1–2):52–69, 2011.

[128] M. Pourahmadi. Canonical Correlation and Reduction of Multiple Time Series. *Ann. Inst. Statist. Math.*, 46(4):625–631, 1994.

[129] T. F. Rangel, J. A. F. Diniz-Filho, and L. M. Bini. Towards an integrated computational tool for spatial; analysis in macroecology and biogeography. *Global Ecology and biogeography*, 15:321–327, 2006.

[130] P. C. Reid, A. C. Fischer, E. Lewis-Brown, M. P. Meredith, M. Sparrow, A. J. Andersson, A. Antia, N. R. Bates, U. Bathmann, G. Beaugrand, H. Brix, S. Dye, M. Edwards, T. Furevik, R. Gangsto, H. Hatun, R. R. Hopcroft, M. Kendall, S. Kasten, R. Keeling, C. Le Quere, F. T. Mackenzie, G. Malin, C. Mauritzen, J. Olafsson, C. Paull, E. Rignot, K. Shimada, M. Vogt, C. Wallace, Z. Wang, and R. Washington. Impacts of the Oceans on Climate Change. In Sims, DW, editor, *Advances in Marine Biology*, volume 56 of *Advances in Marine Biology*, pages 1–150. Advances in Marine Biology, 2009.

[131] P. C. Reid, David G. Johns, M. Edwards, M. Starr, M. Poulin, and P. Snoeijs. A biological consequence of reducing Arctic ice cover: arrival of the Pacific diatom *Neodenticula seminae* in the North Atlantic for the first time in 800,000 years. *Global Change Biology*, 13(9):1910–1921, SEP 2007.

[132] G. Reygondeau and G. Beaugrand. Future climate-driven shifts in distribution of *Calanus finmarchicus*. *Global Change Biology*, 17(2):756–766, FEB 2011.

[133] J. A. Rice and M Rosenblatt. On Frequency Estimation. *Biometrika*, 75(3):477–484, 1988.

[134] P. M. ROBINSON. Generalized Canonical Analysis for Time Series. *Journal of Multivariate Analysis*, pages 141–160, 1973.

[135] M. J. Rodwell, D. P. Rowell, and C. K. Folland. Oceanic forcing of the wintertime North Atlantic Oscillation and European climate. *Letters to Nature*, 1999.

[136] M. A. Rudd. Scientists Opinions on the Global Status and Management of Biological Diversity. *Conservation Biology*, 25(6):1165–1175, 2011.

[137] SAFHOS. CPR data. http://cpr.cisnr.org/default.asp?page=home.

[138] SAFHOS. Sir Alistair Hardy Foundation for Ocean Science. http://www.sahfos.ac.uk/cpr_survey.htm.

[139] SAHFOS. Marine Ecological Status Report. http://www.sahfos.ac.uk/research/publications/ecological-status-report.aspx.

[140] L. M. Sangali, P. Secchi, S. Vantini, and V. Vitelli. K-means aligment for curve clustering. *MOX-Report*, 13, 2008.

[141] M. E. Schlesinger and N. Ramankutty. An Oscillation in the Global Climate System of Period 6570 Years. *Nature*, 367:723–726, 1994.

[142] A. Schmittner, N. M. Urban, J. D. Shakun, N. M. Mahowald, P. U. Clark, P. J. Bartlein, Alan C. Mix, and A. Rosell-Mel. Climate Sensitivity Estimated from Temperature Reconstructions of the Last Glacial Maximum. *Science*, 334(6061):1385–1388, 2011.

[143] A. O. Shelton, E. J. Dick, D. E. Pearson, S. Ralston, and M. Mangel. Estimating species composition and quantifying uncertainty in multispecies fisheries: hierarchical Bayesian models for stratified sampling protocols with missing data. *Canadian Journal of Fisheries and Aquatic Sciences*, 69(2):231–246, FEB 2012.

[144] J. Slingo and T. Palmer. Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Soceity A-Mathematical Physical and Engineering Sciences*, 369(1956):4751–4767, DEC 13 2011.

[145] M. Spalding, H. Fox, G. Allen, N. Davidson, Z. Ferdana, M. Finlayson, B. Halpern, M. Jorge, A. Lombana, S. Lourie, K. Martin, E. McManus, J. Molnar, C. Recchia, and J. Robertson. Marine Ecoregions of the World: A

Bioregionalisation of Coastal and Shelf Areas. *Bioscience*, 57(7):573–583, 2007.

[146] M. Spencer, S. N. R. Birchenough, N. Mieszkowska, L. A. Robinson, S. D. Simpson, M. T. Burrows, E. Capasso, P. Cleall-Harding, J. Crummy, C. Duck, D. Eloire, M. Frost, A. J. Hall, S. J. Hawkins, D. G. Johns, D. W. Sims, T. J. Smyth, and C. L. J. Frid. Temporal change in UK marine communities: trends or regime shifts? *Marine Ecology-An Evolutionary Perspective*, 32(1):10–24, APR 2011.

[147] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

[148] C. A. Sugar and G. M. James. Finding the Number of Clusters in a Dataset. *Journal of the American Statistical Association*, 98(463):750–763, 2003.

[149] R. Tang and H. G. Muller. Time-synchronized clustering of gene expression trajectories. *Biostatistics*, 10(1):32–45, 2009.

[150] D. W. J. Thompson, J. M. Wallace, J. J. Kennedy, and P. D. Jones. An abrupt drop in Northern Hemisphere sea surface temperature around 1970. *Nature*, 467:444–447, 2010.

[151] R. Tibishirani. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society*, 58(1):267–288, 2008.

[152] R. Tibshirani, M. Saunders, Saharon R., J. Zhu, and K. Knight. Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(1):pp. 91–108, 2005.

[153] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distribution (Wiley Series in Probability and Statistics - Applied Probability and Statistics Section)*. Wiley-Blackwell, 1985.

[154] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley Series in Probability and Mathematical Statistics, 1985.

[155] A. Torstensson, M. Chierici, and A. Wulff. The influence of increased temperature and carbon dioxide levels on the benthic/sea ice diatom *Navicula directa*. *Polar Biology*, 35(2):205–214, FEB 2012.

[156] L. R. Tucker. Implications of factor analysis of three-way matrices for measurement of change. *Problems in Measuring Change*, page 122, 1963.

[157] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279, 1966.

[158] M. Unser and I. Daubechies. On the approximation power of convolution-based least squares versus lnterpolation. *IEEE Transactions on Signal Processing*, 45(7):1697–1711, JUL 1997.

[159] F. Vandeperre, R. M. Higgins, J. Sanchez-Meca, F. Maynou, R. Goni, P. Martin-Sosa, A. Perez-Ruzafa, P. Afonso, I. Bertocci, R. Crec'hriou, G. D'Anna, M. Dimech, C. Dorta, O. Esparza, J. M. Falcon, A. Forcada, I. Guala, L. Le Direach, C. Marcos, C. Ojeda-Martinez, C. Pipitone, P. J. Schembri, V. Stelzenmuller, B. Stobart, and R. S. Santos. Effects of no-take area size and age of marine protected areas on fisheries yields: a meta-analytical approach. *Fish and Fisheries*, 12(4):412–426, DEC 2011.

[160] L. Verzulli and P. C. Reid. The CPR Survey (1948-1997): a Gridded Database Browser of Plankton Abundance in the North Sea. *Progress in Oceanography*, 58:327–336, 2003.

[161] P. Vincent. *The Biogeography of the British Isles: An Introduction*. Routledge, 1990.

[162] N. Vlassis and A. Likas. A Greedy EM Algorithm for Gaussian Mixture Learning. *Neural Processing Letters*, 15:77–87, 2002. 10.1023/A:1013844811137.

[163] P. Walker and E. Wood. *Life in the Sea: The Open Ocean*. Facts on File Inc., 2005.

[164] J. M. Wallace. North atlantic oscillatiodannular mode: Two paradigmsone phenomenon. *Quarterly Journal of the Royal Meteorological Society*, 126(564):791–805, 2000.

[165] J. M. Wallace and D. S. Gutzler. Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter. *Monthly Weather Review*, 109:784–812, 1980.

[166] G. R. Walther, E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee6, J. Fromentin, O. Hoegh-Guldberg, and F. Bairlein. Ecological responses to recent climate change. *Nature*, 416:389–395, 2002.

[167] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

[168] L. Wasserman. *All of Statistics*. Springer, 2004.

[169] L. Wasserman. *All of Non-Parametric Statistics*. Springer, 2006.

[170] H. Yndestad. The Influence of the Lunar Nodal Cycle on Arctic Climate. *ICES Journal of Marine Science*, 63:401–420, 2006.

[171] Z. Zhang, H. Zha, and H. Simon. Low-Rank Approximations with Sparse Factors II: Penalized Methods with Discrete Newton-like Iterations. *SIAM Journal of Matrix Analysis*, 25:901–920, 2004.

[172] H. Zou, T. Hastie, and R. Tibshirani. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, 2006.

Number	Species
1	Calanus I-IV
2	Para-Pseudocalanus spp.
3	Temora longicornis
4	Acartia spp. (unidentified)
5	Centropages typicus
6	Centropages hamatus
7	Isias clavipes
8	Oithona spp.
9	Corycaeus spp.
10	Calanus Total Traverse
11	Total Copepods
12	Podon spp.
13	Evadne spp.
14	Chaetognatha Traverse
15	Cyphonautes
16	Echinoderm larvae
17	Calanus finmarchicus
18	Calanus helgolandicus
19	Calanus hyperboreus
20	Rhincalanus nasutus
21	Metridia lucens
22	Candacia armata
23	Labidocera wollastoni
24	Tomopteris spp.
25	Gammaridea
26	Hyperiidea
27	Decapoda larvae (Total)
28	Clione limacina
29	Euphausiacea Total
30	Chaetognatha eyecount
31	Fish eggs
32	Fish larvae
33	Harpacticoida Total Traverse
34	Oncaea spp.
35	Parapontella brevicornis
36	Copepod nauplii
37	Cirripede larvae
38	Euphausiacea calyptopis
39	Anomalocera patersoni
40	Polychaete larvae (unidentified)
41	Cumacea
42	Isopoda
43	Mysidacea
44	Echinoderm post larvae
45	Branchiostoma lanceolatum
46	Salpidae
47	Appendicularia
48	Siphonostomatoida
49	Bivalvia larvae
50	Pseudocalanus spp. Adult Atlantic
51	Caprelloidea
52	Paraeuchaeta hebes
53	Paraeuchaeta norvegica

Table 8.1: Species numbers part 1: zooplankton.

Number	Species
55	Paralia sulcata
56	Skeletonema costatum
57	Thalassiosira spp.
58	Dactyliosolen antarcticus
59	Rhizosolenia imbricata shrubsolei
60	Rhizosolenia styliformis
61	Rhizosolenia hebetata semispina
62	Rhizosolenia alata indica
63	Chaetoceros(Hyalochaete) spp.
64	Chaetoceros(Phaeoceros) spp.
65	Odontella sinensis
66	Thalassiothrix longissima
67	Thalassionema nitzschioides
68	Ceratium fusus
69	Ceratium furca
70	Ceratium lineatum
71	Ceratium macroceros
72	Ceratium horridum
73	Ceratium longipes
74	Actinoptychus spp.
75	Bacillaria paxillifer
76	Bacteriastrum spp.
77	Bellerochea malleus
78	Biddulphia alternans
79	Phaeocystis pouchetii
80	Odontella granulata
81	Odontella regia
82	Odontella rhombus
83	Corethron criophilum
84	Coscinodiscus concinnus
85	Ditylum brightwellii
86	Eucampia zodiacus
87	Fragilaria spp.
88	Guinardia flaccida
89	Gyrosigma spp.
90	Leptocylindrus danicus
91	Navicula spp.
92	Cylindrotheca closterium
93	Rhizosolenia setigera
94	Stephanopyxis spp.
95	Ceratium bucephalum
96	Ceratium minutum
97	Dinophysis spp. Total
98	Prorocentrum spp. Total
99	Coscinodiscus wailesii
100	Proboscia alata
101	Leptocylindrus mediterraneus
102	Proboscia inermis
103	Asterionellopsis glacialis
104	Pseudo-nitzschia delicatissima complex
105	Pseudo-nitzschia seriata complex
106	Guinardia delicatula
107	Dactyliosolen fragilissimus
108	Guinardia striata
109	Lauderia annulata

Table 8.2: Species numbers part 2: phytoplankton.