NONLINEAR PRINCIPAL COMPONENT ANALYSIS OF CLIMATE DATA

by

Adam Hugh Monahan

B. Sc. (Honours Physics) University of Calgary, 1993

M. Sc. (Physics) University of British Columbia, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

EARTH AND OCEAN SCIENCES

We accept this thesis as conforming

to the required standard

....

....

....

....

....

THE UNIVERSITY OF BRITISH COLUMBIA

February 2000

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Earth and Ocean Sciences
The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1Z1

Date:

Abstract

A nonlinear generalisation of Principal Component Analysis (PCA), denoted Nonlinear Principal Component Analysis (NLPCA), is introduced and applied to the analysis of climate data. This method is implemented using a 5-layer feed-forward neural network introduced originally in the chemical engineering literature. The method is described and details of its implementation are addressed. It is found empirically that NLPCA partitions variance in the same fashion as does PCA, that is, that the sum of the total variance of the NLPCA approximation with the total variance of the residual from the original data is equal to the total variance of the original data. An important distinction is drawn between a modal P-dimensional NLPCA analysis, in which P successive 1D approximations are determined iteratively so that the approximation is the sum of P nonlinear functions of one variable, and a nonmodal analysis, in which the P-dimensional NLPCA approximation is determined as a nonlinear non-additive function of P variables.

Nonlinear Principal Component Analysis is first applied to a data set sampled from the Lorenz attractor. It is found that the NLPCA approximations are much more representative of the data than are the corresponding PCA approximations. In particular, the 1D and 2D NLPCA approximations explain 76% and 99.5% of the total variance, respectively, in contrast to 60% and 95% explained by the 1D and 2D PCA approximations.

When applied to a data set consisting of monthly-averaged tropical Pacific Ocean sea surface temperatures (SST), the modal 1D NLPCA approximation describes average variability associated with the El Niño/Southern Oscillation (ENSO) phenomenon, as does the 1D PCA approximation. The NLPCA approximation, however, characterises the asymmetry in spatial pattern of SST anomalies between average warm and cold events

(manifested in the skewness of the distribution) in a manner that the PCA approximation cannot. The second NLPCA mode of SST is found to characterise differences in ENSO variability between individual events, and in particular is consistent with the celebrated 1977 "regime shift". A 2D nonmodal NLPCA approximation is determined, the interpretation of which is complicated by the fact that a secondary feature extraction problem has to be carried out to interpret the results. It is found that this approximation contains much the same information as that provided by the modal analysis. A modal NLPC analysis of tropical Indo-Pacific sea level pressure (SLP) finds that the first mode describes average ENSO variability in this field, and also characterises an asymmetry in SLP fields between average warm and cold events. No robust nonlinear mode beyond the first could be found.

Nonlinear Principal Component Analysis is used to find the optimal nonlinear approximation to SLP data produced by a 1001 year integration of the Canadian Centre for Climate Modelling and Analysis (CCCma) coupled general circulation model (CGCM1). This approximation's associated time series is strongly bimodal and partitions the data into two distinct regimes. The first and more persistent regime describes a standing oscillation whose signature in the mid-troposphere is alternating amplification and attenuation of the climatological ridge over Northern Europe. The second and more episodic regime describes mid-tropospheric split-flow south of Greenland. Essentially the same structure is found in the 1D NLPCA approximation of the 500mb height field itself. In a 500 year integration with atmospheric CO_2 at four times pre-industrial concentrations, the occupation statistics of these preferred modes of variability change, such that the episodic split-flow regime occurs less frequently while the standing oscillation regime occurs more frequently.

Finally, a generalisation of Kramer's NLPCA using a 7-layer autoassociative neural network is introduced to address the inability of Kramer's original network to find P-dimensional structure topologically different from the unit cube in $\Re^P$. The example of an ellipse is considered, and it is shown that the approximation produced by the 7-layer network is a substantial improvement over that produced by the 5-layer network.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AAO Antarctic Oscillation

AO Arctic Oscillation

CCCma Canadian Centre for Climate Modelling and Analysis

EOF Empirical Orthogonal Function

ENSO El Niño/Southern Oscillation

FEV Fraction of Explained Variance

FUV Fraction of Unexplained Variance

GCM General Circulation Model

NAO North Atlantic Oscillation

NLPCA Nonlinear Principal Component Analysis

NMSD Normalised Mean Square Distance

PCA Principal Component Analysis

PCS Principal Curves and Surfaces

PDF Probability Density Function

PNA Pacific-North America

RPCA Rotated Principal Component Analysis

SLP(A) Sea Level Pressure (Anomaly)

SST(A) Sea Surface Temperature (Anomaly)

$\mathbf{Z_{500}}$(A) 500mb geopotential height (Anomaly)

CALVIN: I used to hate writing assignments, but now I enjoy them. I realized that the purpose of writing is to inflate weak ideas, obscure poor reasoning and inhibit clarity. With a little practice writing can be an intimidating and impenetrable fog! Want to see my book report?

HOBBES (reading): "The Dynamics of Interbeing and Monological Imperatives in 'Dick and Jane': A Study in Psychic Transrelational Modes."

CALVIN: Academia, here I come!

Bill Watterson

Acknowledgements

I would like to thank my supervisors Dr. William Hsieh and Dr. Lionel Pandolfo for their help and guidance during the course of this work. Thanks are also due to Dr. John Fyfe and Dr. Greg Flato, who collaborated on much of the work presented in Chapter 5. As well, I would like to acknowledge Dr. Benyang Tang, Dr. Susan Allen, Dr. Phil Austin, Dr. Cindy Greenwood, and Dr. Roland Stull for suggestions and advice.

None of this work could have been accomplished without the computer support provided by Denis Laplante. Merci.

My colleagues, past and present, improved this work with useful comments and suggestions. Thank you, Steven Bograd, Ana Carrasco, Pål Isachsen, Youmin Tang, Fredolin Tangang, and Yuval.

I am grateful to the National Science and Engineering Research Council, the Crisis Points Group of the Peter Wall Institute for Advanced Study, and the University of British Columbia for financially supporting me over the course of this work.

Tom Waits, John Zorn, and Frank Zappa have helped me maintain what has passed for my sanity over the last four years. Gentlemen, I am in your debt.

Most of all, I would like to express my love and gratitude to my friends and family. Thank you Drew, Michelle, James, Anna, Jenn, Jono, and Alison; thank you Mom, Dad, Erin, Sarah, and Darren. Thanks for putting up with me, and thanks for dragging me away from my computer every now and again.

Chapter 1

Introduction

Early work in the field of climate variability involved the consideration of climatic variables averaged over very large spatial and temporal scales. These variables could in consequence be represented in phase spaces of a small number of dimensions. However, the range of spatial and temporal scales under consideration has broadened substantially over the last few decades, so that data sets considered in contemporary research may involve phase spaces with hundreds to thousands of dimensions. While this refinement of scales allows consideration of a richer class of physical phenomena, it also confounds attempts at reaching an holistic understanding of the data under consideration. A typical modern climatic dataset is overwhelming in the amount of information it contains, so statistical techniques to distill massive multivariate datasets down to a phase space of smaller dimension, characterising the essential information contained in the data, are of great importance. In other words, it is necessary to develop methods to extract the signal from noise in climate data. These methods may be described as belonging to the general class of *feature extraction problems*, which attempt to characterise lower-dimensional structure in large multivariate datasets.

One of these feature extraction methods, Principal Component Analysis (PCA), also known as Empirical Orthogonal Function (EOF) analysis, has been widely used in oceanography and meteorology since its introduction to these fields by Lorenz (1956). PCA is an objective technique used to detect and characterise optimal lower-dimensional linear structure in a multivariate data set, and it is one of the most important methods in the geostatistician's multivariate statistics tool-box. Consequently, it has been well-studied, and standard references exist describing the method and its implementation (Preisendorfer, 1988; Wilks, 1995; von Storch and Zwiers, 1999). Its applications include reduction of data dimensionality for data interpretation (e.g. Barnston and Livezey, 1987; Miller et al., 1997) and for forecasting (e.g. Barnston and Ropelewski, 1992; Tangang et al., 1998). Furthermore, the connection between the results of PCA, which are statistical in nature, and the underlying dynamics of the system under consideration are understood in some detail (North, 1984; Mo and Ghil, 1987).

By construction, PCA finds a lower-dimensional hyperplane which optimally characterises the data, such that the sum of squares of orthogonal deviations of the data points from the hyperplane is minimised. If the structure of the data is inherently linear (for example, if the underlying distribution is Gaussian), then PCA is an optimal feature extraction algorithm; however, if the data contain *nonlinear* lower-dimensional structure, it will not be detectable by PCA. An example of a data set with nonlinear low-dimensional structure is the noisy parabola

$$\mathbf{X}(t_n) = (t_n, t_n^2)^T + \epsilon(t_n) \tag{1.1}$$

where $\epsilon(t_n)$ is a 2D iid $\mathcal{N}(\mathbf{0}, \Sigma^2)$ noise process. Underlying this 2D data set is a 1D parabolic curve. Because this curve cannot be described by a single straight line, a 1D PCA approximation would be unable to characterise this underlying 1D structure.

In the early 1990s, a neural-network based generalisation of PCA to the nonlinear feature extraction problem was introduced in the chemical engineering literature by Kramer (1991), who referred to the resulting technique as Nonlinear Principal Component Analysis (NLPCA). Another solution to this problem, coming from the statistics community, was put forward independently by Hastie and Stuetzle (1989), who named their method Principal Curves and Surfaces (PCS). Recently, Malthouse (1998) demonstrated that

NLPCA and PCS are closely related, and are, for a broad class of situations, essentially the same. Kramer's NLPCA has been applied to problems in chemical engineering (Kramer, 1991; Dong and McAvoy, 1996), psychology (Fotheringhame and Baddeley, 1997; Takane, 1998), dynamical systems theory (Kirby and Miranda, 1994; Kirby and Miranda, 1999), biomedical signal processing (Stamkopoulos et al., 1998), satellite remote sensing (Del Frate and Schiavon, 1999), and image compression (De Mers and Cottrell, 1993), but apart from a single unpublished report by Sengupta and Boyle (1995), the results of which were equivocal, it has not been applied to the large multivariate datasets common in oceanic and atmospheric sciences.

The object of this thesis is to investigate the application of NLPCA to climatic data sets. A brief review of PCA and a description of NLPCA are given in Chapter 2. As well, Chapter 2 contains a discussion of subtleties in the implementation of NLPCA. In particular, it addresses the problem of overfitting and presents the approach adopted for its avoidance. In Chapter 3, NLPCA is first applied to a synthetic data set sampled from the Lorenz attractor (Lorenz, 1963). Synthetic data are used to develop intuition about the implementation of NLPCA and its results. As well, the addition of noise to this synthetic data allows an investigation of the ability of this statistical tool to extract structure from noisy data. The application of NLPCA to actual climate data is investigated first in Chapter 4, in which the low-dimensional nonlinear structure of tropical Pacific Ocean sea surface temperatures (SST) and tropical Indo-Pacific sea level pressure (SLP) is investigated. Chapter 5 contains the results of an analysis of Northern Hemisphere extratropical SLP and 500mb geopotential height from the Canadian Centre for Climate Modelling and Analysis (CCCma) coupled general circulation model (GCM). Chapter 6 discusses a generalisation of Kramer's NLPCA to a 7-layer autoassociative neural network, which can solve a broader class of feature extraction problems than can the original 5-layer network. Conclusions are presented in Chapter 7. Appendix A describes feed-forward neural networks and Appendix B briefly introduces Principal Curves. Appendix C discusses the symmetric and antisymmetric components of composites, an idea introduced in Chapter 4.

Chapter 2

Nonlinear Principal Component Analysis: Theory and Implementation

2.1 Introduction

Both traditional Principal Component Analysis (PCA) and Nonlinear Principal Component Analysis (NLPCA) are examples of *feature extraction problems*, the goal of which is to extract from multivariate data sets representative structures of lower dimension. However, unlike PCA, NLPCA does not admit an analytic solution. Its implementation involves the iterative solution of a variational problem, which must be approached with care. This chapter presents PCA and NLPCA as feature extraction problems, and then addresses the methodology adopted for the implementation of NLPCA.

2.2 Feature Extraction Problems

Denote by $\mathbf{X}(t_n) \in \Re^M$ a typical meteorological or oceanographic data set, where $n \in (1, ..., N)$ labels observation times and the individual components of the vector $\mathbf{X}(t_n)$ correspond to individual observing stations. It is usually the case that the field values at different stations do not evolve independently in time, that is, that the temporal variability of $\mathbf{X}(t_n)$ includes contributions from large-scale, spatially-coherent features. In such a circumstance, the data will not be scattered evenly through the M-dimensional phase space of stations, but will tend to cluster around lower-dimensional surfaces; it is then appropriate to describe $\mathbf{X}(t_n)$ by the model:

$$\begin{aligned}\mathbf{X}(t_n) &= (\mathbf{f} \circ \mathbf{s_f})(\mathbf{X}(t_n)) + \epsilon_n \\ &= \hat{\mathbf{X}}(t_n) + \epsilon_n \end{aligned} \tag{2.1}$$

The function $\mathbf{s_f} : \Re^M \to \Re^P$, $1 \leq P < M$ parameterises a manifold of dimensionality lower than that of $\mathbf{X}(t_n)$, $\mathbf{f} : \Re^P \to \Re^M$ is a smooth map from this manifold to the original space, and the ϵ_n are residuals. The estimation of $\mathbf{s_f}$ and $\mathbf{f}$ from the data $\mathbf{X}(t_n)$, subject to an optimality criterion such as minimising the sum of squares of the residuals, is an example of a feature extraction problem: given a noisy data set $\mathbf{X}(t_n)$, it is desired to retrieve the signal $\hat{\mathbf{X}}(t_n) = (\mathbf{f} \circ \mathbf{s_f})(\mathbf{X}(t_n))$. Doing so, the M-dimensionality of $\mathbf{X}(t_n)$ is treated as being in a sense only superficial, as the signal of interest lives on a P-dimensional manifold embedded in $\Re^M$. Because once $\mathbf{f}$ and $\mathbf{s_f}$ have been found, it is no longer necessary no longer need to work with the signal in $\Re^M$ and can concentrate instead on the signal in $\Re^P$, feature extraction can also be thought of as reduction of data dimensionality.

The method of feature extraction most common in the atmospheric and oceanic sciences is Principal Component Analysis (PCA), which optimally extracts linear features from the data. However, if the underlying structure of the data is nonlinear, traditional PCA is suboptimal in characterising this lower-dimensional structure. This deficiency of PCA motivates the definition of Nonlinear Principal Component Analysis (NLPCA). In this section, I provide a brief review of PCA and demonstrate how it generalises naturally to a nonlinear method. I then discuss the NLPCA method in detail.

2.2.1 Principal Component Analysis

Traditional PCA can be formulated variationally as a special case of the feature-extraction problem, in which the data $\mathbf{X}(t_n)$ (assumed, without loss of generality, to have zero mean in time) is fit to the linear P-dimensional model:

$$\mathbf{X}(t_n) = \sum_{k=1}^{P} (\mathbf{X}(t_n) \cdot \mathbf{e}_k)\mathbf{e}_k + \epsilon_n \tag{2.2}$$

for vectors $\mathbf{e}_k \in \Re^M$, $\mathbf{e}_k \cdot \mathbf{e}_j = \delta_{kj}$ such that the sum of squares of the residuals:

$$J = < ||\mathbf{X} - \hat{\mathbf{X}}||^2 > \tag{2.3}$$

is a minimum, where angle brackets denote a sample time average. The vector $\mathbf{e}_k$ is the kth Empirical Orthogonal Function (EOF) and the projection of $\mathbf{X}(t_n)$ on $\mathbf{e}_k$ is the kth Principal Component (PC). The product of the kth EOF with the kth PC defines a vector time series usually referred to as the kth PC *mode.* The P-dimensional PCA approximation $\hat{\mathbf{X}}(t_n)$ to $\mathbf{X}(t_n)$ lives on the P-dimensional hyperplane that passes optimally through the middle of the data (von Storch and Zwiers, 1999). Principal Component Analysis has the variance partitioning property:

$$\sum_{i=1}^{M} \mathrm{var}(X_i) = \sum_{i=1}^{M} \mathrm{var}(\hat{X}_i) + \sum_{i=1}^{M} \mathrm{var}(X_i - \hat{X}_i), \tag{2.4}$$

so it is sensible to say that $\hat{\mathbf{X}}(t_n)$ "explains" a certain fraction of the variance of $\mathbf{X}(t_n)$. In particular, $\hat{\mathbf{X}}(t_n) = (\mathbf{X}(t_n) \cdot \mathbf{e}_1)\mathbf{e}_1$ is the one-dimensional linear approximation to $\mathbf{X}(t_n)$ which explains the highest percentage of the variance. The fraction of variance explained by $\hat{\mathbf{X}}(t_n)$ is a non-decreasing function of the approximation dimension P; increasing the dimensionality of the PCA approximation increases its fidelity to the original data.

Principal Component Analysis is usually thought of in terms of the eigenstructure of the data covariance matrix $C = < \mathbf{X}\mathbf{X}^T >$. In fact, the vectors $\mathbf{e}_k$ are eigenvectors of C corresponding to the P largest eigenvalues:

$$C\mathbf{e}_k = \lambda_k \mathbf{e}_k \tag{2.5}$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_P$. This fact follows from the minimisation of (2.3) subject to the constraint that the vectors $\mathbf{e}_k$ are normalised. While this eigenanalysis approach is the standard approach to calculating the $\mathbf{e}_k$, it has no analogue in the nonlinear generalisation to be considered. The variational formulation of PCA (from which follows its relation to the eigenstructure of the covariance matrix) is emphasised because it generalises naturally to the nonlinear feature extraction problem.

Note that the PCA approximation $\hat{\mathbf{X}}(t_n)$ to $\mathbf{X}(t_n)$ is the composition of two functions:

1. a projection function $\mathbf{s_f} : \Re^M \rightarrow \Re^P$:

$$\begin{aligned} \mathbf{s_f}(\mathbf{X}(t_n)) &= (\mathbf{X}(t_n) \cdot \mathbf{e}_1, ..., \mathbf{X}(t_n) \cdot \mathbf{e}_P)^T \\ &= \Pi \mathbf{X}(t_n) \end{aligned} \tag{2.6}$$

   where Π is the $P \times M$ matrix whose kth row is the vector $\mathbf{e}_k$, and

2. an expansion function $\mathbf{f} : \Re^P \rightarrow \Re^M$:

$$\mathbf{f}(\mathbf{s_f}) = \Pi^T \mathbf{s_f} \tag{2.7}$$

Thus, the PCA approximation $\hat{\mathbf{X}}(t_n)$ to $\mathbf{X}(t_n)$ is given by

$$\begin{aligned} \hat{\mathbf{X}}(t_n) &= (\mathbf{f} \circ \mathbf{s_f})(\mathbf{X}(t_n)) \\ &= \Pi^T(\Pi \mathbf{X}(t_n)) \\ &= (\Pi^T \Pi)\mathbf{X}(t_n) \end{aligned} \tag{2.8}$$

In the language of LeBlanc and Tibshirani (1994), the projection function characterises the *dimension reduction* aspect of PCA, and the expansion function characterises its *function approximation* aspect. In traditional PCA, both the projection and expansion functions are linear. This method is thus optimal if the feature to be extracted is well-characterised by a set of orthogonal, straight axes: that is, if the data cloud is cigar-shaped (e.g. Gaussian). But what if the data cloud is ring-like, or bowed? In such cases, there is a clear lower-dimensional structure to the data, but not one which is linear, and thus can not be extracted by traditional PCA. This motivates the definition of a generalised, nonlinear PCA.

The generalisation of PCA to NLPCA can also be motivated by the following observation. The 1D PCA approximation $\hat{\mathbf{X}}(t_n)$ to $\mathbf{X}(t_n)$ is separable in terms of its spatial and temporal structure. That is, $\hat{\mathbf{X}}(t_n)$ is the product of a function of time, $\mathbf{X}(t_n) \cdot \mathbf{e}_1$, with a function of space, $\mathbf{e_1}$:

$$\hat{\mathbf{X}}(t_n) = (\mathbf{X}(t_n) \cdot \mathbf{e}_1)\, \mathbf{e}_1. \tag{2.9}$$

In consequence, this approximation can only describe standing variability in the data set, that is, a fixed spatial pattern with an amplitude that varies as a function of time. There is no *a priori* reason to believe that the optimal 1D approximation to a climatic data set is a standing oscillation, but that is all PCA can produce. As shall be seen in Chapters 4 and 5, by moving to NLPCA, 1D approximations to climatic data sets which are not standing oscillations may be obtained.

2.2.2 Nonlinear Principal Component Analysis

To circumvent the limitations of linearity inherent in the PCA model (2.2), Kramer (1991) introduced a nonlinear generalisation that solved the general feature extraction problem described by the model (2.1), where $\mathbf{f}$ and $\mathbf{s_f}$ are allowed to be *nonlinear* functions. Given data $\mathbf{X}(t_n) \in \Re^M$, the problem is to estimate functions $\mathbf{s}_f : \Re^M \to \Re^P$ and $\mathbf{f} : \Re^P \to \Re^M$, where $P < M$, such that the approximation

$$\hat{\mathbf{X}}(t_n) = (\mathbf{f} \circ \mathbf{s}_f)(\mathbf{X}(t_n)) \tag{2.10}$$

to $\mathbf{X}(t_n)$ passes through the middle of the data, ie, such that the sum of squared residuals,

$$J =< ||\mathbf{X}(t_n) - \hat{\mathbf{X}}(t_n)||^2 > \tag{2.11}$$

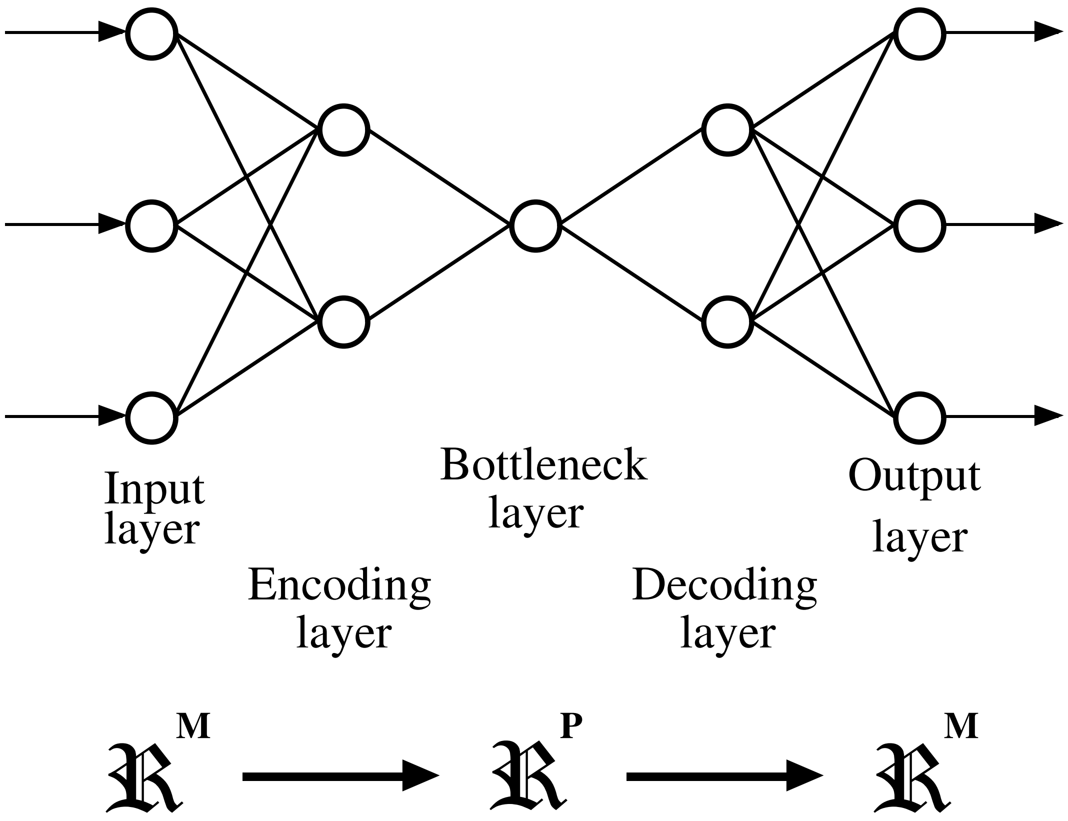


Figure 2.1: The 5-layer feed-forward autoassociative neural network used to perform NLPCA.

is a minimum. Called *Nonlinear Principal Component Analysis* (NLPCA), Kramer implemented his solution using a 5-layer feed-forward neural network. Neural networks are nonlinear, nonparametric statistical tools for function estimation. They are described in detail in Appendix A.

Figure 2.1 shows the architecture of the 5-layer network used to extract the 1D NLPCA approximation to the data set $\mathbf{X}(t_n) \in \Re^M$; this network is unusual in that the third layer contains only a single neuron. This third layer as is referred to as the *bottleneck layer*. The first (*input*) and fifth (*output*) layers each contain M neurons. Layers

2 and 4 are called respectively the *encoding* and *decoding* layers; they contain L neurons, the transfer functions of which are hyperbolic tangents. The transfer functions of the bottleneck and output layers are linear. As input, the network is presented the vector $\mathbf{X}(t_n)$ for each time t_n; the corresponding network output is denoted $\mathcal{N}(\mathbf{X}(t_n))$. The weights and biases are adjusted ("trained"), using a conjugate gradient algorithm (Press et al., 1992) until the sum of squared differences between input and output:

$$J = < ||\mathbf{X}(t_n) - \mathcal{N}(\mathbf{X}(t_n))||^2 > \tag{2.12}$$

is minimised (subject to certain robustness criteria discussed in the next section). Because the network is trained to approximate as closely as possible the input data itself, it is said to be *autoassociative*. It was proved by Sanger (1989) that if the transfer functions of the neurons in the second and fourth layers are linear, the resulting network performs classical PCA, such that the output of the bottleneck layer is the time series $s_{\mathbf{f}}(t_n)$ of equation (2.6) (up to a normalisation factor).

Now consider the manner by which this network solves the feature extraction problem for $P = 1$. The first three layers, considered alone, form a map from $\Re^M$ to $\Re$, and the last three layers alone form a map from $\Re$ to $\Re^M$. All five layers together are simply the composition of these two maps. Because the bottleneck layer contains only a single neuron, the network must compress the input down to a single one-dimensional time series before it produces its final M-dimensional output. Once the network has been trained, the output $\mathcal{N}(\mathbf{X}(t_n))$ is the optimal one-dimensional approximation to $\mathbf{X}(t_n)$, embedded in $\Re^M$. As is discussed in Appendix A, it is known from a result due to Cybenko (1989) that if L is sufficiently large, then the first three layers can approximate any continuous $\mathbf{s}_f$, and the last 3 layers any continuous $\mathbf{f}$, to arbitrary accuracy. Thus, the network illustrated in Figure 2.1 should be able to recover optimally, in a least-squares sense, any one-dimensional nonlinear structure present in $\mathbf{X}(t_n)$ for which the projection and expansion functions are continuous. It is not required that the encoding and decoding layers each have the same number of neurons, but the numbers are fixed to be the same so as to have only one free parameter in the model architecture. That the network must be composed of (at least) 5 layers follows from the fact that each of the functions $\mathbf{s_f}$ and $\mathbf{f}$ requires a network with (at least) 3 layers for its approximation. The composition $\mathbf{f} \circ \mathbf{s_f}$ of the two must then have at least 5 layers, as one layer is shared.

The network illustrated in Figure 2.1 will extract the optimal one-dimensional curve characterising $\mathbf{X}(t_n)$. To uncover higher-dimensional structure, the number of neurons in the bottleneck layer can be increased. For example, if two neurons are used, the network will determine the optimal two-dimensional characterisation (by continuous functions) of $\mathbf{X}(t_n)$. In general, a P-dimensional NLPCA approximation to $\mathbf{X}(t_n)$ can be obtained by setting to P the number of neurons in the bottleneck layer.

Another solution to the general feature extraction problem was introduced independently by Hastie and Stuetzle (1989). Their method, termed *Principal Curves and Surfaces* (PCS), is described in Appendix B. Principal Curves and Surfaces is based on a rather different set of ideas than NLPCA. In practice, however, because both minimise the sum of squared errors (2.11), the two methods both boil down to iterative solutions to the variational formulation of a feature extraction problem. In fact, Malthouse (1998) argued that NLPCA and PCS are quite similar for a broad class of feature extraction problems. A primary difference between NLPCA and PCS is that in the former the projection function $\mathbf{s}_f$ is constrained to be continuous, while in the latter it may have a finite number of discontinuities. Although here I will investigate the use of Kramer's NLPCA, because its implementation is straightforward, PCS has a stronger theoretical underpinning. In Chapter 6 a generalisation of Kramer's NLPCA that can model discontinuous projection and expansion functions is introduced, and is thus closer to PCS.

As noted by LeBlanc and Tibshirani (1994), Hastie and Stuetzle's PCS partitions variance in the same fashion as does traditional PCA: if $\hat{\mathbf{X}}(t_n)$ is the PCS approximation to $\mathbf{X}(t_n)$, then

$$\sum_{i=1}^{M} \mathrm{var}(X_i) = \sum_{i=1}^{M} \mathrm{var}(\hat{X}_i) + \sum_{i=1}^{M} \mathrm{var}(X_i - \hat{X}_i). \tag{2.13}$$

As with PCA, it is therefore sensible to describe a PCS approximation as explaining a certain fraction of variance in the original dataset. From the close relationship between NLPCA and PCS demonstrated by Malthouse (1998), it is tempting to hypothesise that NLPCA also partitions variance in such a fashion. While I am not aware of a rigorous proof of this result, this partitioning of variance in fact occurs in all of the examples I have considered, and in the following discussion it shall be assumed that equation (2.13) holds for $\hat{\mathbf{X}}(t_n)$ the NLPCA approximation to $\mathbf{X}(t_n)$.

Yet a third nonlinear generalisation of PCA was introduced by Oja and Karhunen (1993) and by Oja (1997), in which the map $\mathbf{s_f}$ is allowed to be nonlinear while $\mathbf{f}$ remains linear. Such a generalisation can be carried out using a two-layer recursive neural network. Because only the projection function is nonlinear, this approach is distinct from the class of feature extraction problems addressed by Kramer's NLPCA and Hastie and Stuezle's PCS.

Because the traditional PCA model has the additive structure (2.2), the optimal linear P-dimensional substructure of $\mathbf{X}(t_n)$ can be found all at once, or mode by mode; both methods yield the same result. In the iterative approach, the first mode $\hat{\mathbf{X}}^{(1)}(t_n)$ of $\mathbf{X}(t_n)$ is calculated from the entire data set, and then the second mode is calculated from the residual $\mathbf{X}(t_n) - \hat{\mathbf{X}}^{(1)}(t_n)$, taking advantage of the fact that the second PC mode of $\mathbf{X}(t_n)$ is the first PC mode of this residual. The two approaches are equivalent for PCA because the most general linear function of P variables has the additive structure:

$$g(u_1, u_2, ..., u_P) = \sum_{i=1}^{P} \alpha_i u_i. \tag{2.14}$$

They are generally distinct, however, for NLPCA, as an arbitrary smooth function f of

P variables cannot be decomposed as a sum of smooth functions of one variable. That is, in general

$$f(u_1, u_2, ..., u_P) \neq f_1(u_1) + f_2(u_2) + ... + f_P(u_P) \quad (2.15)$$

for some functions $f_1, f_2, ..., f_P$: f cannot usually be written as a Generalised Additive Model (Hastie and Tibshirani, 1990). The iterative approach will be referred to as a *modal* analysis, and to the all-at-once approach as *nonmodal*. Naturally enough, in a modal analysis, each 1D approximation will be referred to as a *mode*, and ordered in terms of decreasing fraction of variance explained. I will compare both the modal and nonmodal approaches in this thesis. Theoretically, the nonmodal P-dimensional approximation should be superior to the modal approximation, because it is drawn from a broader class of functions, although the modal analysis is more amenable to interpretation. Of course, a general P-dimensional analysis could involve both modal and nonmodal decompositions at various stages; such mixed modal/nonmodal analyses will not be considered here.

Malthouse (1998) pointed out two limitations of NLPCA as formulated by Kramer (1991). First, Kramer's NLPCA is unable to characterise low-dimensional structure which is self-intersecting. Because the projection $\mathbf{s_f}$ must be discontinuous for a self-intersecting surface, there will be open neighbourhoods in $\Re^M$ that are mapped by $\mathbf{s_f}$ into non-open neighbourhoods in $\Re^P$. Consider the example of a circle in $\Re^2$. It is a 1D surface, a natural parameterisation of which is the interval $\theta \in [0, 2\pi]$, with the points $\theta = 0$ and $\theta = 2\pi$ identified (S^1 topology). Clearly, for any small ϵ, there will exist an open neighbourhood on the circle which maps onto the non-open set $[0, \epsilon) \cup (2\pi - \epsilon, 2\pi]$. This limitation of Kramer's NLPCA is not of great importance to the analysis of climate data, as precisely cyclic variability is not characteristic of climatic systems. An exception is perhaps the annual cycle, but it is typically removed from climate data before analysis. The limitation can be removed by moving to a 7-layer neural network, which can approximate *discontinuous* projection and expansion functions. This issue is addressed in Chapter 6.

The second limitation of Kramer's NLPCA highlighted by Malthouse is that the parameterisation $\mathbf{s_f}$ of the NLPCA approximation is only determined up to an arbitrary homeomorphism (i.e., a continuous, one-to-one, and onto function with a continuous inverse). That is, for an arbitrary homeomorphism $\mathbf{g} : \Re^P \mapsto \Re^P$, the time series $\mathbf{g}(\mathbf{s_f}(\mathbf{X}(t_n)))$ is also an acceptable parameterisation of the surface, because $\mathbf{f} \circ \mathbf{s_f} = (\mathbf{f} \circ \mathbf{g}^{-1}) \circ (\mathbf{g} \circ \mathbf{s_f})$. This degeneracy is a potentially serious complication in the interpretation of the time series produced by the bottleneck layer of Kramer's network.

Based on the results presented in this thesis, it is apparent that this degeneracy is not problematic in a modal NLPC analysis. Homeomorphisms from $\Re$ to $\Re$ are functions which can only stretch and compress (locally or globally), or translate globally, and thus do not radically change the information present in the time series $\mathbf{s_f}(\mathbf{X}(t_n))$. The time series arising from the modal analyses in Chapters 4 and 5 are amenable to natural interpretation in terms of familiar phenomena in the systems under consideration.

This degeneracy is substantially more significant in the case of a nonmodal analysis, because homeomorphisms from $\Re^P$ to $\Re^P$ for $P > 1$ can include rotations as well as dilations and compressions. Generally, the time series produced by the P different neurons in the bottleneck layer will thus not be independent, or even uncorrelated, because of mixing induced by this arbitrary rotation. Any P-dimensional surface can be parameterised by a set of independent variables γ_i, $i = 1, ..., P$. Determining such a set from the set of P time series $\beta_i(t_n)$ determined empirically by NLPCA is another problem of feature extraction in the space of the variables parameterising the surface. In principle, PCA or modal NLPCA can be used to calculate the $\gamma_i(t_n)$. An example of such an approach is considered in Chapter 5. The fact that nonmodal NLPCA leads to a second feature extraction problem reduces its utility, despite its aesthetic appeal.

Another common generalisation of PCA is Rotated Principal Component Analysis (RPCA; Richman, 1986). As was pointed out by Buell (1975,1979), because PCA is designed to maximise the global variance explained by the leading mode, and because successive modes are constrained to have orthogonal spatial patterns, its results can be strongly affected by the shape of the data domain. Rotated PCA addresses this problem by modifying the cost function (2.3) to include a "simple structure criterion" so that the resulting approximation strikes a compromise between maximising its explained variance and minimising its spatial scale. In doing so, either the orthogonality of the spatial patterns, the uncorrelatedness of the time series, or both, must be sacrificed.

Rotated PCA and NLPCA are both generalisations of PCA, but they address substantially different issues. Rotated PCA allows the detection of localised variability in the data, but still variability that is linear. Consequently, 1D RPCA approximations share with 1D PCA approximations the problem that they can only describe standing variability. On the other hand, NLPCA is concerned with detecting and characterising nonlinear structure in data sets. The general 1D NLPCA approximation cannot be expressed as a separable function of space and time such as (2.9), so it is able to describe variability more general than standing oscillations. Because of the lack of a "simple structure criterion" in the cost function (2.11), NLPCA also maximises global variance, and its results will presumably suffer to some degree from the same sensitivity to domain boundaries that PCA does. However, this problem is presumably less serious with NLPCA because of the absence of orthogonality constraints between different modes. In fact, such constraints cannot be naturally formulated for the nonlinear approximations. A further generalisation of NLPCA to encourage regionalisation of the approximation could be introduced by modifying the cost function (2.11) to include a simple structure criterion. Such a generalisation, while an interesting direction for future research, is beyond the scope of the present study.

2.3 Implementation of NLPCA

Neural networks are powerful tools for function approximation. Given input and target data sets $\mathbf{u}(t_n)$ and $\mathbf{v}(t_n)$, $n = 1, ..., N$, a neural network, denoted by $\mathcal{N}$, can be trained until $\mathcal{N}(\mathbf{u}(t_n))$ is an arbitrarily good approximation to $\mathbf{v}(t_n)$, if the number of neurons in the hidden layer is sufficiently large. That is, a network can always be built so that the total sum of squared errors

$$< ||\mathbf{v}(t_n) - \mathcal{N}(\mathbf{u}(t_n))||^2 > \tag{2.16}$$

is as small as desired. Another important property of the neural network is that it *generalises*, that is, that given new data $\mathbf{u}(t_{N+1}), \mathbf{v}(t_{N+1})$, the network error on these data is about the same size as the errors over the training set:

$$||\mathbf{v}(t_{N+1}) - \mathcal{N}(\mathbf{u}(t_{N+1}))||^2 \sim < ||\mathbf{v}(t_n) - \mathcal{N}(\mathbf{u}(t_n))||^2 > . \tag{2.17}$$

The two goals of minimising network error and maximising its ability to generalise are often incompatible, and a subtle balance must be struck between the two. This situation arises, for example, in the case when $\mathbf{u}(t_n)$ and $\mathbf{v}(t_n)$ are of the form

$$\mathbf{u}(t_n) = \mathbf{z}(t_n) + \epsilon_n \tag{2.18}$$

$$\mathbf{v}(t_n) = \mathbf{f}(\mathbf{z}(t_n)) + \eta_n \tag{2.19}$$

where ϵ_n and η_n are noise terms, and it is desired that $\mathcal{N}$ learn the deterministic relationship $\mathbf{f}$ between $\mathbf{u}(t_n)$ and $\mathbf{v}(t_n)$. In such a case, care must be taken to avoid allowing the network to fit the noise as well. If $\mathcal{N}$ is trained until it maps particular details of a given realisation of $\mathbf{u}(t_n)$ into those of a given realisation of $\mathbf{v}(t_n)$, and thus will not generalise, the network has *overfit*. An overfit network is not truly representative of the structure underlying a data set. The avoidance of overfitting by neural networks is a primary issue in their implementation (Finnoff et al., 1993; Yuval, 1999).

To avoid overfitting, a simple technique called *early stopping* has been used. Because the neural network is nonlinear in the model parameters, they must be determined iteratively in a process referred to as *training*. In early stopping, the training is terminated before the error function is minimised, according to a well-defined stopping criterion. In essence, the idea behind early stopping is that the training is allowed to continue sufficiently long to fit the structure underlying the data, but not long enough to fit the noise. The strategy employed was to hold aside a fraction of the data points, selected randomly, in a validation set not used to train the network. While network training proceeded, the network performance on the validation set was monitored, and training was stopped when this performance began to degrade, or after a fixed large number of iterations, whichever came first.

The use of early stopping along with the deterministic conjugate gradient algorithm to minimise the error function confers on the training results a degree of sensitivity to the network parameters (the weights and biases) used to initialise the iterative training procedure. This sensitivity is exacerbated by the possible existence of multiple minima in the error function (2.11). To address this problem, an ensemble of training runs starting from different, randomly chosen, initial parameter values was carried out for each analysis performed. The training results from these runs were examined, and those members of the ensemble for which the final error over the validation set was greater than that over the training set were discarded. The remaining members of the ensemble are referred to as *candidate models*.

The number of neurons L in the encoding and decoding layers determines the class of functions that $\mathbf{s_f}$ and $\mathbf{f}$ can take. As L increases, there is more flexibility in the forms of $\mathbf{s_f}$ and $\mathbf{f}$, but the model also has more parameters, implying both that the error surface becomes more complicated and that the parameters are less constrained by data. Consequently, for L large, the scatter among the candidate models can be large, as measured by the *normalised mean square distance* (NMSD). The NMSD between approximations $\hat{\mathbf{X}}^{\{1\}}(t_n)$ and $\hat{\mathbf{X}}^{\{2\}}(t_n)$ is defined as

$$\text{NMSD} = \frac{< ||\hat{\mathbf{X}}^{\{1\}} - \hat{\mathbf{X}}^{\{2\}}||^2 >}{< ||\mathbf{X}||^2 >} \tag{2.20}$$

This statistic was introduced in Monahan (1999), in which it was found that NLPCA approximations for which the NMSD was less than about 2% were essentially indistinguishable. In the end, the number L of neurons used in the encoding and decoding layers was the maximum such that the NMSD between NLPCA approximations to $\mathbf{X}(t_n)$ in the candidate model set was less than 5%. This threshold value of NMSD was chosen on the basis of experience and intuition; there is no existing rigorous sampling theory for this test statistic. In other words, for any given analysis, the value of L used in the NLPCA network is the largest that produces a robust set of candidate models. The early stopping technique ensures that the NLPCA approximation is robust to the introduction of new data, and the existence of a set of similar candidate models (as measured by NMSD) ensures that the approximation is robust with respect to the initial parameter values used in the training.

Finally, once a maximal L was determined and a set of candidate models obtained, the model selected as "the" NLPCA approximation was the one with the highest *Fraction of Explained Variance* (FEV):

$$\text{FEV} = \frac{< ||\hat{\mathbf{X}}||^2 >}{< ||\mathbf{X}||^2 >} \tag{2.21}$$

which is a meaningful statistic because NLPCA partitions variance as described in equation (2.13). Alternately, the candidate model selected was that which minimised the

Fraction of Unexplained Variance (FUV):

$$\mathrm{FUV} = 1 - \mathrm{FEV} \tag{2.22}$$

Typically, the FEV differed little between candidate models.

2.4 Dynamical Significance of Low-Dimensional Approximations

The lower-dimensional approximations obtained by feature extraction methods are statistical in nature. A natural question concerns the relation they bear to the dynamics of the system under investigation.

North (1984) considered the dynamical system

$$L\psi(\mathbf{x}, t) = \zeta(\mathbf{x}, t) \tag{2.23}$$

where L is a space-time linear differential operator and $\zeta(\mathbf{x}, t)$ represents stochastic forcing, and concluded that the EOFs of $\psi(\mathbf{x}, t)$ coincide with its dynamical modes if and only if the operator L is normal (i.e. it commutes with its adjoint) and the noise ζ is white in space and stationary in time:

$$E(\zeta(\mathbf{x}_1, t_1)\zeta(\mathbf{x}_2, t_2)) = g(|t_1 - t_2|)\delta(\mathbf{x}_1 - \mathbf{x}_2) \tag{2.24}$$

where $g(\tau)$ is a lag autocovariance function. These requirements greatly restrict the class of dynamical systems for which the connection between the statistics and dynamics is clear-cut, as it does not even include the geophysically-relevant class of linear models for which non-modal variance growth is important (Penland and Sardeshmukh, 1995; Farrell and Ioannou, 1996; Whitaker and Sardeshmukh, 1998).

Mo and Ghil (1987) attempted to assess the connection between the results of EOF analysis and the dynamics of the system under consideration in the context of nonlinear dynamics. They concluded that, "the dynamical interpretation of EOFs is their pointing from the time mean to the most populated regions of the system's phase space". If the distribution of the system in phase space is Gaussian, then this direction will lie along the distribution's principal axis. On the other hand, if the distribution is not Gaussian, but is characterised by an inhomogeneous density with a small number of local extrema corresponding to preferred regimes of behaviour (associated, e.g., with slow manifolds of the dynamics (Ghil and Childress, 1987)), then the leading EOFs will characterise the distribution of these extrema.

Unlike PCA, NLPCA approximations are not characterised by unique "directions" through phase space, but rather by curved surfaces. Interpretation of the results of an NLPC analysis must then be couched in rather different terms than that of PCA. A natural interpretation, using the language of dynamical systems theory is that NLPCA approximations characterise the *attractor* of the system under consideration, as was noted by Kirby and Miranda (1994). Many naturally occurring systems possess a stable attractor, which is a manifold typically of smaller dimension than the Cartesian phase space in which it must be embedded to preclude spurious self-intersections (Ott, 1993). These attractors are generally complicated surfaces of non-integer dimension; only in very special cases are they planar. Because PCA produces an orthogonal coordinate system in the phase space, it can at best eliminate the degrees of freedom in the data associated with noise, thereby producing an embedding space for the attractor. Nonlinear PCA, however, can characterise the curved structures associated with these attractors (although it is restricted to approximations of integer dimension), and produce what Kirby and Miranda denote the "optimal coordinates" of the system. The results of NLPCA are thus best considered as characterising the underlying attractor of the system under consideration, as will be illustrated in the following Chapter when an NLPCA approximation of the Lorenz attractor is constructed. The estimation of the attractor underlying a data set provides insight into the governing physics. Knowledge of the dominant forms of variability in the data can act as a guide to reductionism, helping to develop mechanistic models of the system under investigation, which then provide insight into the underlying dynamics.

Chapter 3

Nonlinear Principal Component Analysis of the Lorenz Attractor

3.1 Introduction

As a preliminary investigation into the implementation of NLPCA, consider a synthetic data set consisting of a set of points sampled from the Lorenz attractor (Lorenz, 1963). This familiar object is the attractor on which (as $t \to \infty$) live solutions $\mathbf{x}(t)$ of the system of coupled nonlinear ODEs

$$\dot{x_1} = -\sigma x_1 + \sigma x_2 \tag{3.1}$$

$$\dot{x_2} = -x_1 x_3 + r x_1 - x_2 \tag{3.2}$$

$$\dot{x_3} = x_1 x_2 - b x_3, \tag{3.3}$$

with parameter values $r = 28, b = 8/3$, and $\sigma = 10$. Synthetic data is used to test the NLPCA method because

- by adding random noise to the signal $\mathbf{x}(t)$, the sensitivity of the method to noise level can be tested, and
- the structure of the Lorenz attractor is well-known, and of sufficiently low dimension that visualisation of results is straightforward.

Figure 3.1 displays the projections of the Lorenz attractor (as determined by numerical integration of equations (3.1)-(3.3)) on the (x_1, x_2), (x_2, x_3), and (x_3, x_1) planes. It turns out that the Lorenz attractor is fractal, with a box-counting dimension of about 2.04

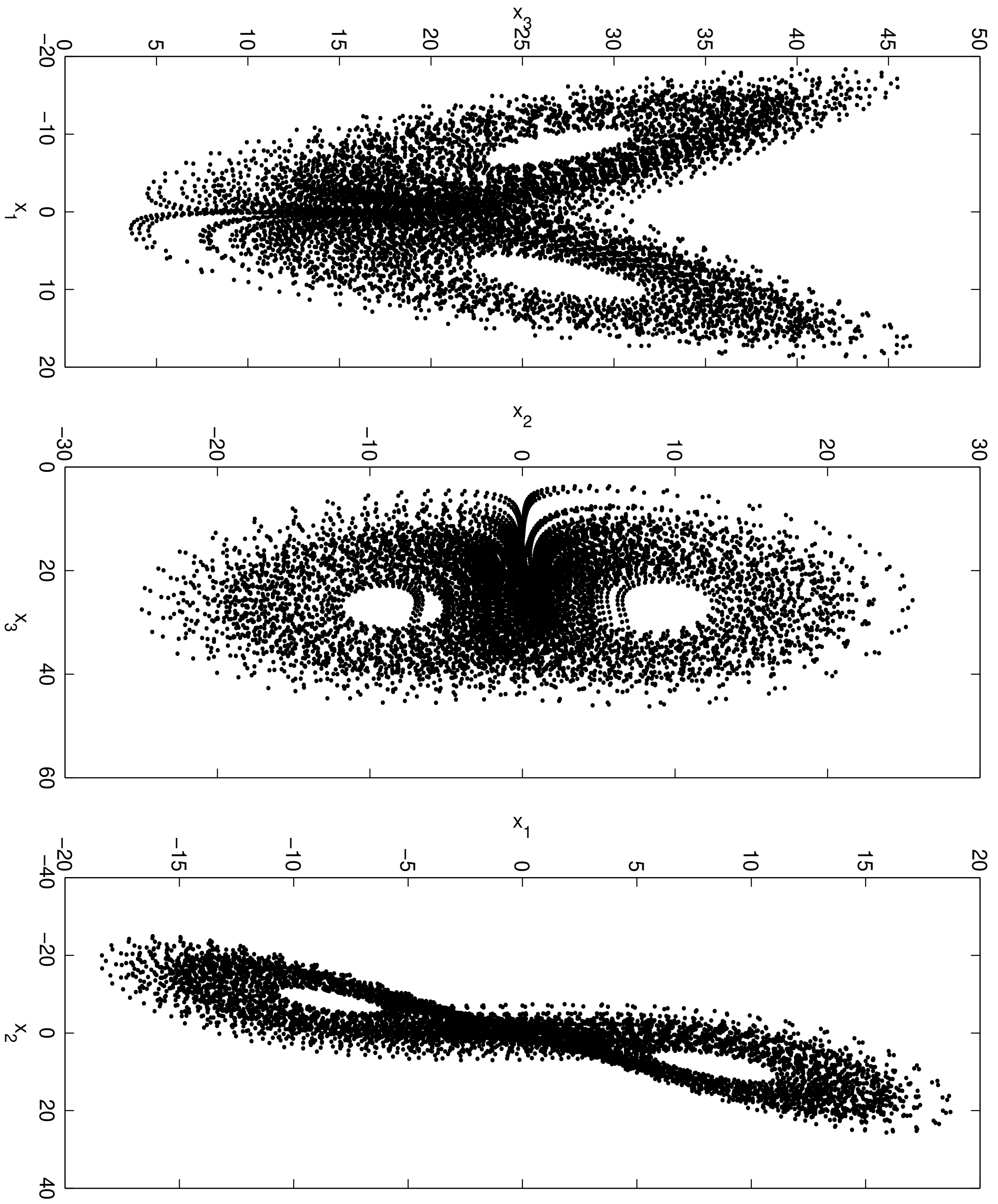


Figure 3.1: The Lorenz attractor, projected on the (x_1, x_3), (x_3, x_2), and (x_2, x_1) planes.

(Berliner, 1992). However, inspection of the butterfly-shaped attractor indicates that a one-dimensional U-shaped curve passing through the centres of the two lobes should explain a substantial fraction of the variance.

To produce a dataset of size similar to that typically encountered in climate applications (e.g. 600 points in length, corresponding to 50 years of monthly data), the data displayed in Figure 3.1 were subsampled at uniform intervals in time to produce a 3D time series 584 points in length, to be denoted $\mathbf{z}(t_n)$. The subsampled data set is displayed in Figure 3.2. Clearly, it retains the gross structure of the original attractor. To investigate the effects of noise on the NLPCA results, constructed the datasets

$$\mathbf{x}(t_n) = \mathbf{z}(t_n) + \eta\epsilon(t_n) \tag{3.4}$$

were constructed, where $\epsilon(t_n)$ is a 584-point 3D series of Gaussian iid random deviates with zero mean and unit standard deviation, and η is a tunable parameter for the noise level. This noise is added in an effort to model measurement error; the stochasticity is not intrinsic to the dynamics.

3.2 Model Building

The early stopping algorithm described in the previous chapter was used to carry out the NLPC analysis of the Lorenz data. A validation set containing 30% of the data points was set aside, and network performance over this set was monitored as training progressed. Training was stopped when this validation set error started to increase, or after 500 iterations, whichever was the first to occur.

3.3 Results

The 1D PCA approximation to $\mathbf{x}(t_n)$ when $\eta = 0$ is displayed in Figure 3.3; it is a straight line passing through the centres of the two lobes of the attractor, and explains

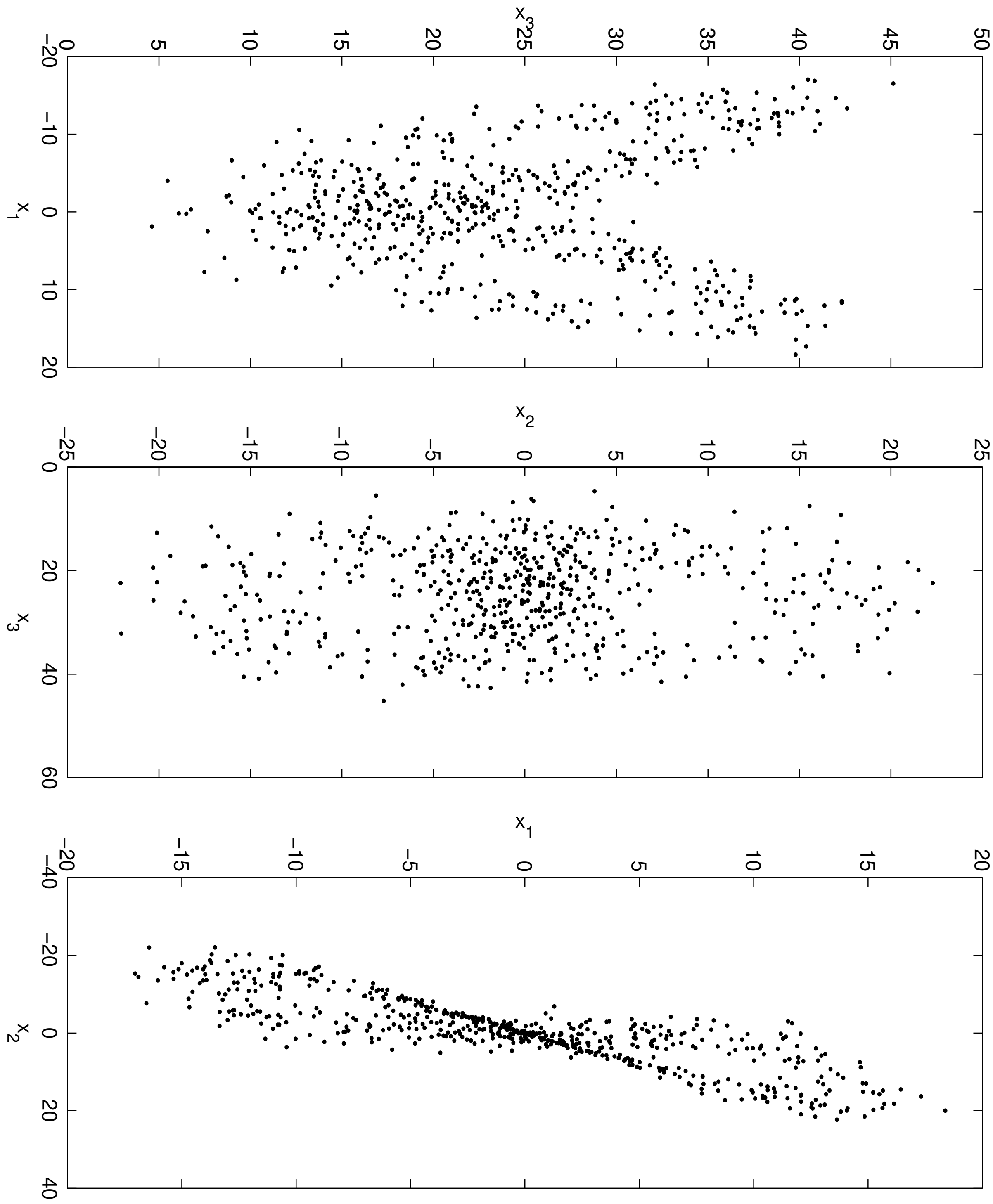


Figure 3.2: As in Figure 3.1, for a subsample of 584 points.

60% of the variance of $\mathbf{x}(t_n)$. Figure 3.4 displays the 1D NLPCA approximation to $\mathbf{x}(t_n)$. As anticipated, it is a U-shaped curve passing through the middle of the data. This curve explains 76% of the variance, an improvement of 16% over the PCA results. Clearly, the 1D NLPCA approximation is substantially closer to $\mathbf{x}(t_n)$ than is the 1D PCA approximation. The network used to perform the NLPCA had 3 input and output neurons for x_1, x_2, and x_3, 1 bottleneck neuron, and $L = 3$ neurons in the encoding and decoding layers. Experimentation indicated that the NLPCA results improved using $L = 3$ over using $L = 2$ (ie, the former had a smaller FUV than the latter), but that for $L > 3$, the results did not improve. Turning now to the issue of robustness of results, the NMSD between 6 different 1D NLPCA curves (not shown) varies between 0.5% and 2%. These curves differ only in small details, and agree in their essential structure with the curve shown in Figure 3.4. Thus, the 1D NLPCA approximation to $\mathbf{x}(t_n)$ displayed in Figure 3.4 is a robust result that improves substantially over the 1D PCA approximation.

Figures 3.3 and 3.4 illustrate the strength of NLPCA relative to PCA. It can be proven analytically (Lücke, 1976) that x_3 is uncorrelated with x_1 and x_2. Consequently, the covariance matrix takes on the form

$$\Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & 0 \\ \Gamma_{21} & \Gamma_{22} & 0 \\ 0 & 0 & \Gamma_{33} \end{pmatrix}, \tag{3.5}$$

where $\Gamma_{12} = \Gamma_{21}$ because the covariance matrix is symmetric. One eigenvector of Γ thus lies along the x_3 axis while the other two span the x_1x_2 plane. One of the latter appears as the leading PCA approximation displayed in Figure 3.3. In the PCA description of the Lorenz attractor, variability along x_3 appears in a separate mode from variability in the x_1x_2 plane. However, while variability along x_3 is uncorrelated with variability in the x_1x_2 plane, it is clear upon inspection of Figure 3.2 that these are not *independent* modes of variability. Indeed, large values of $|x_1|$ are associated with large values of x_3.

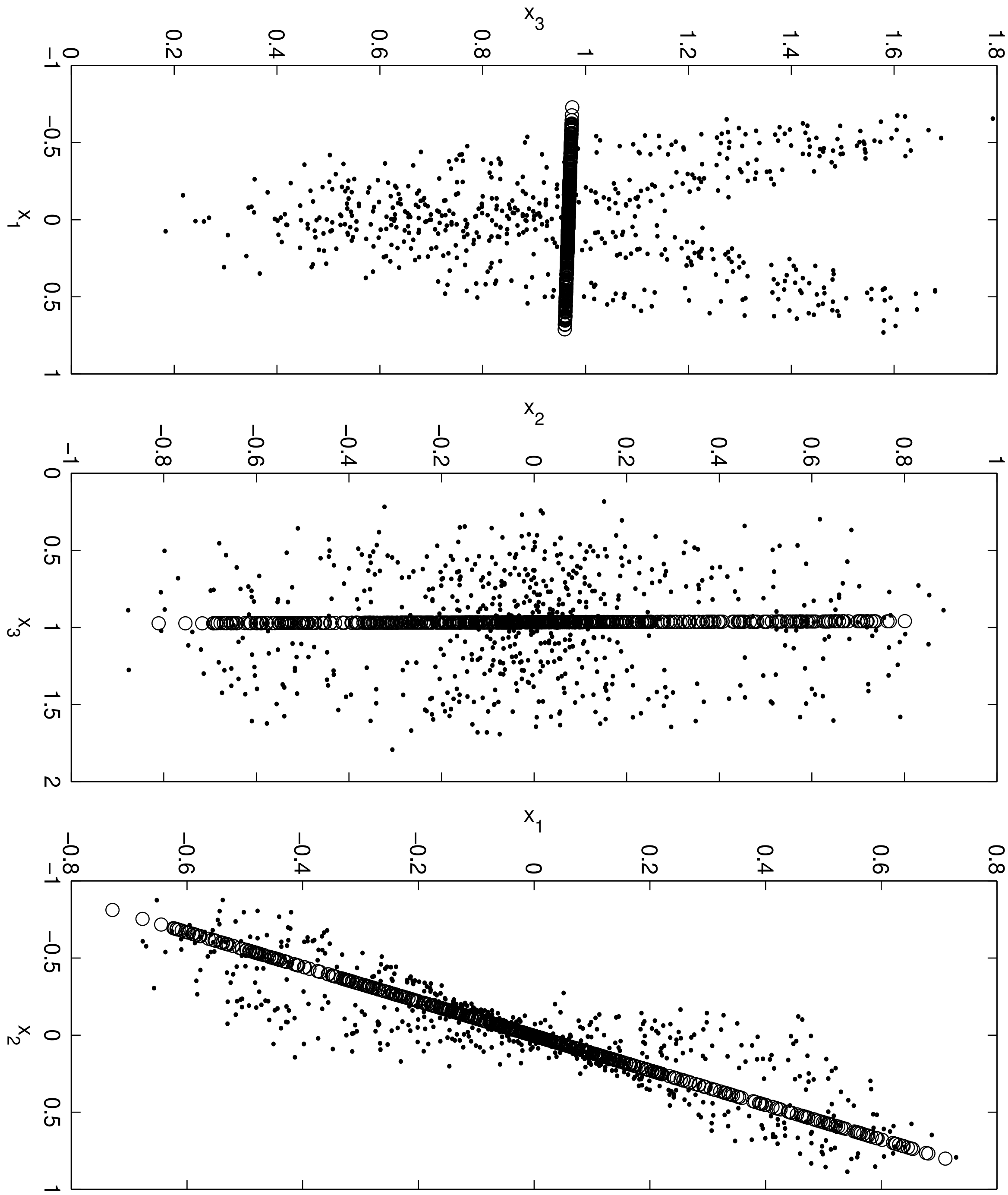


Figure 3.3: Noise-free Lorenz data for a subsample of 584 points and their 1D PCA approximation, projected as in Figure 3.1 (note axes have been rescaled). The dots represent the original data points, the open circles represent points of the approximation.

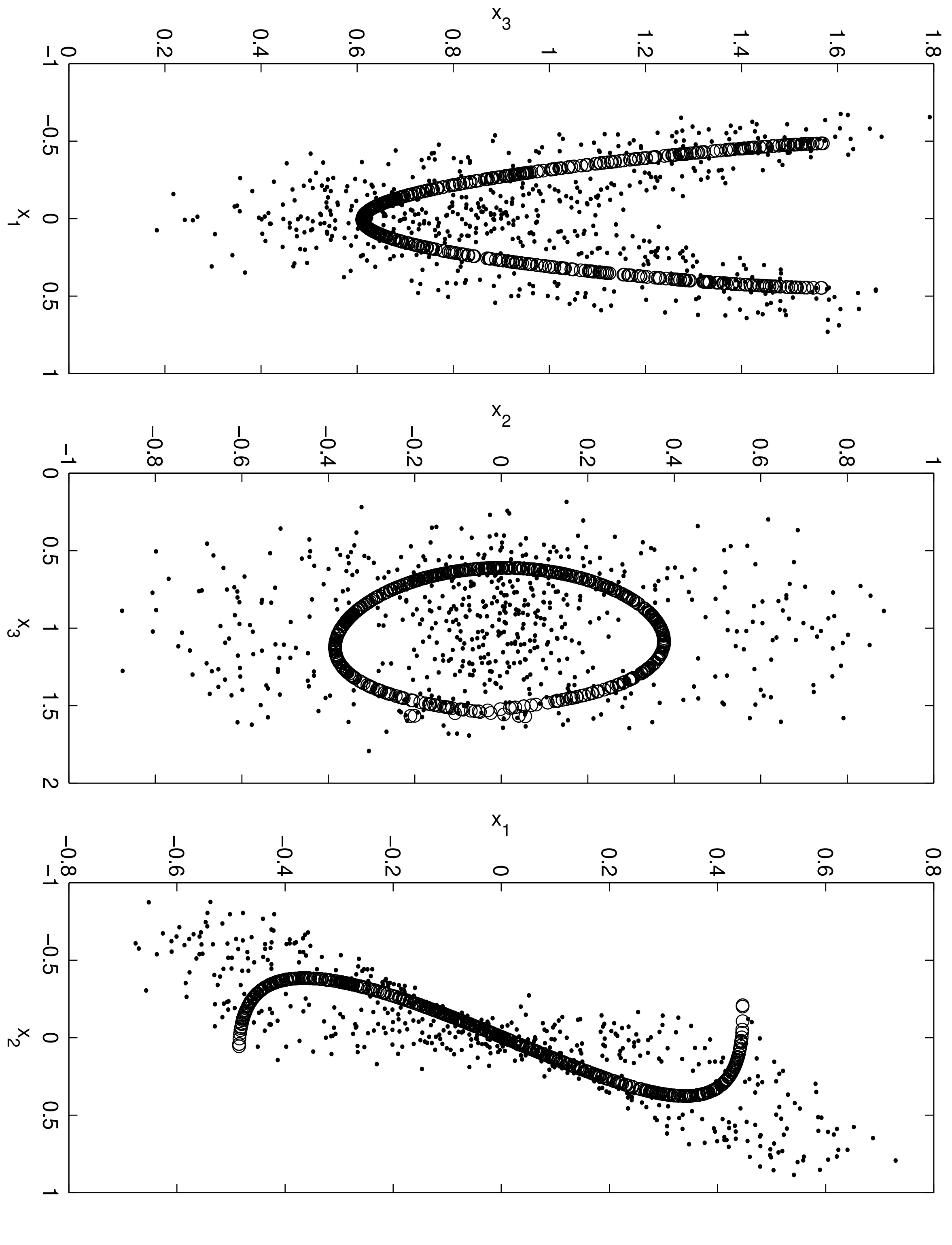


Figure 3.4: As in Figure 3.3, but for the 1D NLPCA approximation.

The 1D NLPCA approximation characterises this dependence between x_1 and x_3, while also describing the covariability of x_1 and x_2 described by the 1D PCA approximation. The power of NLPCA is that it can characterise covariability between variables that are uncorrelated, but not independent, which PCA cannot.

Figure 3.5 displays the 2D PCA approximation of the data $\mathbf{x}(t_n)$ when $\eta = 0$; this surface explains 95% of the variance. The 2D PCA approximation is a flat sheet that characterises the structure of the data as projected in the (x_1, x_3) and (x_2, x_3) planes well but fails to reproduce the structure seen in the projection on the (x_1, x_2) plane. In Figure 3.6, the result of a 2D nonmodal NLPCA of $\mathbf{x}(t_n)$ is shown. This surface explains 99.5% of the variance, implying an order of magnitude reduction in FUV as compared to the PCA result. The network used to perform the NLPCA had 2 neurons in the bottleneck layer and $L = 6$ neurons in the encoding and decoding layers. It was found that decreasing L below 6 also decreased the fraction of variance explained, and increasing it above $L = 6$ had little effect upon the results. The 2D nonmodal NLPCA result is highly robust: a sample of 4 NLPCA models (not shown) has NMSD between curves of at most 0.1%. As with the 1D example considered above, the NLPCA approximation is a substantially better approximation to the original data set than is the PCA approximation.

Consider now a dataset $\mathbf{x}(t_n)$ obtained from equation (3.4) with $\eta = 2.0$. The 1D PCA approximation to $\mathbf{x}(t_n)$ (not shown) explains 59% of the variance. The 1D NLPCA approximation (Figure 3.7), explains 74% of the variance. The curve in Figure 3.7 is very similar to that shown in Figure 3.4 for the $\eta = 0$ case; the two-lobed structure of the data is still manifest at a noise level of $\eta = 2.0$, and the NLPCA is able to recover it. Addressing again the issue of robustness of results, 6 different NLPCA approximations to $\mathbf{x}(t_n)$ were found to have NMSD varying from 0.5% to 3%. These 6 curves agree in their essential details, although the set displays more variability between members than did the corresponding set for $\eta = 0$. Figure 3.8 shows the results of a 2D nonmodal

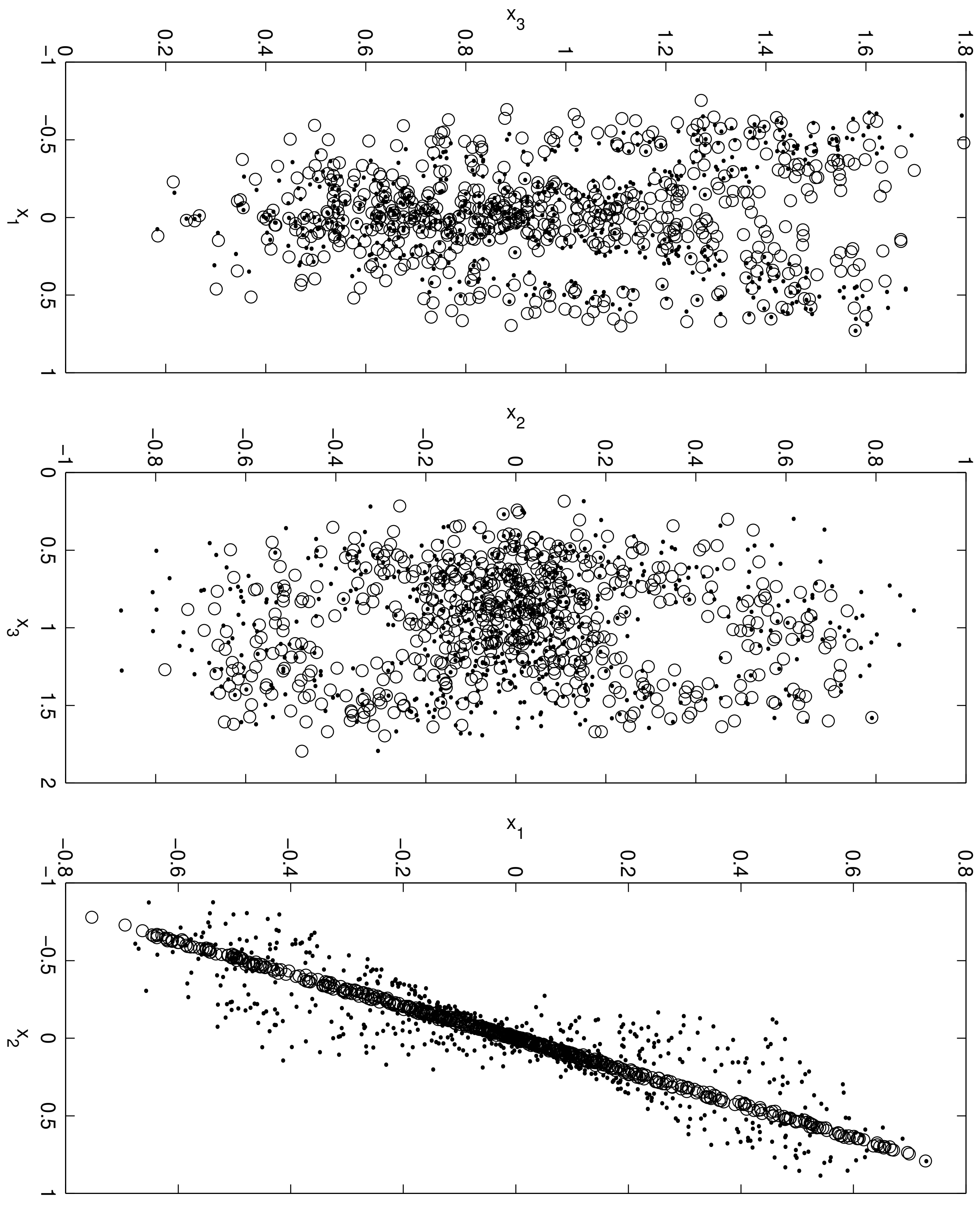


Figure 3.5: Noise-free Lorenz data for a subsample of 584 points and their 2D PCA approximation. The dots represent the original data points, and the open circles the points of the approximation.

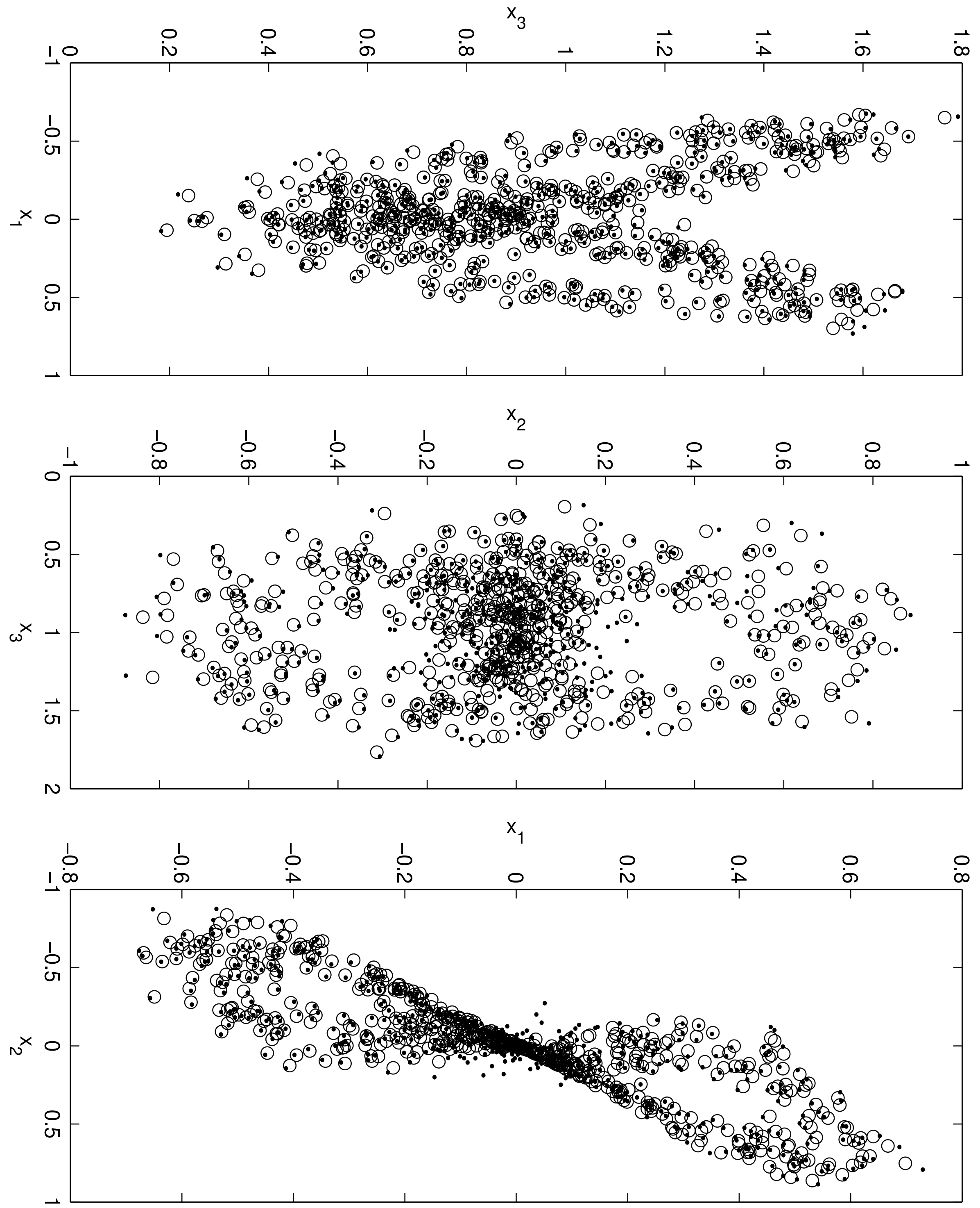


Figure 3.6: As in Figure 3.5, but for the nonmodal 2D NLPCA approximation.

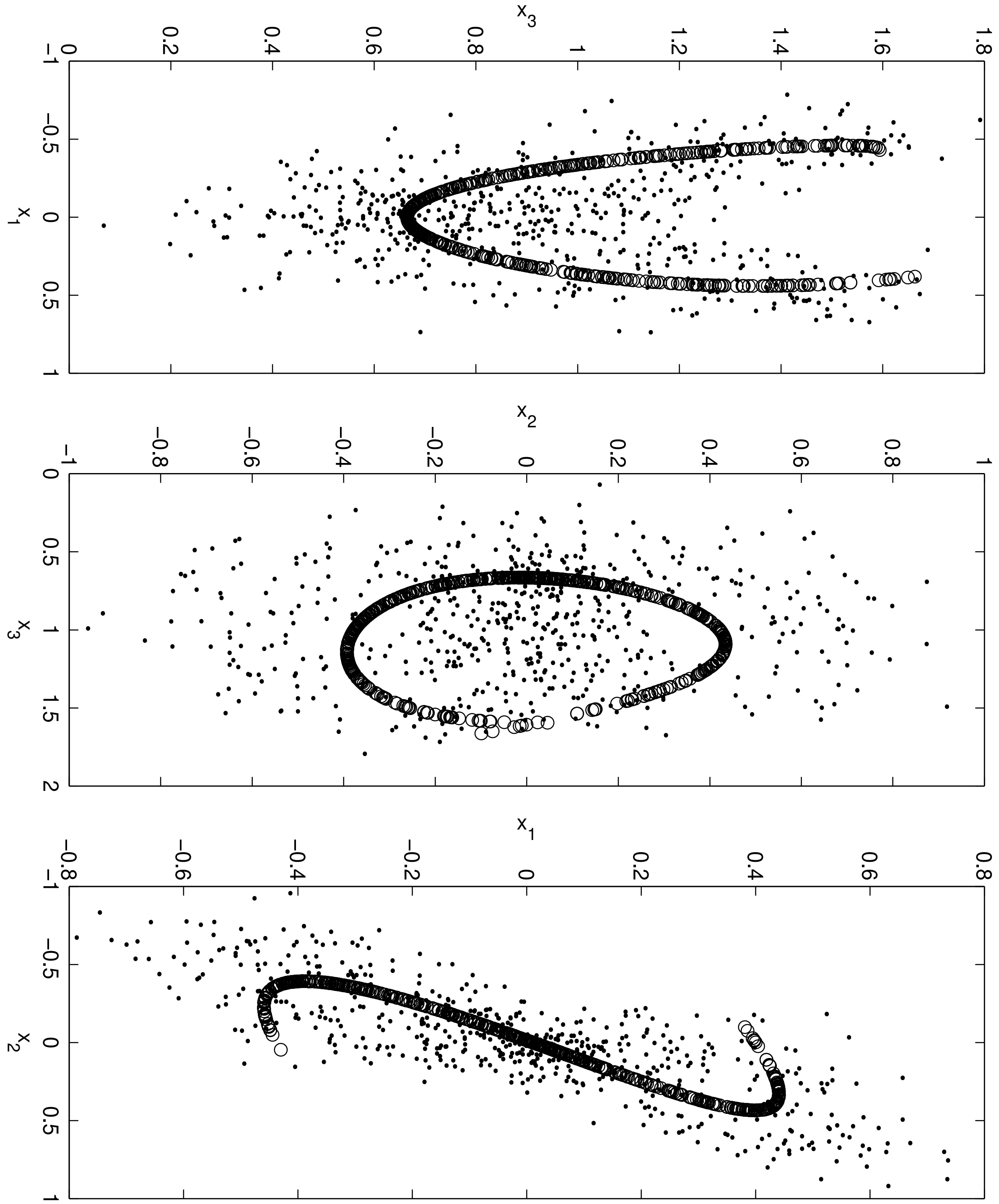


Figure 3.7: Lorenz data with noise level $\eta = 2.0$ for a subsample of 584 points and its 1D NLPCA approximation. Dots represent the data points and the open circles represent points of the approximation.

NLPCA performed on this data set. This explains 97.4% of the variance, in contrast to the 2D PCA approximation (not shown), which explains 94.2%. Thus, the FUV of the 2D nonmodal NLPCA approximation is about half that of the 2D PCA approximation. These results too are robust; the NMSD between different 2D nonmodal NLPCA models was about 0.2%. The 2D nonmodal NLPCA approximation is again an improvement over the 2D PCA approximation, but not by such a substantial margin as was the case when $\eta = 0$. The noise-free Lorenz attractor is very nearly two-dimensional, so the 2D nonmodal NLPCA was able to account for almost all of the variance. The addition of noise acted to smear out this fine fractal structure and made the data cloud more 3 dimensional. The 2D NLPCA applied to this quasi-3D structure could not produce as close an analogue as was the case when $\eta = 0$.

At a noise strength of $\eta = 5.0$, the data set $\mathbf{x}(t_n)$ still has a discernible two-lobed structure, but it is substantially obscured. The 1D PCA approximation (not shown) explains 54% of the variance, whereas the 1D NLPCA approximation (shown in Figure 3.9) explains 65%. Again, the 1D NLPCA approximation to $\mathbf{x}(t_n)$ is qualitatively similar to that obtained in the noise-free case (Figure 3.4). The $\eta = 0$ and $\eta = 5.0$ 1D NLPCA approximations differ at the ends of the curves. Presumably, the structure represented in the former is somewhat washed out by noise in the latter. Four different NLPCA curves for the data obtained with $\eta = 5.0$ share their gross features, but differ fairly substantially in detail. In this case, the NMSD between curves varies between 5% and 10%. The 2D nonmodal NLPCA approximation to these data (not shown) explains 90% of the variance, only slightly more than the 2D PCA approximation, which explains 88% of the variance.

Finally, at a noise level of $\eta = 10.0$ the two-lobed structure of $\mathbf{x}(t_n)$ is no longer obvious, and the data cloud appears as a fairly homogeneous, vaguely ellipsoidal blob. The results of NLPCA by this noise level are no longer robust, tending to be asymmetric,

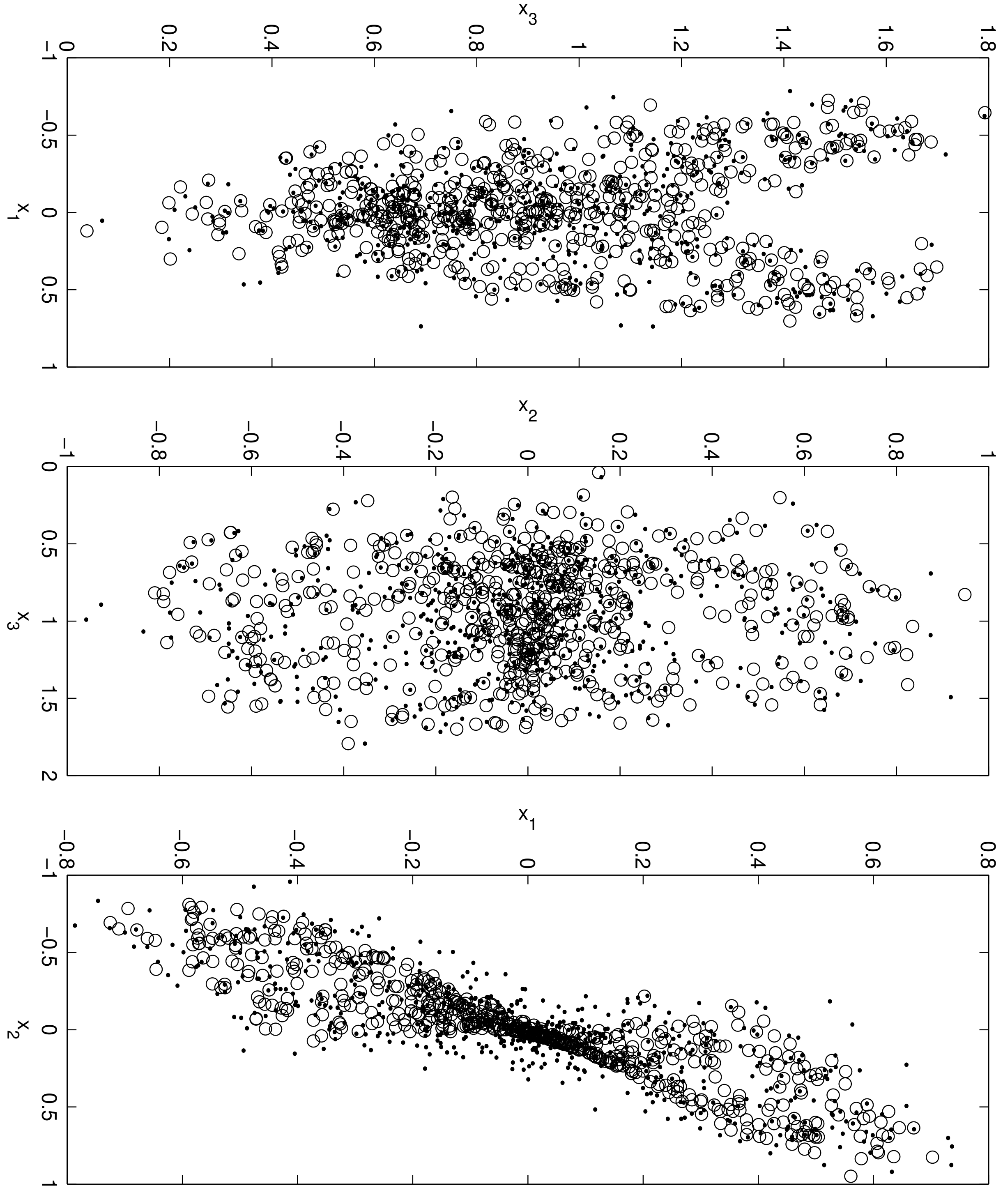


Figure 3.8: As in Figure 3.7, but with the 2D nonmodal NLPCA approximation of the Lorenz data with noise level $\eta = 2.0$.

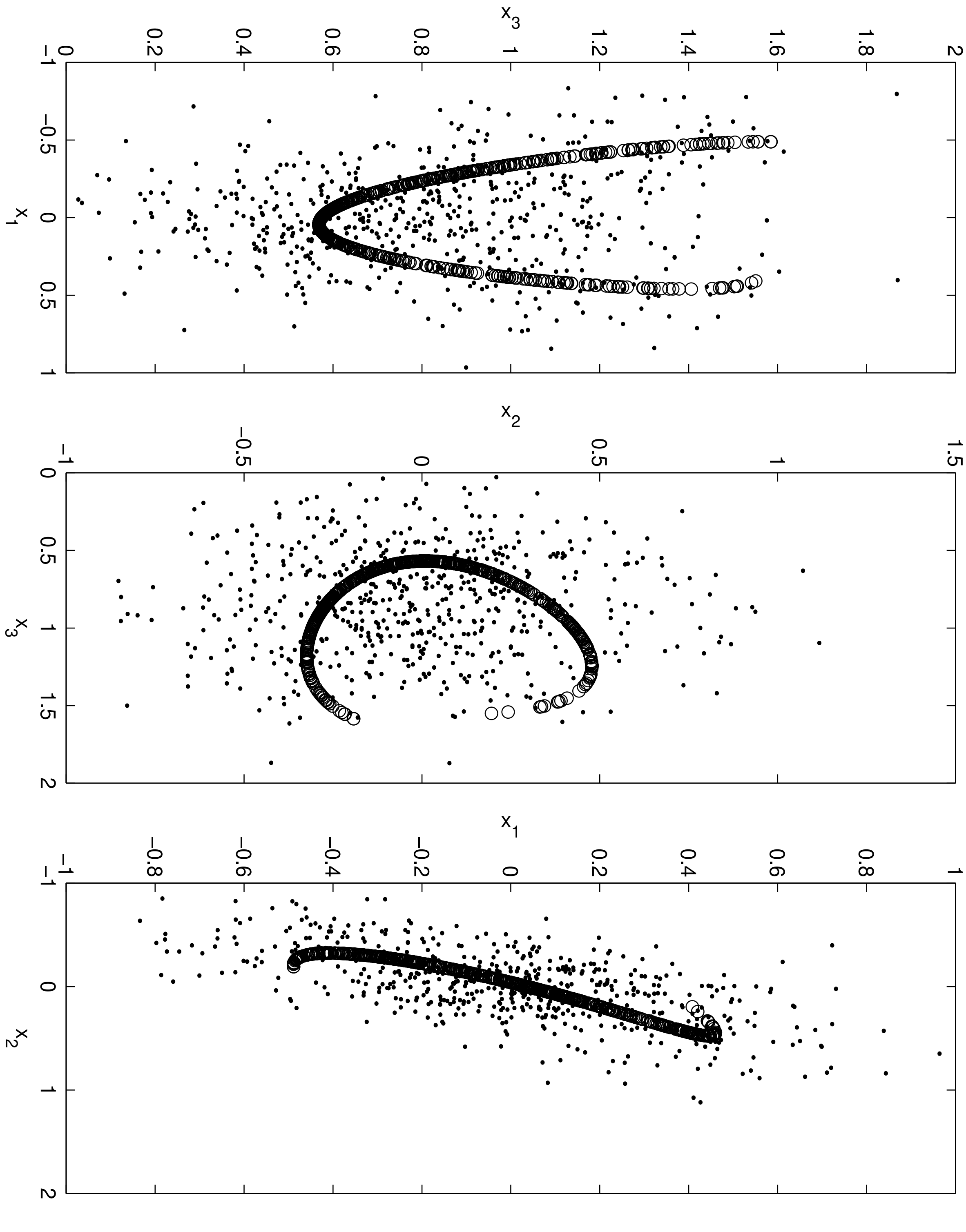


Figure 3.9: As in Figure 3.7, for Lorenz data with noise level $\eta = 5.0$.

convoluted curves. At this noise level, then, NLPCA seems no longer to be a useful technique for characterising low-dimensional nonlinear structure of the data set, precisely because the addition of noise has destroyed this structure.

3.4 Conclusion

The results contained in this chapter demonstrate that NLPCA is able to produce low-dimensional approximations to multivariate data sets that are more representative of the data than the corresponding PCA approximations. The 1D NLPCA approximation to a data set sampled from the Lorenz attractor explains 76% of the variance, in contrast to 60% explained by the 1D PCA approximation, and characterises the two-lobed structure of the data. A 2D nonmodal NLPCA approximation explains 99.5% of the variance, where the 2D PCA approximation explains 95%. As the box-counting dimension of the attractor from which the data is sampled is 2.04, the 2D nonlinear approximation is able to capture almost the entire structure of the data; because the attractor is embedded in $\Re^3$, the 2D PCA approximation cannot do this. With the addition of Gaussian noise of small to moderate strength to the data, NLPCA remains superior to PCA in its ability to detect low-dimensional structure. As the noise level increases, the improvement of NLPCA over PCA decreases, until eventually the noise dominates the signal and NLPCA cannot improve upon PCA. Table 3.1 presents a summary of these results.

Consideration of synthetic data was useful because it is of low dimension and easily visualised, and because the sensitivity of the results of NLPCA to noise level can be assessed through manipulation of the strength of the noise. Real climate data sets however are of a high dimensionality and of a fixed noise level, and it is the potential of NLPCA to produce robust, enlightening approximations to climate data that determines the utility of the method. The analysis of such data sets is the subject of the next two Chapters.

	1D		2D	
	PCA	NLPCA	PCA	NLPCA
$\eta = 0$	60	76	95	99.5
$\eta = 2$	59	74	94.2	97.4
$\eta = 5$	54	65	88	90

Table 3.1: Percentages of variance explained by the 1D and 2D NLPCA approximations to the Lorenz data for the three noise levels η considered.

Chapter 4

Nonlinear Principal Component Analysis of Tropical Indo-Pacific Sea Surface Temperature and Sea Level Pressure

4.1 Introduction

Interannual variability of the Earth's climate system is dominated by the tropical Pacific basin-wide phenomenon known as El Niño and the Southern Oscillation (ENSO) (Philander, 1990). This phenomenon is characterised by alternating periods of anomalously warm or cold water in the eastern equatorial Pacific, alternately weakening or strengthening the zonal sea surface temperature (SST) gradient across the Pacific ocean. These phases of the phenomenon are referred to respectively as El Niño and La Niña. Associated with these changes in SST are pronounced changes in the zonal gradient of thermocline depth and sea surface height. In the atmosphere, El Niño (La Niña) events are associated with a slackening (strengthening) of the zonally-oriented Walker circulation, implying a reduction (increase) in wind stress applied to the ocean surface associated with the easterly Trade Winds, and an eastward (westward) shift in the region of deep convection. Variability of the Walker circulation manifests itself as an east-west dipolar fluctuation in sea level pressure (SLP) known as the Southern Oscillation. As was originally described by Bjerknes (1969), these changes in atmospheric circulation feed back on the ocean and reinforce the original SST anomalies. In the late 1980's, a mechanism was proposed for the transition between El Niño and La Niña events that invoked the dynamics of oceanic baroclinic waves in the so-called "equatorial waveguide" (Suarez and Schopf, 1988; Battisti and Hirst, 1989). Denoted the "delayed oscillator" mechanism, this has become the dominant paradigm for the negative feedback that terminates ENSO events (Battisti and Sarachik, 1995). ENSO variability is aperiodic, with power primarily in the 4-7 year band (Tangang et al., 1998). Models invoked to explain the aperiodicity of the variability range from the stochastic forcing of a linear system (Penland, 1996) to low-dimensional dynamics of a chaotic system (Jin et al, 1996); the actual character of ENSO dynamics is still a subject of some debate, although recent evidence favours the former of the above models (Penland et al., 1999).

While the physical mechanisms producing ENSO are thought to be mainly confined to the equatorial Pacific, its effects are global in scale (Philander, 1990; Trenberth et al., 1998). In consequence, forecasts of ENSO variability have been attempted by a number of researchers, and throughout the 1990s it has been the paradigmatic problem of climate prediction (Barnston et al., 1994; Barnston et al., 1999). ENSO dynamics has quite certainly been the most intensively studied problem in climate physics for the last decade. ENSO variability is often diagnosed from observations using linear statistical tools, in particular PCA; the dominant ENSO signal in tropical Pacific SST and SLP is usually identified with the leading PCA approximation to these data sets.

In this chapter, NLPCA is applied to climatic data sets relevant to ENSO variability: tropical Pacific Ocean sea surface temperature and tropical Indo-Pacific sea level pressure. In the case of SST, NLPCA is able to produce one- and two-dimensional approximations that are of greater fidelity to the original data than the corresponding one- and two-dimensional PCA approximations. In particular, the 1D SST NLPCA describes ENSO variability in a manner that characterises the asymmetry in spatial distribution of temperature anomalies between El Niño and La Niña events, which are treated symmetrically in the 1D PCA approximation. The improvement of the NLPCA approximations over PCA are more modest, but still notable, in the case of SLP.

In Chapter 3, I considered the application of NLPCA to synthetic data sets of sufficiently low dimension, and of sufficiently low noise level, that their underlying low-dimensional structure was manifest. Nonlinear PCA was able to recover this structure, even in the presence of moderate noise levels. Fundamentally, however, NLPCA is only of practical use in climate research if it is able to robustly characterise low-dimensional structure in real data sets arising from the climate system, and improve upon the results obtained by traditional linear methods. I show here that this is indeed the case, and thereby demonstrate the potential utility of NLPCA in the analysis of climatic data sets.

4.2 Data and Model Building

The SST data considered consist of monthly-averaged NOAA sea surface temperatures for the tropical Pacific Ocean. The data are on a $2^{\mathrm{O}} \times 2^{\mathrm{O}}$ grid from 19S to 19N, and from 125E to 69W, and span the period from January 1950 to December 1998. This data set was produced using the PCA-based interpolation method developed by Smith et al. (1996). A climatological annual cycle was calculated by averaging the data for each calendar month, and monthly SST anomalies (SSTA) were defined relative to this annual cycle.

The SLP data were COADS monthly-averaged sea level pressure (SLP) over the tropical Indo-Pacific area (Woodruff et al., 1987) on a $2^{\mathrm{O}} \times 2^{\mathrm{O}}$ grid from 27S to 19N, and from 31E to 67W, covering the period from January 1950 to December 1998. The annual cycle was removed in the same fashion as for the SST data to produce sea level pressure anomalies (SLPA). The SLPA field was then smoothed in time using a 3-month running mean filter and a 1-2-1 filter in each spatial direction.

In the analysis of both the SSTA and SLPA data, the early stopping algorithm described in the previous chapter was used. In all analyses, 20% of the data was held aside in a validation set, for which network performance was monitored as training proceeded. Training was stopped when this performance began to degrade, or after 5000 iterations, whichever came first. It was found that increasing the maximum number of iterations beyond 5000 did not affect the results of the analysis.

4.3 Tropical Pacific Sea Surface Temperature

To render the NLPCA problem tractable, the SSTA data set was pre-processed by projecting it on the space of its first 10 EOF modes $\{\mathbf{e}_k\}_{k=1}^{10}$, in which 91.4% of the total variance is contained. Doing so takes advantage of the *data compression* aspect of PCA, which is a feature distinct from *feature extraction*, for which NLPCA shall be used. Such pre-processing of data to reduce the problem to a manageable size is common in rotated PCA (Barnston and Livezy, 1987) and in statistical forecasting (Barnston, 1994; Tangang et al., 1998). The first 3 EOF spatial patterns of SSTA are displayed in Figure 4.1; these explain, respectively, 57.6%, 10.9%, and 6.8% of the *total* SSTA variance. A scatterplot of the two leading principal component time series is shown in Figure 4.2. It can also be considered to be a plot of the projection of the data into the plane spanned by the first two SSTA EOF modes. The time series corresponding to these two PCA modes are uncorrelated, but they are clearly not *independent*; the distribution of the data appears to be markedly non-Gaussian. Indeed, Figure 4.2 indicates that there is an inverted U-shape underlying the data, such that both strongly positive and negative values of the first PCA time series are associated with negative values of the second. Physically, this coupling of the PC1 and PC2 time series describes the fact that the most positive SST anomalies during an average El Niño event lie closer to the eastern boundary of the

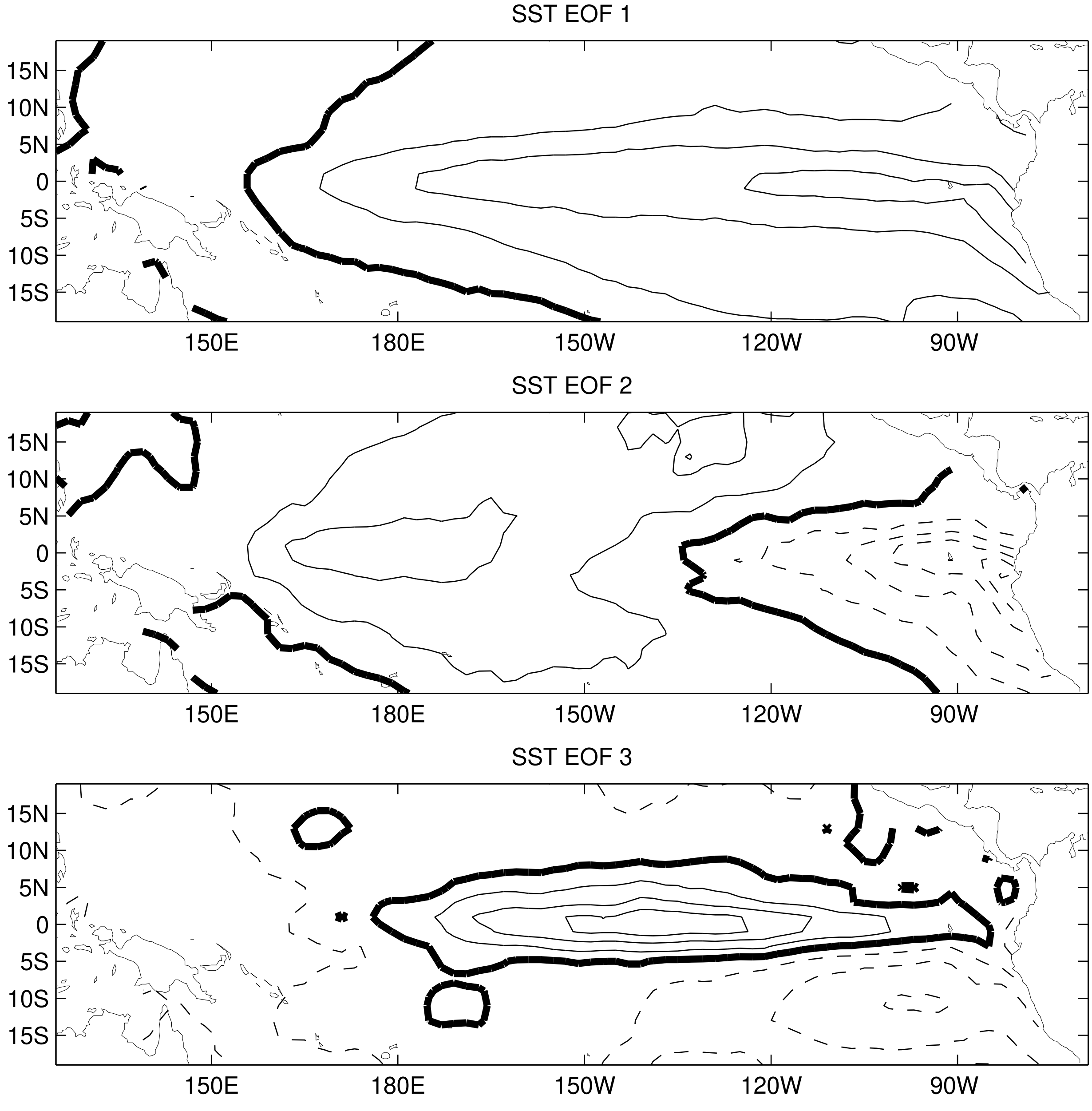


Figure 4.1: Spatial patterns of the first three SSTA EOF patterns, normalised to unit magnitude. The contour interval is 0.02, the zero contour is in bold, and negative contours are dashed.

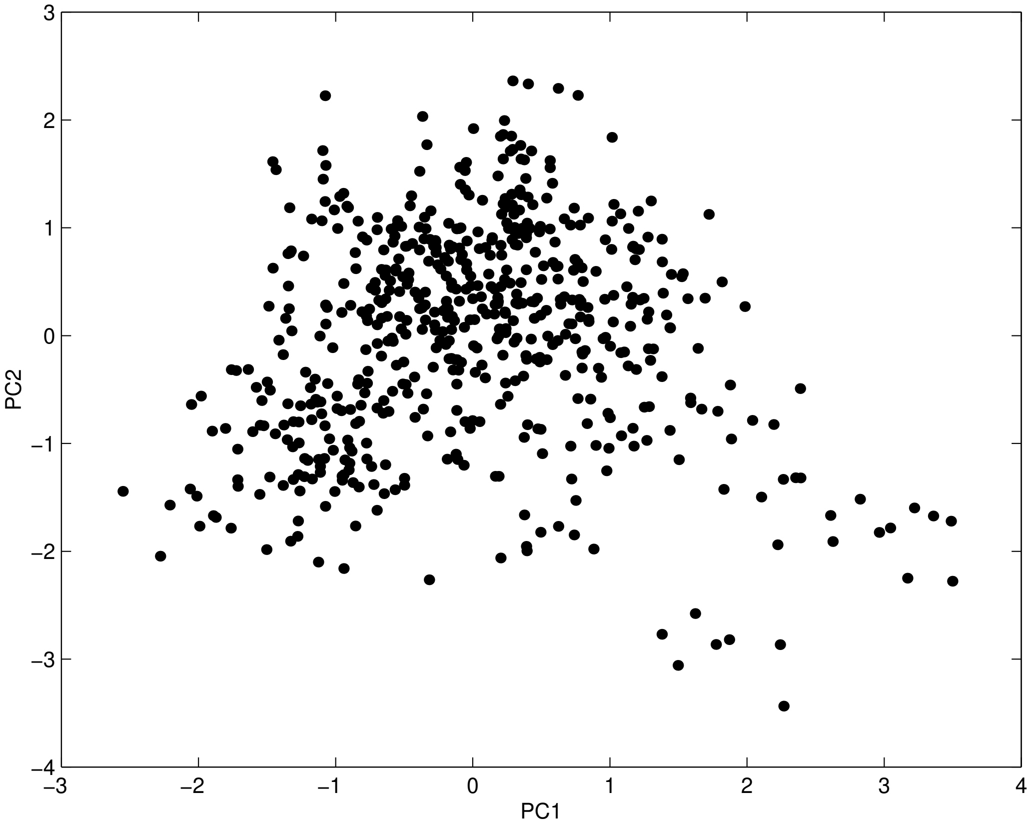


Figure 4.2: Scatterplot of SSTA data projected onto the plane spanned by the first two EOF patterns.

Pacific than do the coldest anomalies during an average La Niña event. I shall return to this point later.

Consider first a modal NLPCA decomposition of this SSTA data. Mode 1, found using a network with $L = 4$ nodes in the encoding and decoding layers and a single neuron in the bottleneck layer, explains 69.1% of the variance in the 10-dimensional EOF space, and thus 63.3% of the variance in the total SSTA data, as compared to 57.6% explained by the 1D PCA approximation. Four candidate models were obtained from an ensemble of 20; these models differed with NMSD of at most 1%. Projections of the first NLPCA mode onto the subspaces spanned by the SSTA EOFs $(\mathbf{e}_1, \mathbf{e}_2)$, $(\mathbf{e}_2, \mathbf{e}_3)$, $(\mathbf{e}_1, \mathbf{e}_3)$, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ are shown in Figure 4.3 (a)-(d), respectively. All four projections are shown because it is difficult to understand the structure of the NLPCA approximation from a single projection. This difficulty is particularly evident in Figure 4.3(b): the curve, viewed edge-on, appears to be self-intersecting, when in fact the other projections demonstrate this is not the case. Figure 4.3(a) indicates that NLPCA mode 1 characterises the U-shaped structure discussed in the previous paragraph; NLPCA mode 1 is primarily a mixture of PCA modes 1 and 2. Associated with this mode is the standardised time series

$$\alpha_1(t_n) = \frac{s_{\mathbf{f}}(\mathbf{X}(t_n)) - < s_{\mathbf{f}} >}{< (s_{\mathbf{f}} - < s_{\mathbf{f}} >)^2 >^{1/2}}, \tag{4.1}$$

corresponding to the output of the single neuron in the bottleneck layer. A plot of $\alpha_1(t_n)$ appears in Figure 4.4(a). This time series bears a strong resemblance to the Niño 3.4 time series (defined as the average SSTA over a box from 7S to 7N, and from 119W to 171W) displayed in Figure 4.4(b); the correlation coefficient between the two series is 0.88.

In contrast to PCA, no single spatial pattern is associated with any given NLPCA

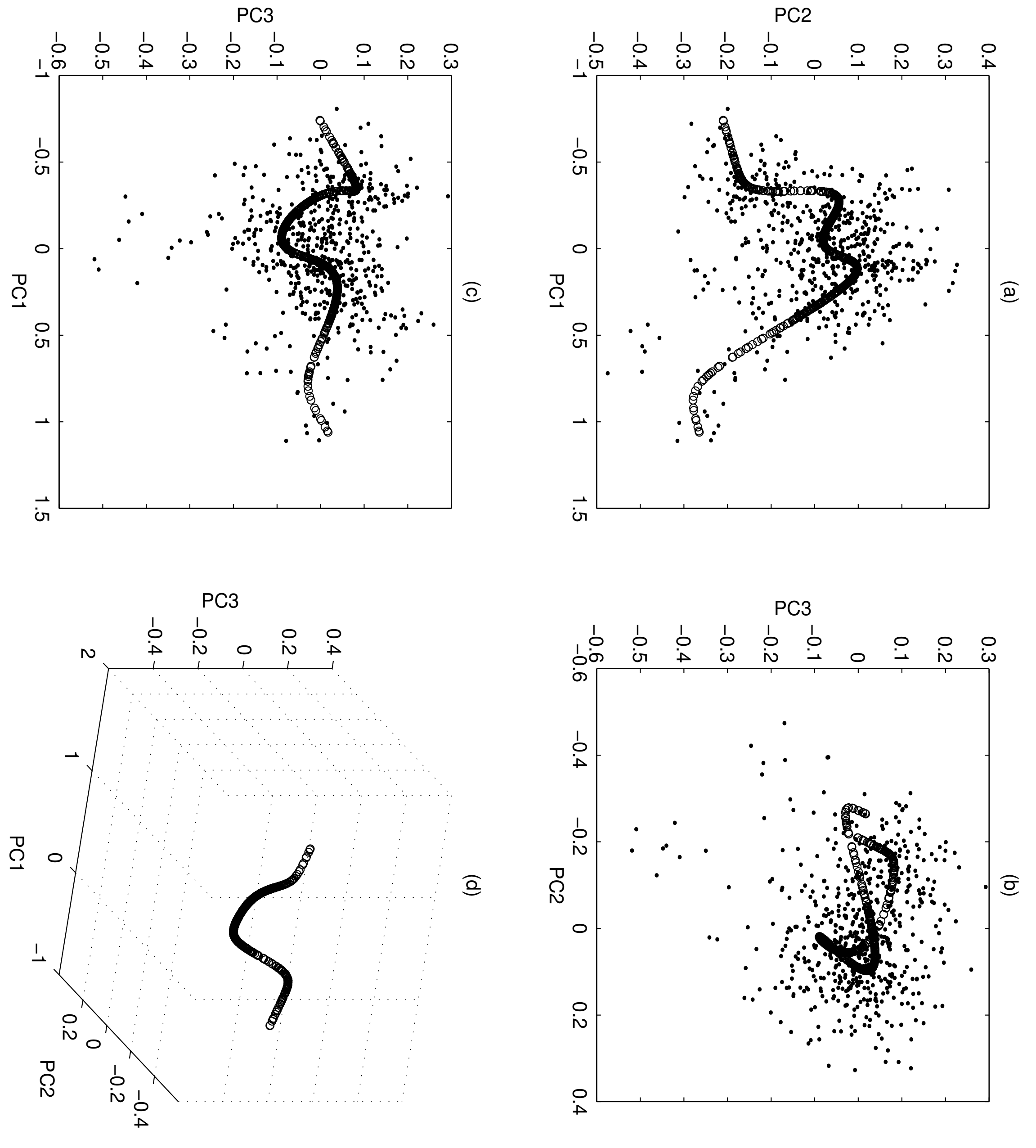


Figure 4.3: Scatterplot of SSTA data (points) and SSTA NLPCA mode 1 approximation (open circles) projected onto the planes spanned by EOFs (a) ($\mathbf{e}_1$,$\mathbf{e}_2$), (b) ($\mathbf{e}_2$,$\mathbf{e}_3$), (c) ($\mathbf{e}_1$,$\mathbf{e}_3$). (d) shows a scatterplot of the 1D NLPCA approximation projected into the subspace ($\mathbf{e}_1$, $\mathbf{e}_2$,$\mathbf{e}_3$).

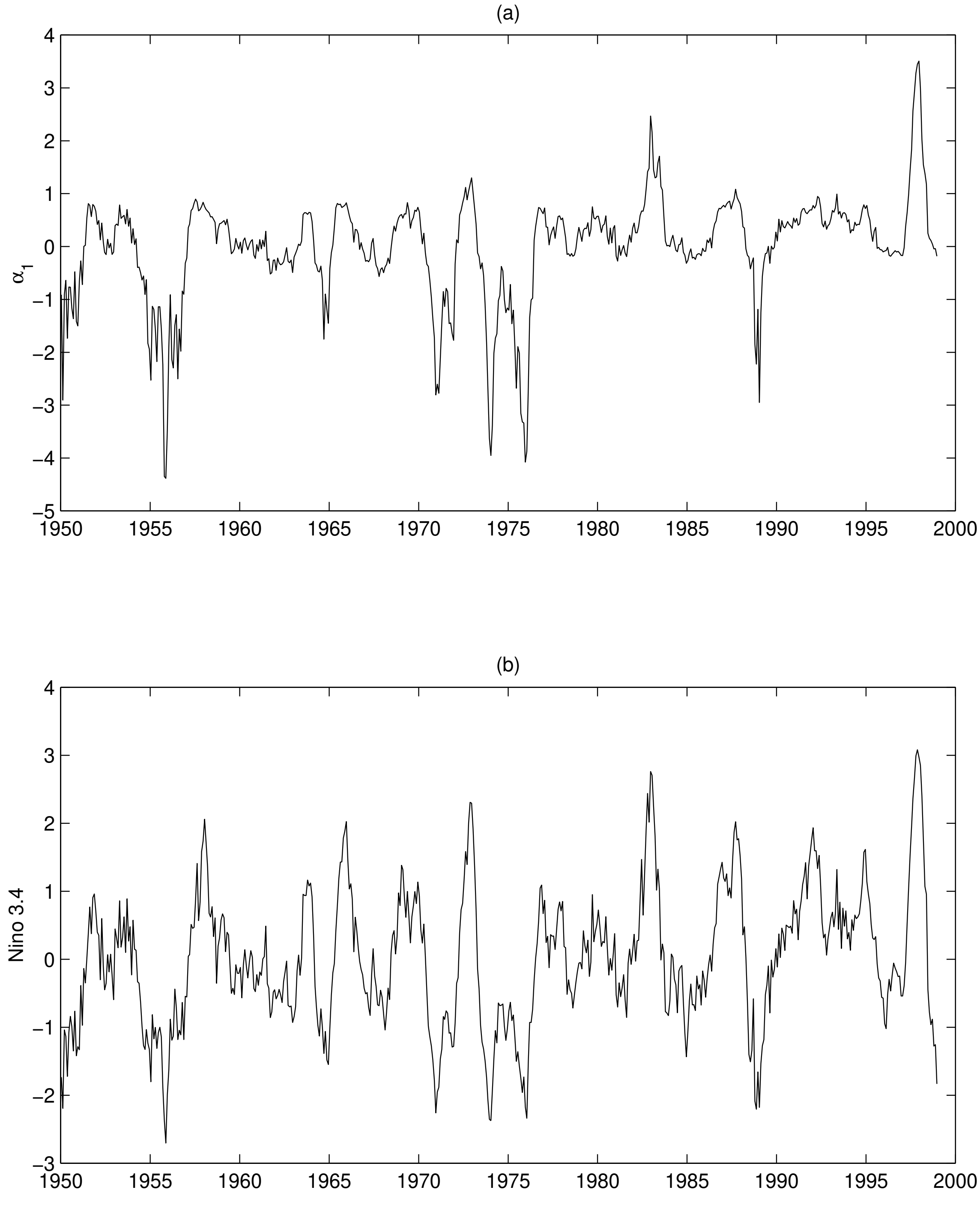


Figure 4.4: (a) Plot of $\alpha_1(t_n) = s_f(\mathbf{X}(t_n))$, the standardised time series associated with SSTA NLPCA mode 1. (b) Plot of the Niño 3.4 index normalised to unit variance.

mode. The approximation $\hat{\mathbf{X}}$, however, provides a sequence of patterns that can be visualised cinematographically. This cinematographic interpretation is implicit in traditional PCA: the 1D PCA approximation $\hat{\mathbf{X}}(t_n) = (\mathbf{e}_1 \cdot \mathbf{X}(t_n))\mathbf{e}_1$ describes the evolution in time of a *standing oscillation*. This oscillation has a fixed spatial structure with an amplitude that varies in time. The more general approximation $\hat{\mathbf{X}}(t_n) = (\mathbf{f} \circ s_{\mathbf{f}})(\mathbf{X}(t_n))$, with $s_{\mathbf{f}}$ and $\mathbf{f}$ nonlinear, is not so constrained, and can characterise more complex lower-dimensional variability. There is no *a priori* reason to expect the optimal 1D approximation to a spatial field to be a standing wave - but standing waves are the only such approximations that traditional PCA can produce. The power of NLPCA lies in its ability to characterise more general lower-dimensional structure.

Figure 4.5 displays maps characterising the first NLPCA mode $\hat{\mathbf{X}}^{(1)}$ corresponding to values of the time series $\alpha_1 = -3.5, -1.5, -0.75, -0.25, 0.25, 0.75, 1.5, 3.5$. These values were chosen to provide a representative sample of spatial structures associated with the NLPCA approximation. Clearly, NLPCA mode 1 describes the evolution of average ENSO events, in contrast with PCA mode 1, which describes only the standing oscillation associated with average ENSO variability. This difference between NLPCA and PCA modes 1 results from the spatial asymmetry between the average warm and cold phases of ENSO. In particular, warm events described by NLPCA mode 1 display the strongest anomalies near the Peruvian coast, whereas the cold events are strongest near 150W. This asymmetry in the evolution of NLPCA mode 1 arises because NLPCA mode 1 mixes PCA modes 1 and 2: for both El Niño and La Niña events, the PCA mode 2 spatial map (Figure 4.1(b)) enters into the NLPCA mode 1 approximation with the same (negative) sign.

This spatial asymmetry between average El Niño and La Niña events is manifest in a composite study. Figure 4.6(a) is a composite of NDJ (November, December, January)

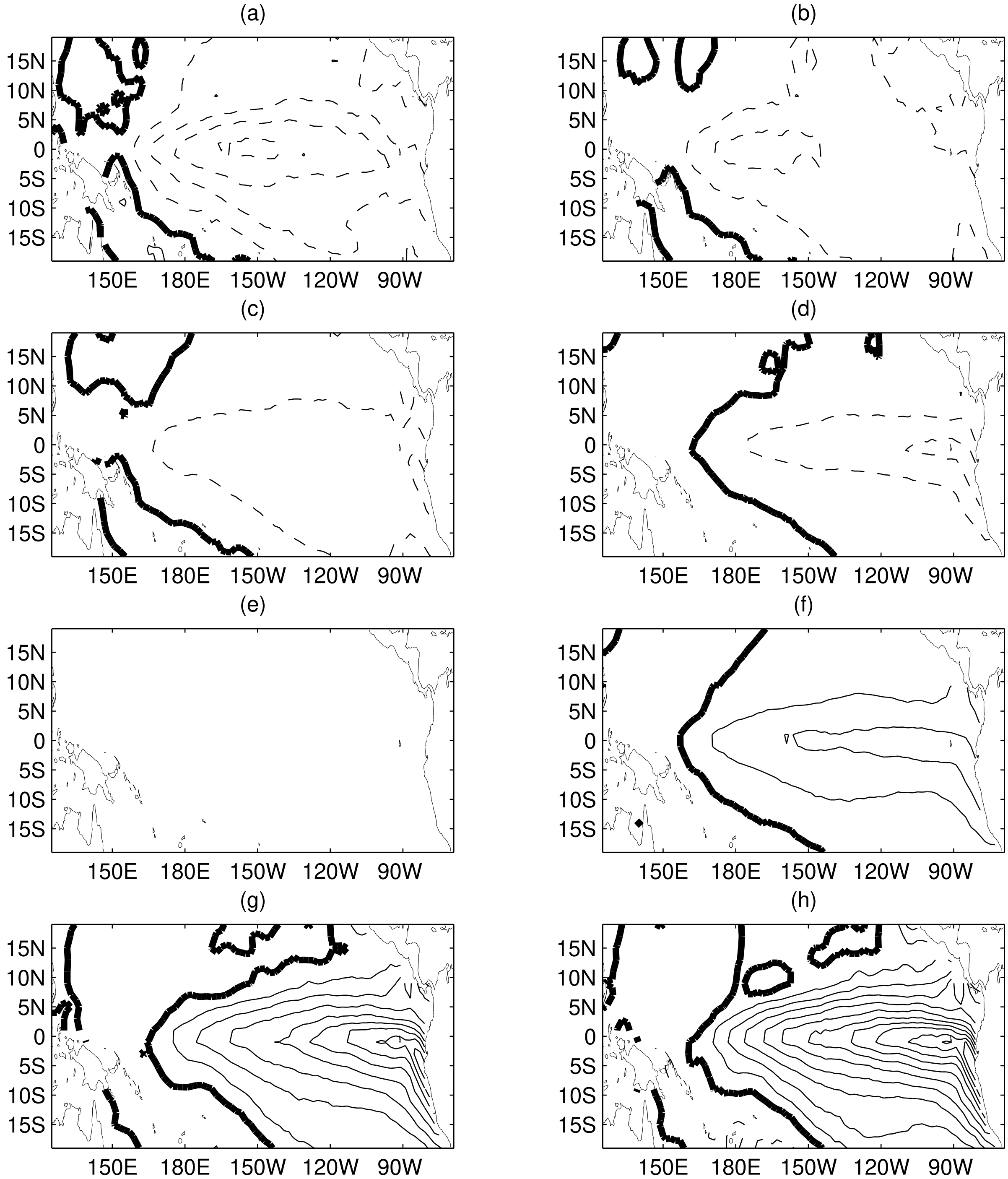


Figure 4.5: Sequence of spatial maps characterising SSTA NLPCA mode 1 for (a) $\alpha_1 = -3.5$ (b) $\alpha_1 = -1.5$ (c) $\alpha_1 = -.75$ (d) $\alpha_1 = -.25$ (e) $\alpha_1 = .25$ (f) $\alpha_1 = 0.75$ (g) $\alpha_1 = 1.5$ (h) $\alpha_1 = 3.5$. Contour interval: $0.5^{\circ}C$.

averaged SSTA for those years in which the NDJ Niño 3.4 index is greater than one standard deviation above the long-term average; Figure 4.6(b) is the same for those NDJ for which Niño 3.4 is less than one standard deviation below the long-term average. This averaging period was used for the composites because NDJ displays the largest variance of all 3-month seasons. These two maps correspond to the SSTA patterns of an average El Niño and an average La Niña, respectively. Note that, consistent with the maps corresponding to the 1D NLPCA approximation (Figure 4.5), the largest SST anomalies tend to be located in the central Pacific during average La Niña events and in the eastern Pacific during average El Niño events. This asymmetry in the composite fields was previously noted by Hoerling et al. (1997). The symmetric component of the composite El Niño and La Niña maps, as defined in Appendix B, is displayed in Figure 4.6(c). This map, which in a rough sense characterises the pattern in the composites that is related nonlinearly to the Niño 3.4 time series, bears a strong resemblance to SST EOF mode 2 (Figure 4.1(b)). In fact, the spatial correlation between the two maps is -0.90. The antisymmetric component of the composite (not shown) bears a strong resemblance to EOF 1; the pattern correlation between these two maps is 0.975. Thus, PCA mode 1 may be interpreted as characterising the component of average ENSO behaviour that is antisymmetric between average El Niño and La Niña events. By mixing EOF modes 1 and 2, NLPCA mode 1 is able to characterise the spatial asymmetry between average El Niño and La Niña events. The bias of SST toward warm anomalies in the eastern part of the basin and toward cold anomalies in the western and central parts is also evident in the study of Burgers and Stephenson (1999) , who calculate the skewness of the observed SSTA distribution. It is interesting to note the striking similarity between their map of the spatial distribution of skewness (their Figure 3(a)) and the symmetric component of the SSTA composite displayed in Figure 4.6(c).

A final comparison of the 1D NLPCA and 1D PCA approximations is given in Figures

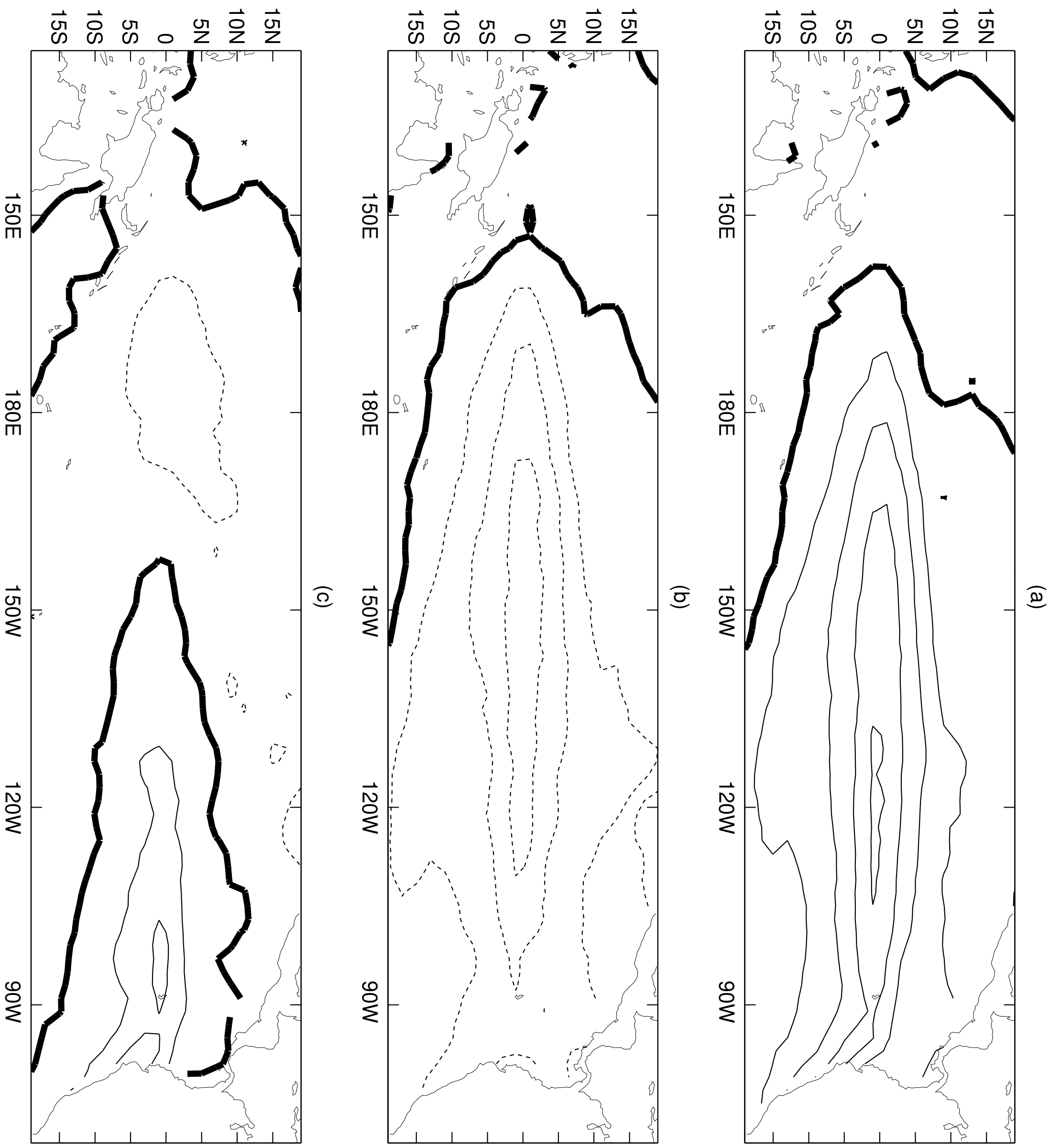


Figure 4.6: SSTA composite maps for "average" (a) El Niño and (b) La Niña events. Contour interval: $0.5^{o}C$. (c) Symmetric component of composites (a) and (b). Contour interval: $0.1^{o}C$. See text for definition of composites and of the symmetric component.

4.7 (a)-(c), which show respectively maps of the pointwise correlation between the original SSTA data and the 1D NLPCA approximation, the pointwise correlation between SSTA and the 1D PCA approximation, and the difference between these two pointwise correlations. The two approximations are equally well correlated with the original data over the eastern-central half of the Pacific Ocean, except near the Ecuadorian coast, where the NLPCA correlations are somewhat higher than those of PCA. In the western Pacific, and in particular in the neighbourhood of the EOF mode 1 zero line, the 1D NLPCA approximation displays greater fidelity to the original data, as measured by pointwise correlation, than does the 1D PCA approximation.

Now consider SSTA NLPCA mode 2, which was calculated using a network containing $L = 3$ neurons in the encoding and decoding layers. Figure 4.8 displays mode 2 projected onto the subspaces spanned by SSTA EOFs $(\mathbf{e}_1, \mathbf{e}_2)$, $(\mathbf{e}_2, \mathbf{e}_3)$, $(\mathbf{e}_1, \mathbf{e}_3)$, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. The 5 candidate models from an ensemble of 20 trials differed from each other with NMSD always less than 4%. NLPCA mode 2 explains 11.1% of the variance in the original SSTA data. The associated standardised time series, $\alpha_2(t_n)$ is shown in Figure 4.9. Interestingly, the correlation coefficient between $\alpha_1(t_n)$ and $\alpha_2(t_n)$ is -0.06; the two time series are uncorrelated. Figure 4.10 displays maps of SSTA NLPCA mode 2, $\hat{\mathbf{X}}^{(2)}$, corresponding to $\alpha_2 = -4, -1, -0.25, 0, 0.15, 0.25, 0.3, 0.35, 0.4, 0.5, 0.75, 1.5$. These values of α_2 were selected to present a representative sample of maps describing NLPCA mode 2. When α_2 is strongly negative, SSTA NLPCA mode 2 is characterised by negative anomalies in the central and western Pacific and positive anomalies in the eastern Pacific. As α_2 increases through zero, the anomalies decrease in magnitude, while the positive anomalies in the eastern part of the basin become increasingly concentrated in the equatorial region. Eventually, the region of positive SSTA breaks off from the coast of South America and migrates into the central Pacific. As α_2 increases further, the SSTA pattern becomes the opposite of that for α_2 near zero, with positive anomalies in the central and western

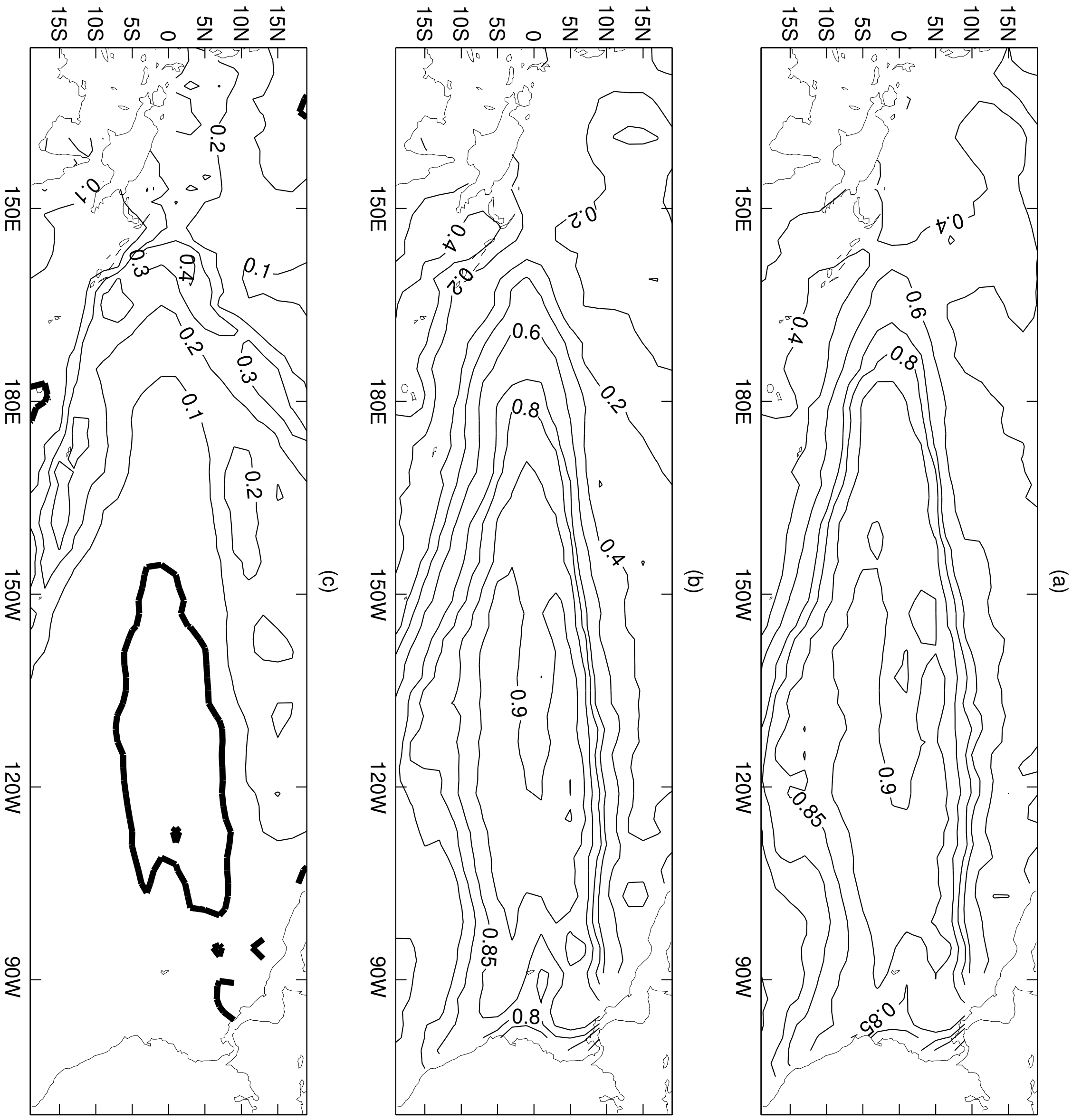


Figure 4.7: Map of pointwise correlation coefficient between observed SSTA and (a) 1D NLPCA approximation, (b) 1D PCA approximation, and (c) difference between (a) and (b).

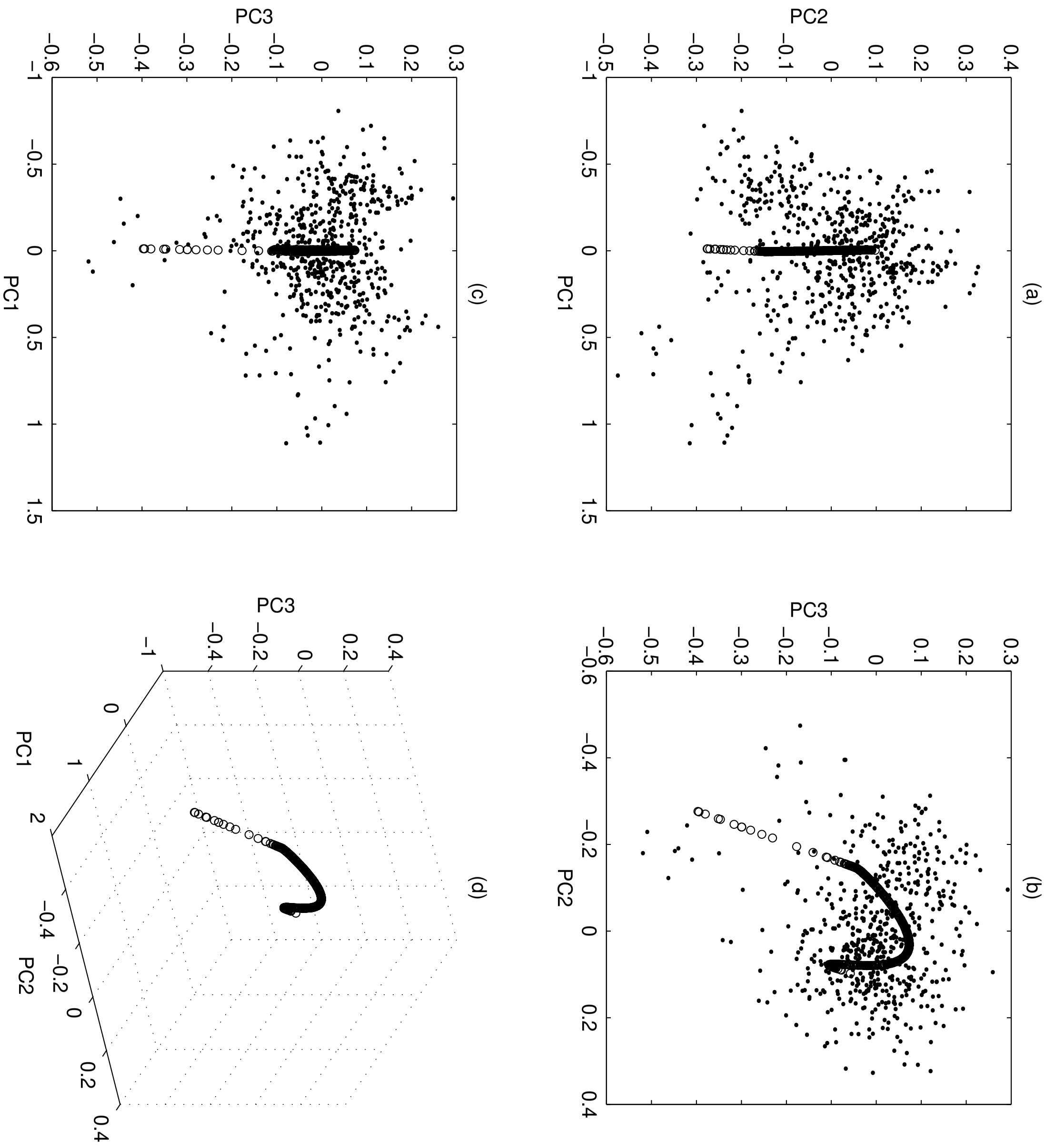


Figure 4.8: As for Figure 4.3, but for SSTA NLPCA mode 2.

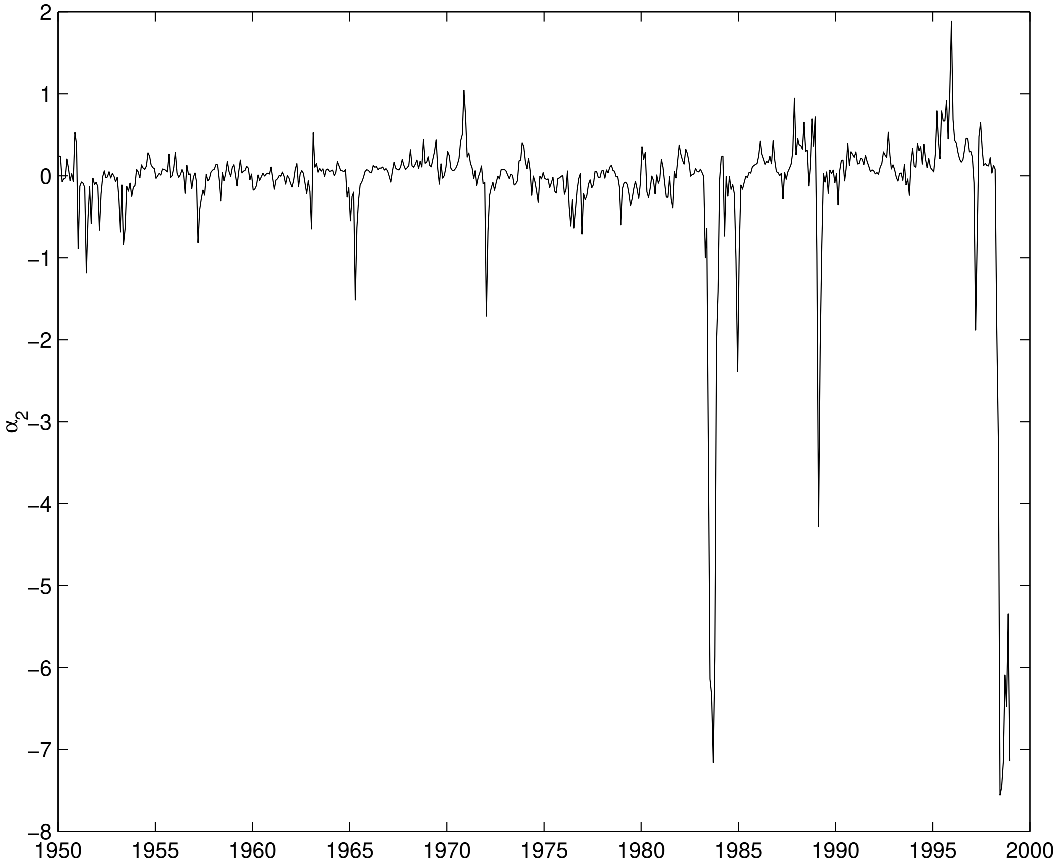


Figure 4.9: As for Figure 4.4(a), but for $\alpha_2(t_n)$, the time series corresponding to SSTA NLPCA mode 2.

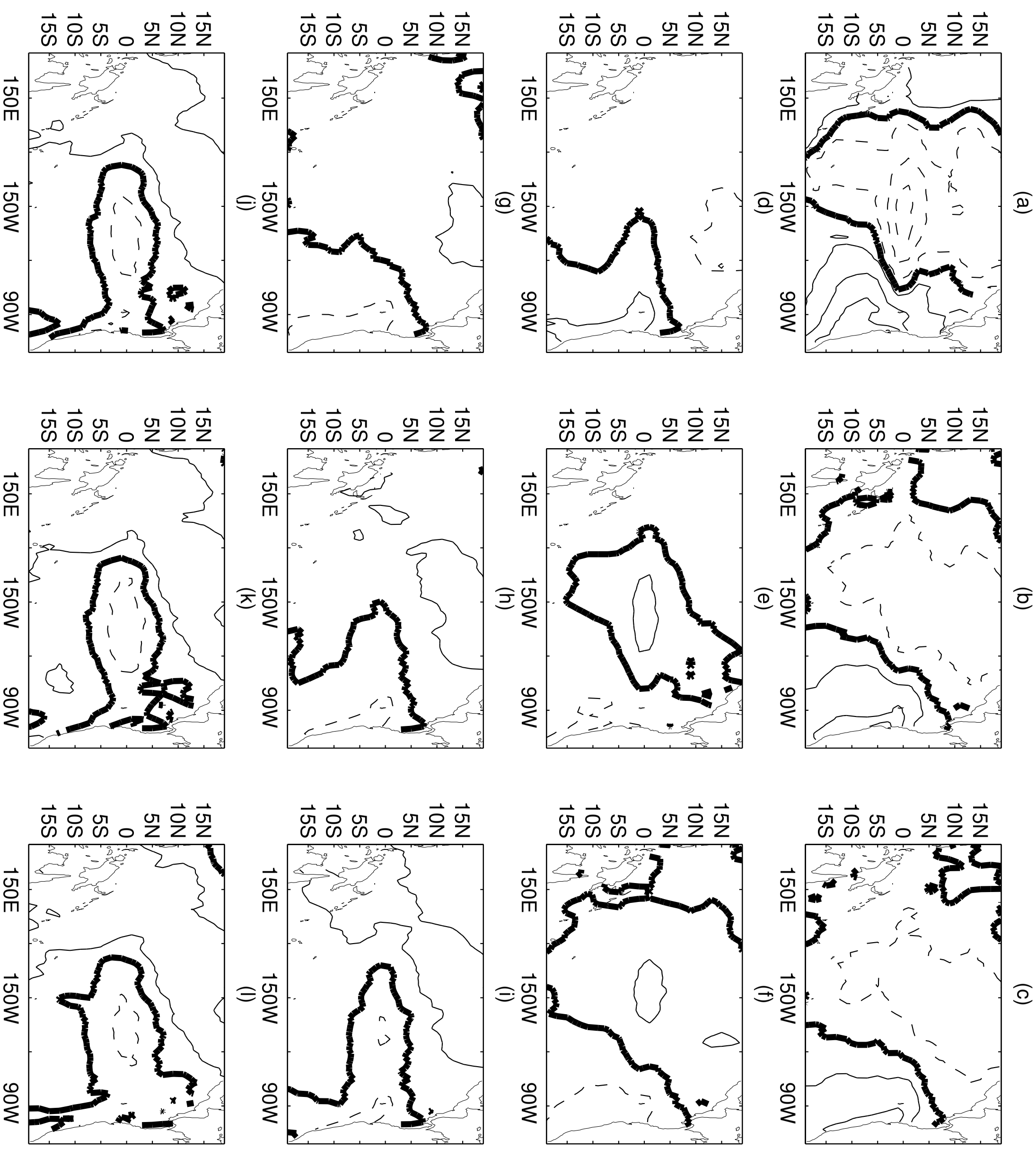


Figure 4.10: Maps corresponding to SSTA NLPCA mode 2 for (a) $\alpha_2 = -4$ (b) $\alpha_2 = -1$ (c) $\alpha_2 = -0.25$ (d) $\alpha_2 = 0$ (e) $\alpha_2 = 0.15$ (f) $\alpha_2 = 0.25$ (g) $\alpha_2 = 0.3$ (h) $\alpha_2 = 0.35$ (i) $\alpha_2 = 0.4$ (j) $\alpha_2 = 0.5$ (k) $\alpha_2 = 0.75$ (l) $\alpha_2 = 1.5$. Contour interval: $0.5^{O}C$.

Pacific and negative anomalies in the east. Finally, for α_2 near the extreme positive part of its range, SSTA NLPCA mode 2 is characterised by negative anomalies along the equator, extending to the dateline, with positive anomalies throughout the rest of the basin. Because the anomalies are often concentrated along the equator, it is reasonable to associate this mode with certain aspects of ENSO variability not captured by NLPCA mode 1. Indeed, it is interesting to note from Figure 4.4 that NLPCA mode 2 is more active in the later part of the record than in the earlier. The two strong minima in $\alpha_2(t_n)$ coincide with the decay phases of the large El Niño events of 1982/83 and 1997/98, describing the lingering patches of warm water in the eastern tropical Pacific observed during these periods. Two of the three weaker minima in the more active period are associated with the peaks of the La Niña events of 1985/85 and 1988/89. Indeed, the cold anomalies during La Niñas in the later period are somewhat stronger, more concentrated in the central Pacific Ocean, and weaker in the eastern Pacific Ocean than La Niña events in the earlier part of the record, as indicated by a composite analysis (Figure 4.11). A number of studies have noted a shift in ENSO variability in 1977 (see, e.g., Wang, 1995). The apparent nonstationarity of $\alpha_2(t_n)$ is consistent with a shift at this time, although the precise timing of the shift in $\alpha_2(t_n)$ is not obvious. The 1977 shift is in fact manifest in SSTA NLPCA mode 1 time series $\alpha_1(t_n)$ (Figure 4.4); the time series is biased toward negative extrema before 1977 and toward positive extrema after 1977. It should however be noted that this shift may simply be an artifact of the technique used to reconstruct the gridded SST data set.

Thus, the first mode of the modal NLPCA decomposition of tropical Pacific SSTA describes the average variability associated with the ENSO phenomenon, and nicely characterises the asymmetry in spatial structure between average El Niño and La Niña events. The second mode characterises the difference in evolution between different ENSO events, and in particular, displays a nonstationarity consistent with the observed "regime

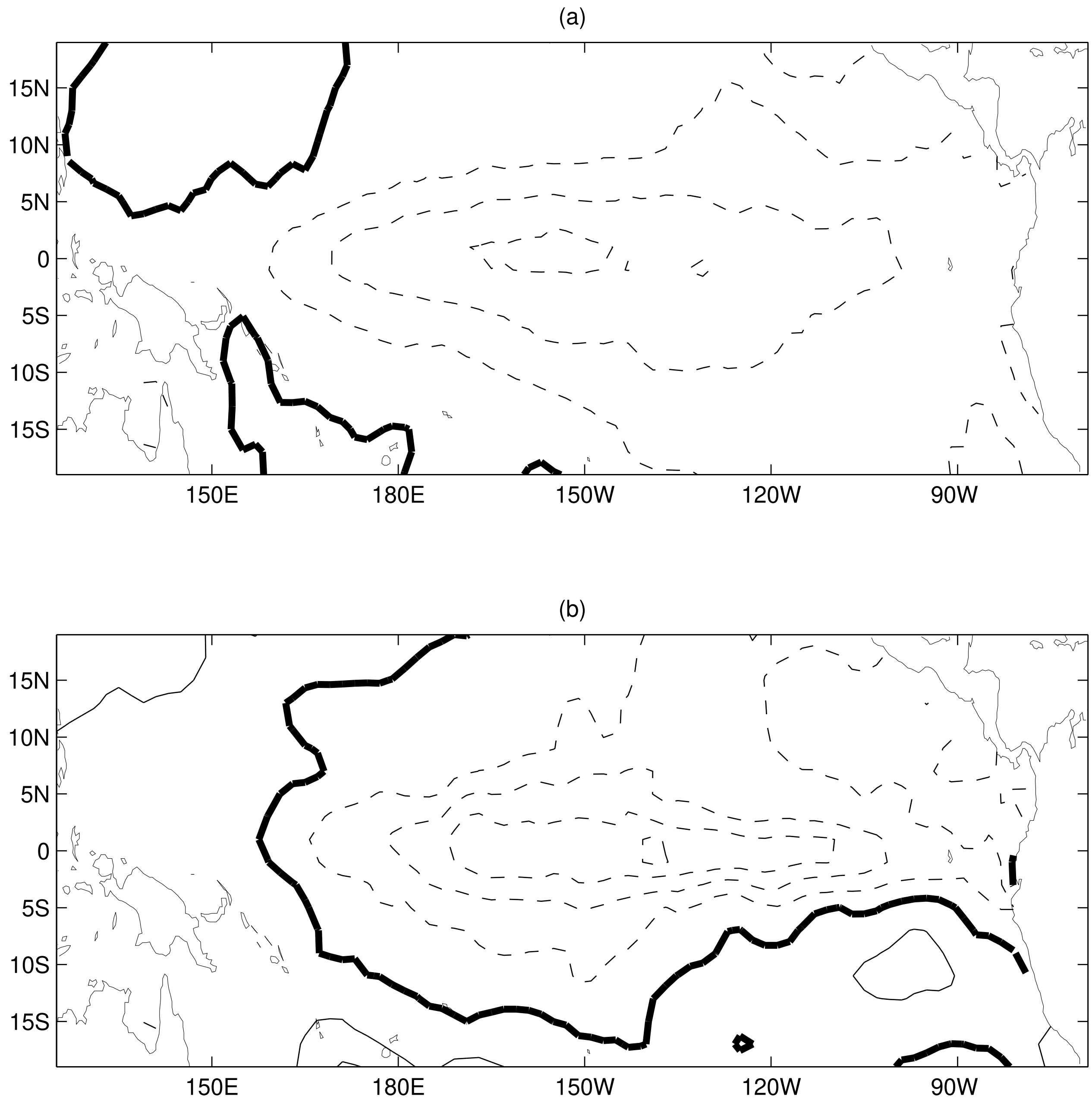


Figure 4.11: Composites of SSTA for La Niña events (a) before 1977 and (b) after 1977. Contour interval $0.5^{\circ}C$.

shift" in ENSO variability.

Plots of a 2D nonmodal NLPCA approximation of the SSTA data (ie, using 2 neurons in the bottleneck layer), projected in the subspaces spanned by SSTA EOFs $(\mathbf{e}_1, \mathbf{e}_2)$, $(\mathbf{e}_2, \mathbf{e}_3)$, $(\mathbf{e}_1, \mathbf{e}_3)$, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, are shown in Figure 4.12 (a)-(d). The associated network used $L = 6$ neurons in the encoding and decoding layers, and the NMSD between candidate models (8 out of an ensemble of 20) varied between 1% and 3%. The 2D nonmodal NLPCA approximation explains 79.0% of the variance in the truncated data set, and thus 72.2% of the variance of the original data. The time series corresponding to the output of the bottleneck layers, denoted $(\beta_1, \beta_2)(t_n) = \mathbf{s_f}(\mathbf{X}(t_n))$, are shown in Figure 4.13. These two time series are highly correlated with each other (r = -.835) and with the Niño 3.4 index ($r_1 = -.879$ and $r_2 = .889$, respectively). Because the 2D nonmodal NLPCA depends on 2 parameters, β_1 and β_2, it is difficult to visualise the results using a sequence of maps as was done with the modal NLPCA in Figures 4.5 and 4.10. Figures 4.14(a) and (b) display maps of the pointwise correlation coefficient between the SSTA data and the 2D PCA and 2D nonmodal NLPCA approximations, respectively. The 2D nonmodal NLPCA approximation produces higher correlations than the 2D PCA approximation in the central equatorial, western, and southeastern Pacific, and slightly lower correlations in the eastern equatorial Pacific.

It is worth considering the time series $\beta_1(t_n)$ and $\beta_2(t_n)$ in more detail. As was discussed in Chapter 2, the parameterisation $\mathbf{s_f}(\mathbf{X}(t_n))$ of the P-dimensional surface determined by NLPCA is only defined up to an arbitrary homeomorphism (i.e., a continuous, one-to-one, and onto function with continuous inverse). That is, for an arbitrary homeomorphism $\mathbf{g} : \Re^P \mapsto \Re^P$, the time series $\mathbf{g}(\mathbf{s_f}(\mathbf{X}(t_n)))$ is also an acceptable parameterisation of the surface, because $\mathbf{f} \circ \mathbf{s_f} = (\mathbf{f} \circ \mathbf{g}^{-1}) \circ (\mathbf{g} \circ \mathbf{s_f})$. In particular, for any homeomorphism $\mathbf{g} : \Re^2 \mapsto \Re^2$, $(g_1(t_n), g_2(t_n)) = \mathbf{g}(\beta_1(t_n), \beta_2(t_n))$ parameterises the surface found by 2D nonmodal NLPCA. Which parameterisation is determined by the

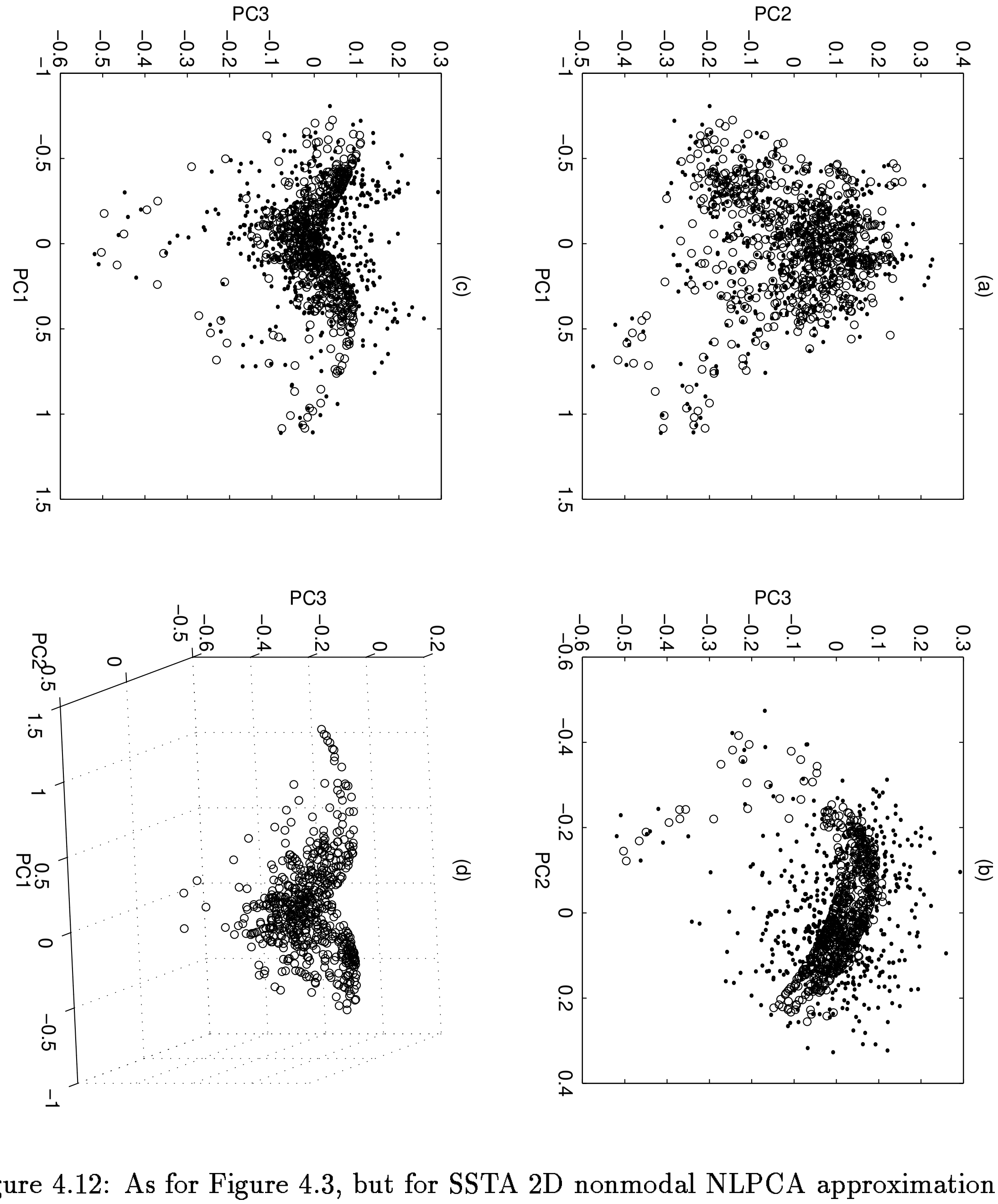


Figure 4.12: As for Figure 4.3, but for SSTA 2D nonmodal NLPCA approximation.

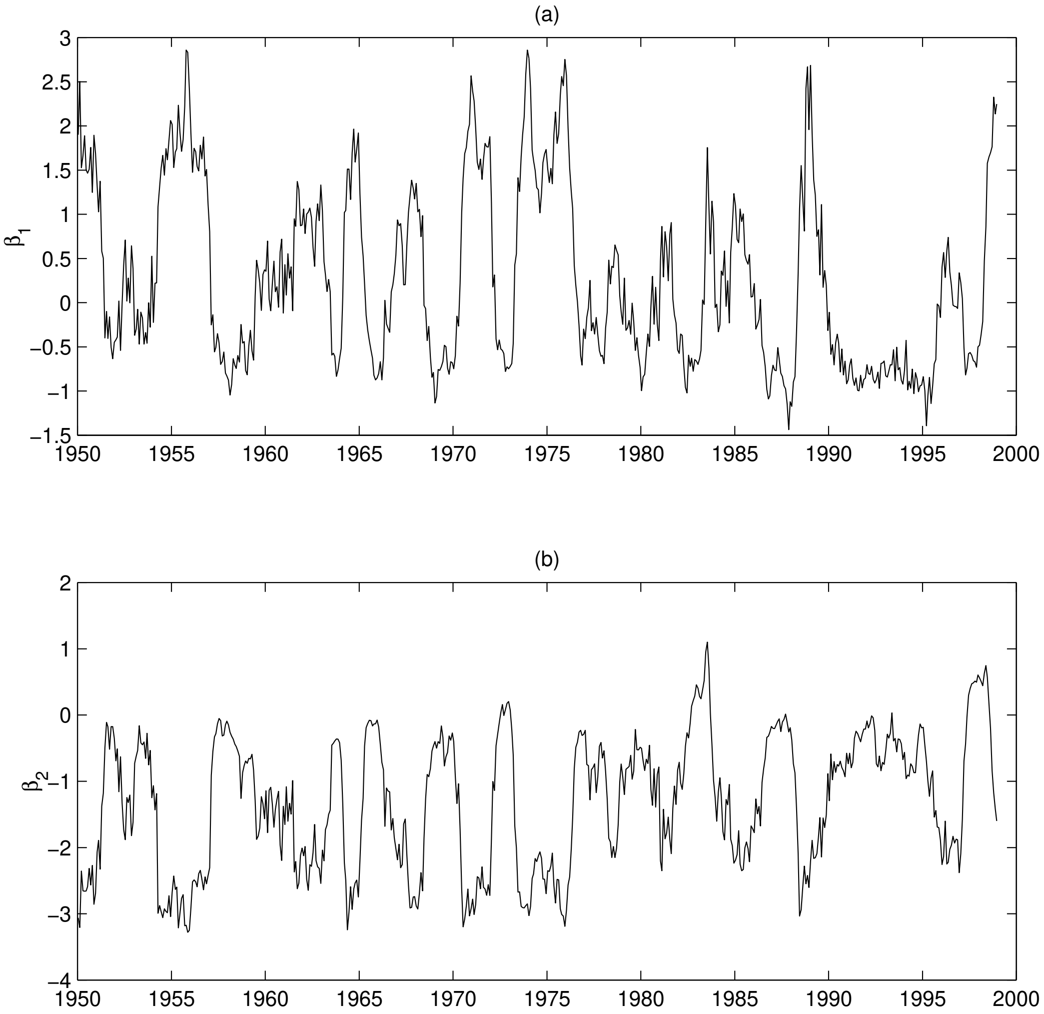


Figure 4.13: Time series (a) $\beta_1(t_n)$ and (b) $\beta_2(t_n)$, where $(\beta_1, \beta_2)(t_n) = \mathbf{s_f}(\mathbf{X}(t_n))$ is the pair of time series associated with the SSTA 2D nonmodal NLPCA approximation. Both time series have been normalised to unit variance.

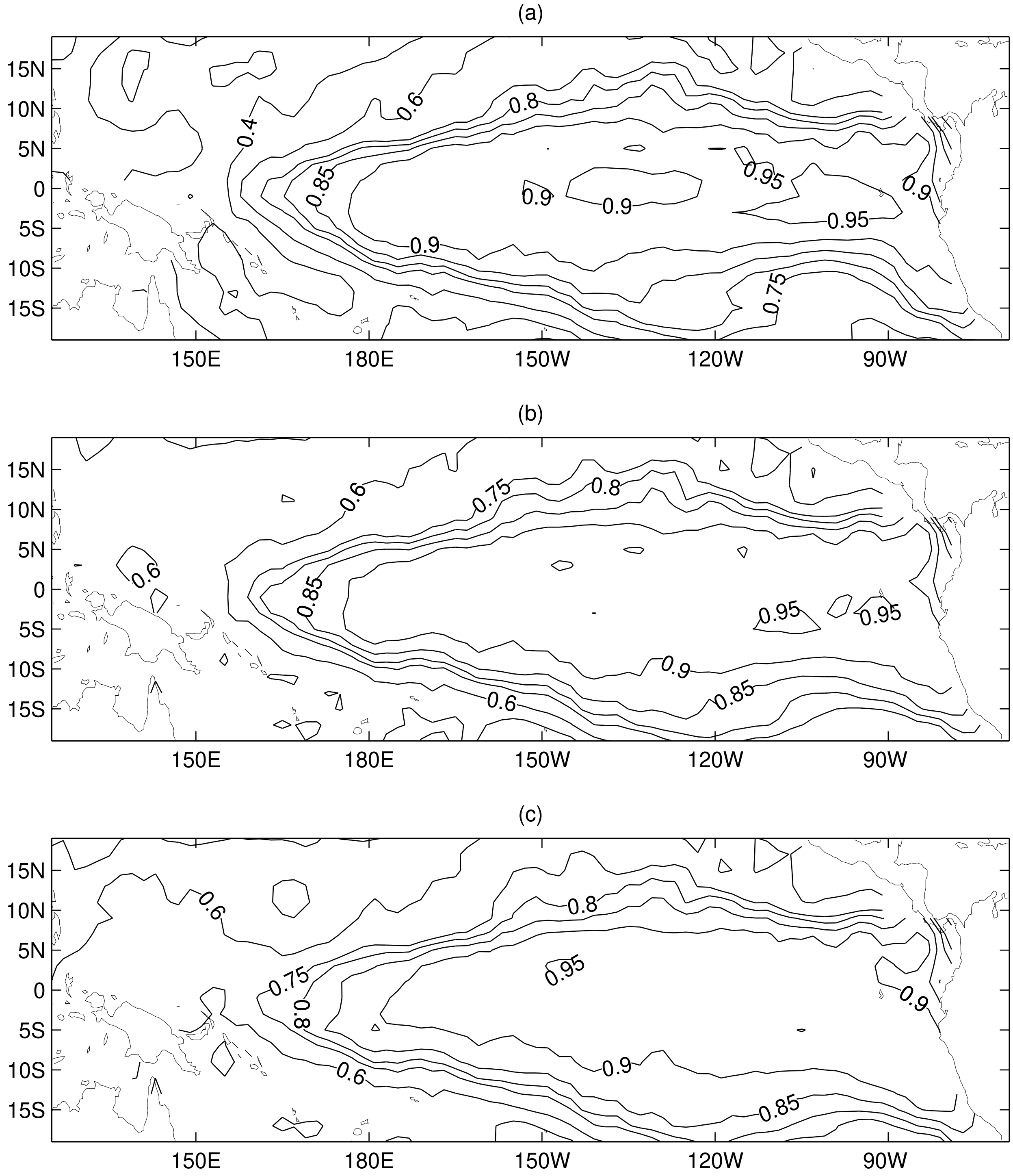


Figure 4.14: Maps of pointwise correlation between observed SSTA and (a) 2D PCA approximation (b) 2D nonmodal NLPCA approximation, and (c) 2D modal NLPCA approximation.

NLPCA algorithm is a matter of chance. This degeneracy complicates the interpretation of the time series determined by nonmodal NLPCA. In particular, the time series $\beta_i(t_n)$ need not be independent, or even uncorrelated.

Determining a set of P independent variables γ_i parameterising the surface from the set of P time series $\beta_i(t_n)$ determined empirically by NLPCA is another problem of feature extraction, in the space of the variables parameterising the surface. Therefore, PCA or modal NLPCA can be used to calculate the $\gamma_i(t_n)$. In the case at hand, inspection of the scatterplot of $\beta_1(t_n)$ with $\beta_2(t_n)$ indicated that PCA was appropriate for the separation of the correlated time series $\beta_1(t_n)$ and $\beta_2(t_n)$ into two uncorrelated time series $\gamma_1(t_n)$ and $\gamma_2(t_n)$. The homeomorphism $\mathbf{g}$ is thus simply a linear function. The first PCA mode explained 92.7% of the variance in β-space, and the associated time series (not shown) describes average ENSO variability. Its correlation coefficient with the Niño 3.4 time series is 0.92 and with $\alpha_1(t_n)$ is 0.87. The second PCA mode explains the remaining 7.3% of the variance in β-space, and the associated time series $\gamma_2(t_n)$ (not shown) is rather similar to $\alpha_2(t_n)$. The two time series have a correlation of 0.7, and in particular $\gamma_2(t_n)$ displays the same shift in activity from the pre-1977 to the post-1977 period as does $\alpha_2(t_n)$, with the same prominent peaks appearing in both time series. The parameterisation $(\beta_1(t_n), \beta_2(t_n))$ thus contains essentially the same information as the two time series $\alpha_1(t_n)$ and $\alpha_2(t_n)$. The fact that an extra step of processing is required to allow interpretation of the time series produced by nonmodal NLPCA indicates a distinct advantage of modal over nonmodal NLPCA.

Figure 4.15 displays the 2D modal approximation to the SSTA data, obtained by adding the two 1D modal approximations obtained before, projected in SSTA EOF subspaces $(\mathbf{e}_1, \mathbf{e}_2)$, $(\mathbf{e}_2, \mathbf{e}_3)$, $(\mathbf{e}_1, \mathbf{e}_3)$, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. Because of the variance-partitioning property of NLPCA, the fraction of variance explained by this approximation is the sum of the fractions explained by the two 1D modal approximations, i.e., 74.4%, which is

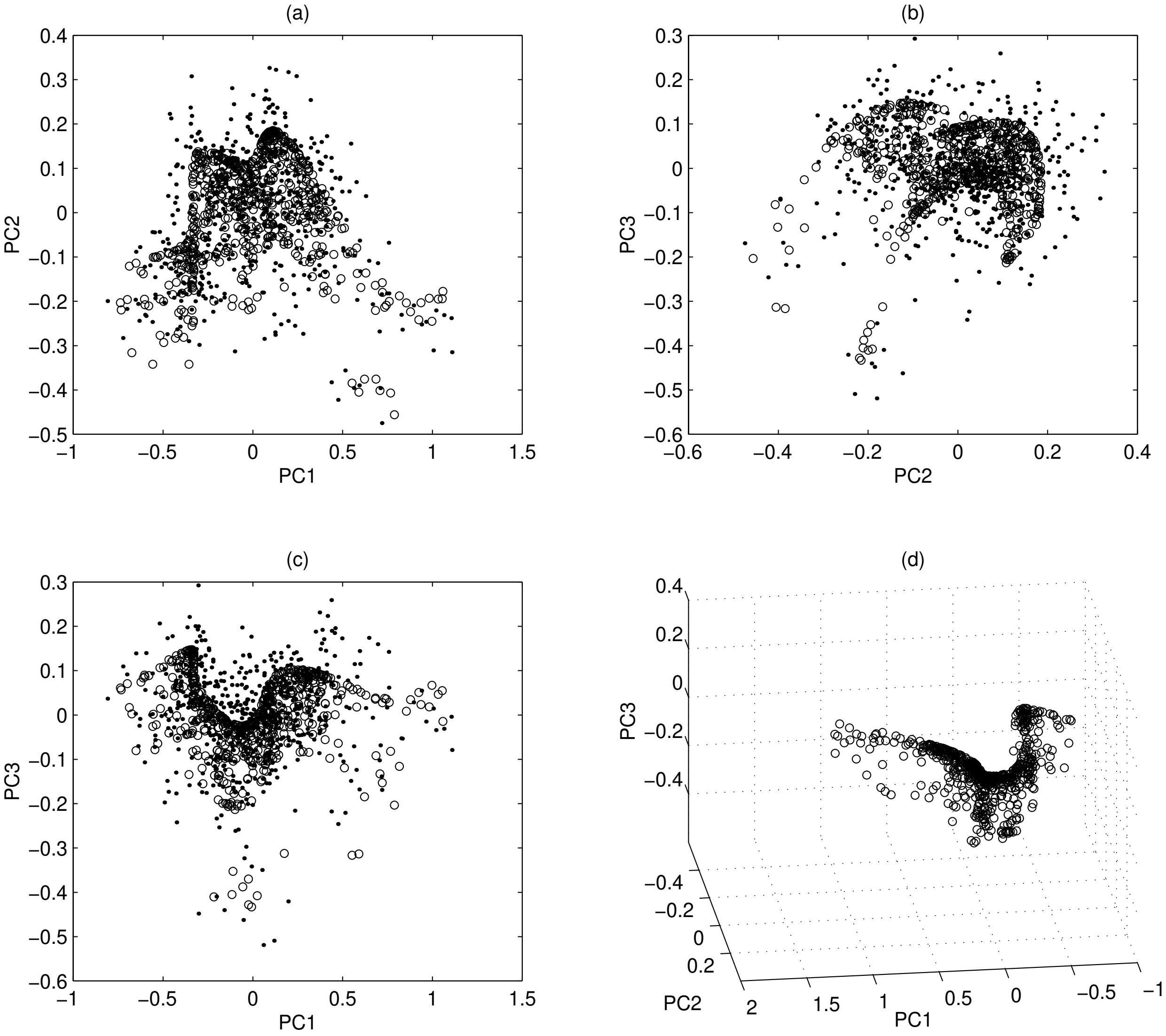


Figure 4.15: As for Figure 4.3, but for SSTA 2D modal NLPCA approximation.

slightly greater than that obtained with the 2D nonmodal approximation. This approximation differs in detail from the nonmodal approximation displayed in Figure 4.12, but the two agree broadly in their general features. Figure 4.14(c) displays a map of the pointwise correlation coefficient between the 2D modal NLPCA approximation and the original SSTA data; correlations are somewhat higher than those of the 2D nonmodal approximation in the western Pacific ocean and somewhat lower in the eastern equatorial Pacific, but by and large the differences between the two pointwise correlation maps are small.

Note that while in principle one would expect a 2D nonmodal NLPCA approximation to be superior to a 2D modal NLPCA approximation, because it should have access to a broader class of functions, in the case of SSTA the former explains 72.2% of the variance while the latter explains 74.4%. In fact, the model corresponding to the modal NLPCA had 13% more free parameters than that corresponding to the nonmodal NLPCA. It seems that this difference allowed the modal model more flexibility than the nonmodal, leading to the slight improvement in the fraction of variance explained.

Thus, in both 1D and 2D, and for both modal and nonmodal approaches, NLPCA produces approximations of greater fidelity to the tropical Pacific SSTA than does PCA. In particular, both the 1D PCA and the 1D NLPCA approximations describe "average" ENSO variability, but the 1D modal NLPCA approximation was able to characterise the spatial asymmetry between average El Niño and La Niña events in a fashion that 1D PCA cannot.

4.4 Tropical Indo-Pacific Sea Level Pressure

As was done with the SSTA data, the SLPA data was preprocessed by projecting it onto the space spanned by its 10 leading EOF modes, which together explain 60% of the total variance in the data. Figure 4.16 displays maps of the three leading SLPA EOF modes, which explain 24.2%, 10.7%, and 6.0% of the total variance, respectively.

Figure 4.17 displays the 1D NLPCA approximation of the SLPA data projected on the subspaces spanned by SLPA EOFs $(\mathbf{e}_1, \mathbf{e}_2)$, $(\mathbf{e}_2, \mathbf{e}_3)$, $(\mathbf{e}_1, \mathbf{e}_3)$, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. This curve was obtained using a network with $L = 2$ neurons in the encoding and decoding layers; the NMSD between the 8 candidate models from an ensemble of 50 ranged between 0.1% and 0.3%. The 1D NLPCA approximation explains 27.0% of the total variance in the SLPA data, a slight improvement over the fraction of variance explained by the 1D PCA approximation. Figures 4.18(a) and (b) display respectively a plot of $\alpha_1(t_n)$, the time series associated with the 1D NLPCA approximation, and the Southern Oscillation Index (SOI), calculated by subtracting the SLPA at Darwin, Australia (131E,12S) from that at Tahiti (149W,17S) (see Figure 4.16(a)), and then applying a 5-month running average smoother. The two time series bear a strong resemblance to each other on interannual and longer timescales; their correlation coefficient is 0.72. The 1D modal NLPCA approximation thus seems to describe ENSO variability in the SLPA field. This association is reinforced by inspection of the sequence of maps $\hat{\mathbf{X}}^{(1)}$ for $\alpha_1 = -3, -2, -1, -0.5, 0, 0.5, 1, 2$ (Figure 4.19). This sequence of spatial patterns describes the east-west SLPA dipole associated with average Southern Oscillation variability. Figure 4.20 displays composites of SLPA averaged over those DJF (December, January, February) periods in which the SOI was less than 1 standard deviation below the long-term average (Figure 4.20(a)) or was more than one standard deviation above the long-term average (Figure 4.20(b)). This 3-month period was selected because of all 3-month seasons in the record, it displayed the greatest variance. Figures 4.20(a) and 4.20(b) represent "average" El Niño and La Niña patterns, respectively. Clearly, the maps in Figure 4.19 for $\alpha_1 < 0$ correspond to the El Niño composite and those for $\alpha_1 > 0$ correspond to the La Niña composite. Comparison of Figures (4.19) and (4.20) indicates that the 1D NLPCA approximation characterises

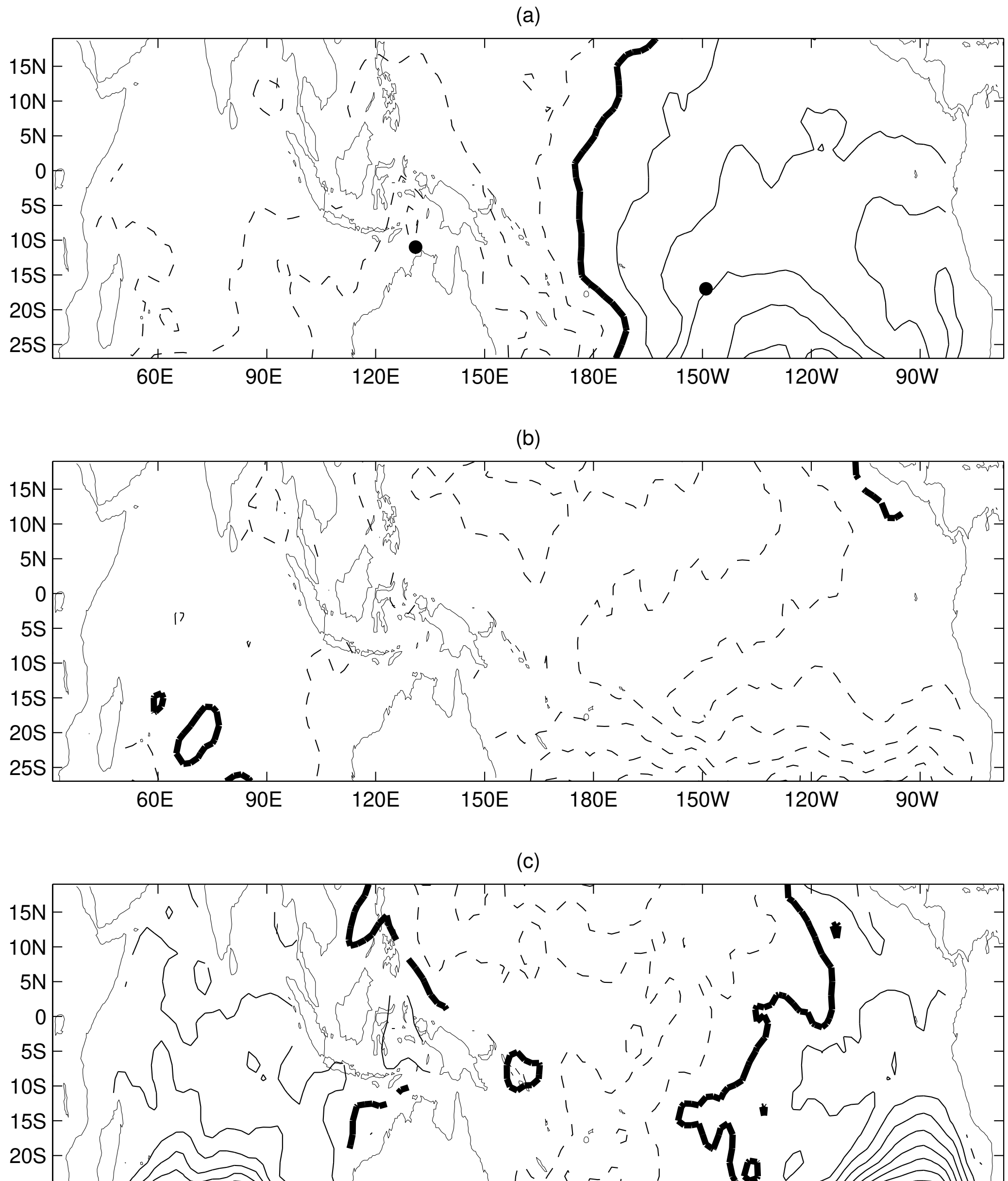


Figure 4.16: Spatial patterns of SLPA (a) EOF mode 1, (b) EOF mode 2, (c) EOF mode 3. The black dots in (a) designate the positions of Tahiti and Darwin, Australia.

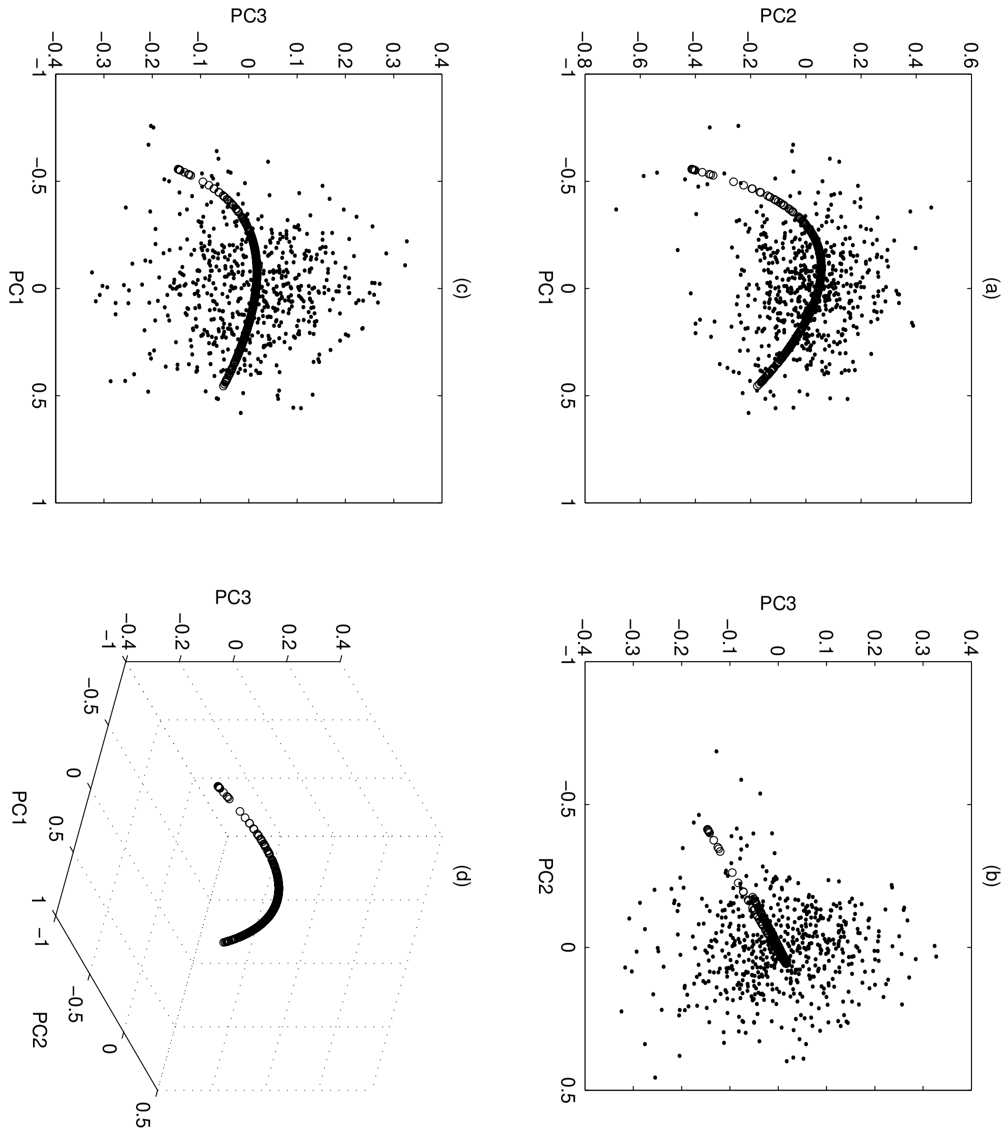


Figure 4.17: As for Figure 4.3, but for SLPA NLPCA mode 1.

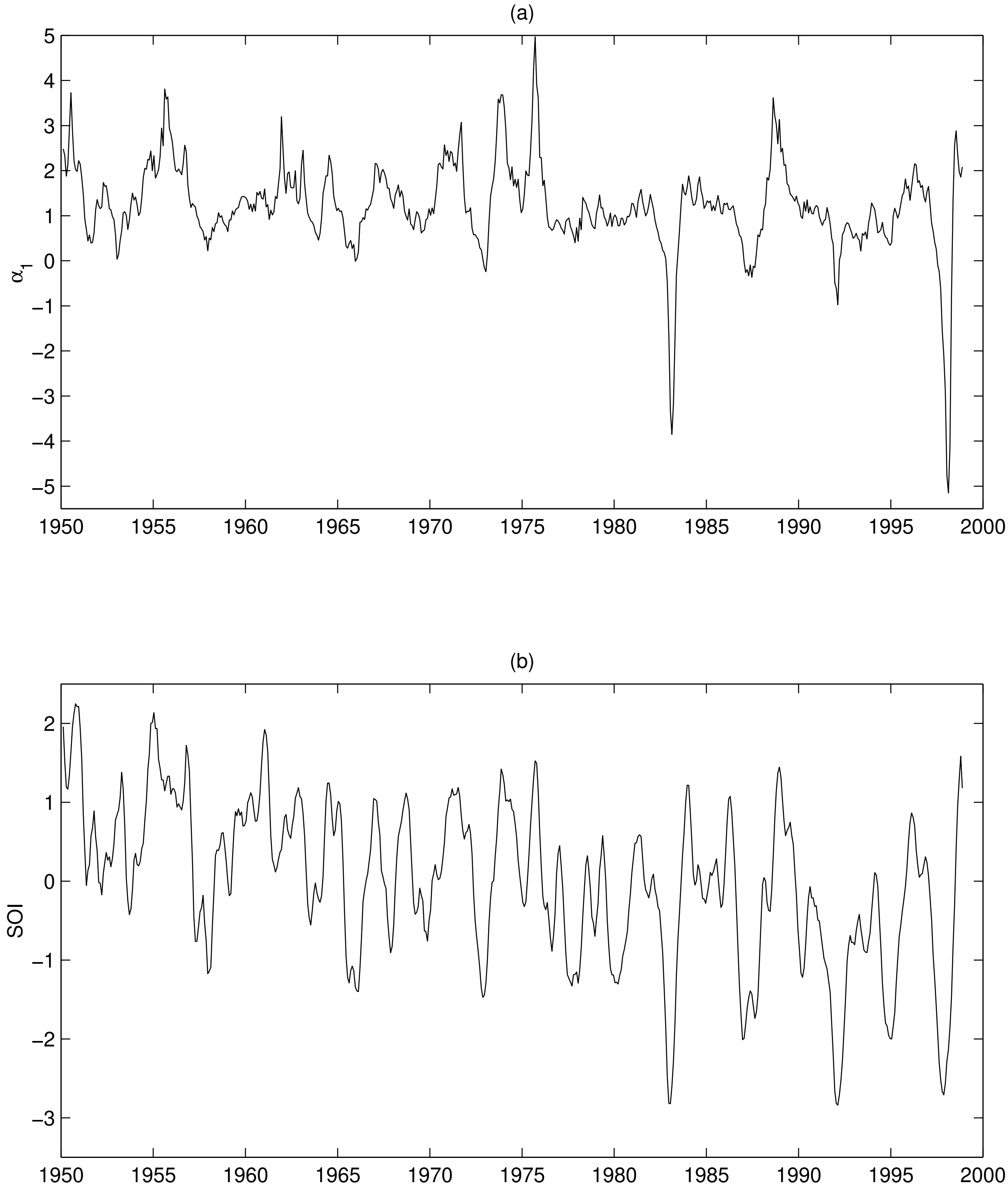


Figure 4.18: (a) Plot of $\alpha_1(t_n) = s_{f_1}(\mathbf{X}(t_n))$, the standardised time series associated with SLPA NLPCA mode 1. (b) Plot of 5-month running mean of SOI, standardised to unit variance.

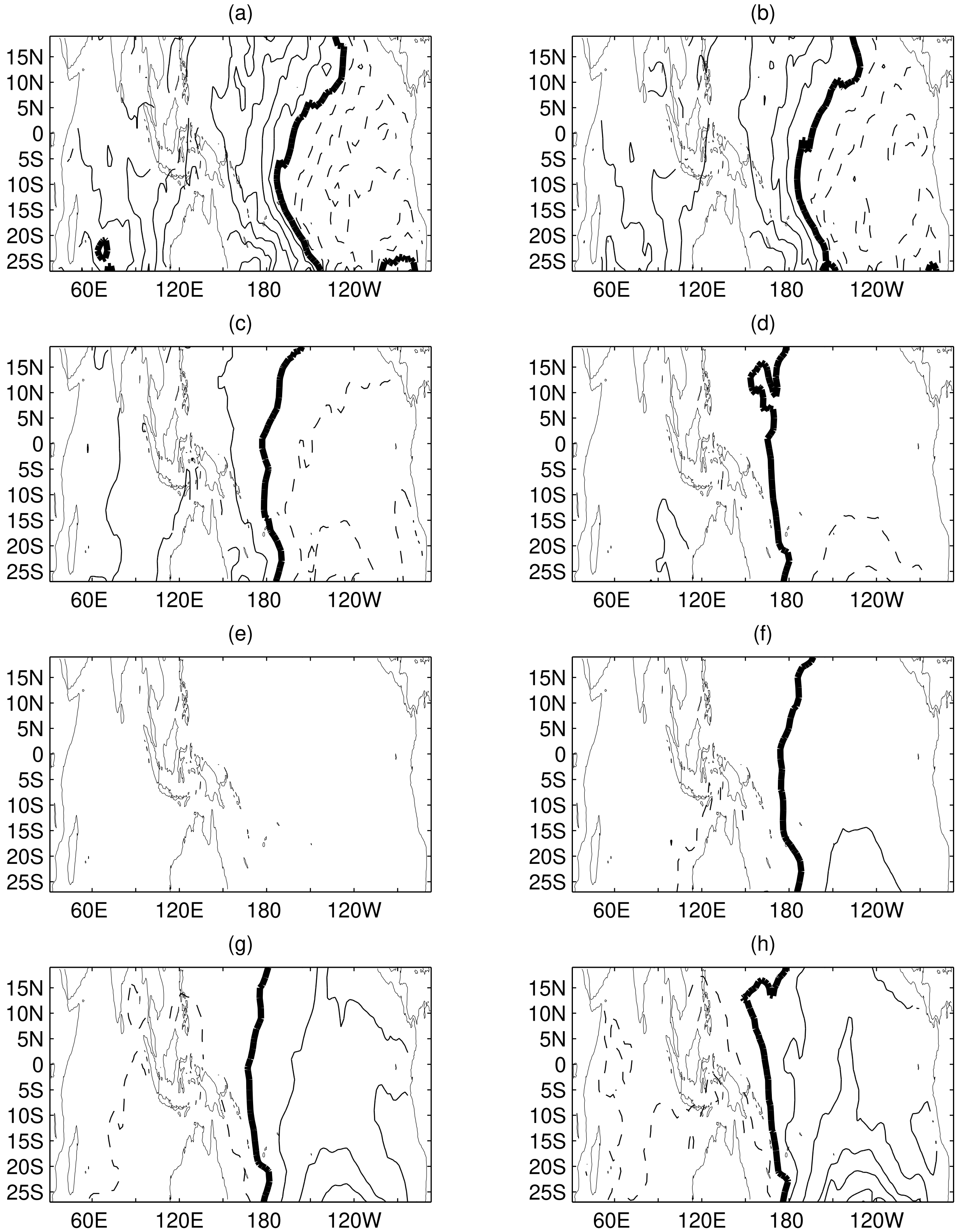


Figure 4.19: Plot of a sequence of spatial maps characterising SLPA NLPCA mode 1 for (a) $\alpha_1 = -3$ (b) $\alpha_1 = -2$ (c)$\alpha_1 = -1$ (d) $\alpha_1 = -0.5$ (e) $\alpha_1 = 0$ (f) $\alpha_1 = 0.5$ (g) $\alpha_1 = 1$ (h) $\alpha_1 = 2$. Contour interval: 0.5 hPa.

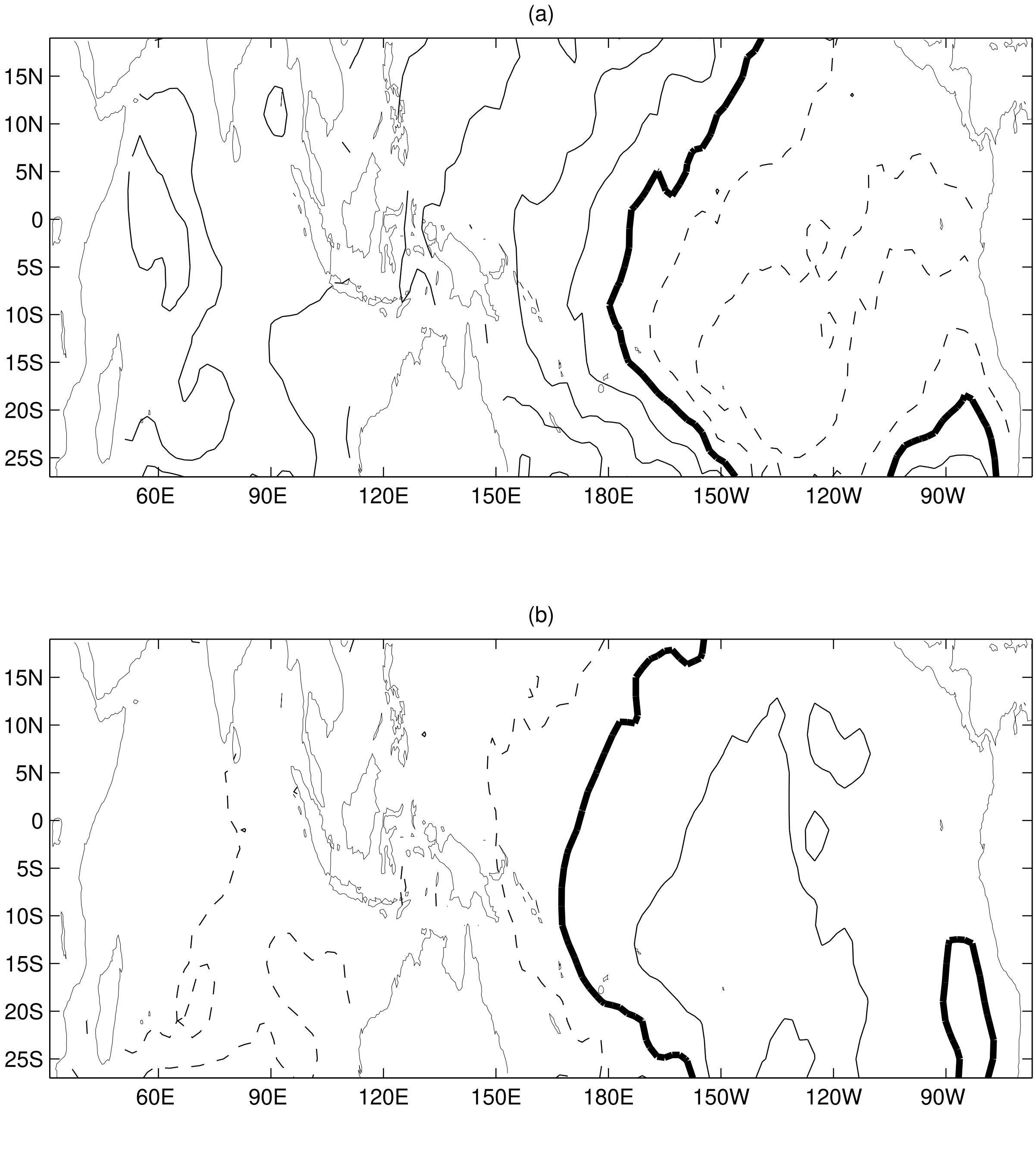


Figure 4.20: Composites of SLPA during (a) El Niño and (b) La Niña. Contour Interval: 0.5 hPa.

the asymmetry in SLPA pattern between average El Niño and La Niña events, particularly in the eastern half of the domain. Figure 4.21 displays the spatial structure of the pointwise correlation coefficient between SLPA and the 1D NLPCA approximation (Figure 4.21(a)), and the 1D PCA approximation (Figure 4.21(b)). As was the case in the previous section, the 1D SLPA NLPCA approximation produces higher correlations than the 1D PCA approximation particularly around nodal lines of the latter.

In calculating the second mode of the modal NLPCA decomposition of the SLPA data, it was determined that only for $L = 1$ neuron in the encoding and decoding layers could robust results be obtained. Neural-network based NLPCA can only find nonlinear structure if there are two or more neurons in the encoding and decoding layers. Thus, the optimal 1D characterisation of the residual data, obtained by subtracting from the original SLPA data the 1D NLPCA approximation, is a straight line. The 1D NLPCA approximation of the residual data coincides with the 1D PCA approximation of these data, and hence no lower-dimensional nonlinear structures can be found in these residuals. Our calculation (not shown) shows that NLPCA mode 2 explains 15.9% of the variance of the original data. The spatial pattern and associated time series are shown in Figure 4.22. In fact, SLPA NLPCA mode 2 bears a strong resemblance to SLPA PCA mode 2; the correlation coefficient between the two time series is 0.96 and the pattern correlation between the associated spatial patterns is 0.93. The similarity between the two is not surprising, as SLPA NLPCA mode 1 does not differ substantially from SLPA PCA mode 1. The results of a 2D nonmodal NLPCA of the SLPA data (not shown) did not yield particularly interesting results.

Thus, apart from a weakly nonlinear 1D NLPCA approximation corresponding to average ENSO variability and characterising a slight asymmetry between average El Niño and La Niña events, the robust low-dimensional structure of the SLPA data is linear.

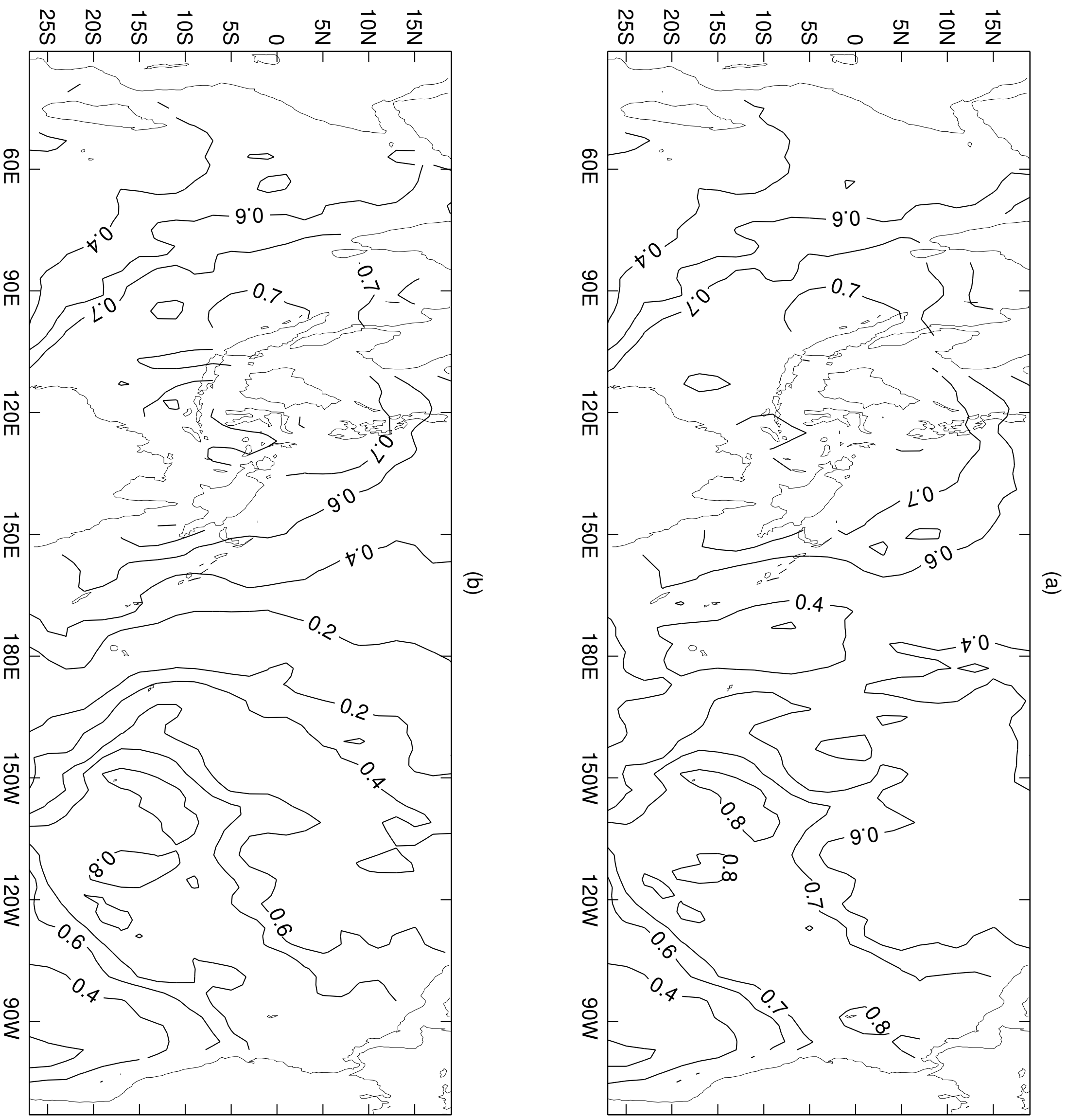


Figure 4.21: Spatial pattern of pointwise correlation coefficient between SLPA and (a) 1D NLPCA approximation and (b) 1D PCA approximation.

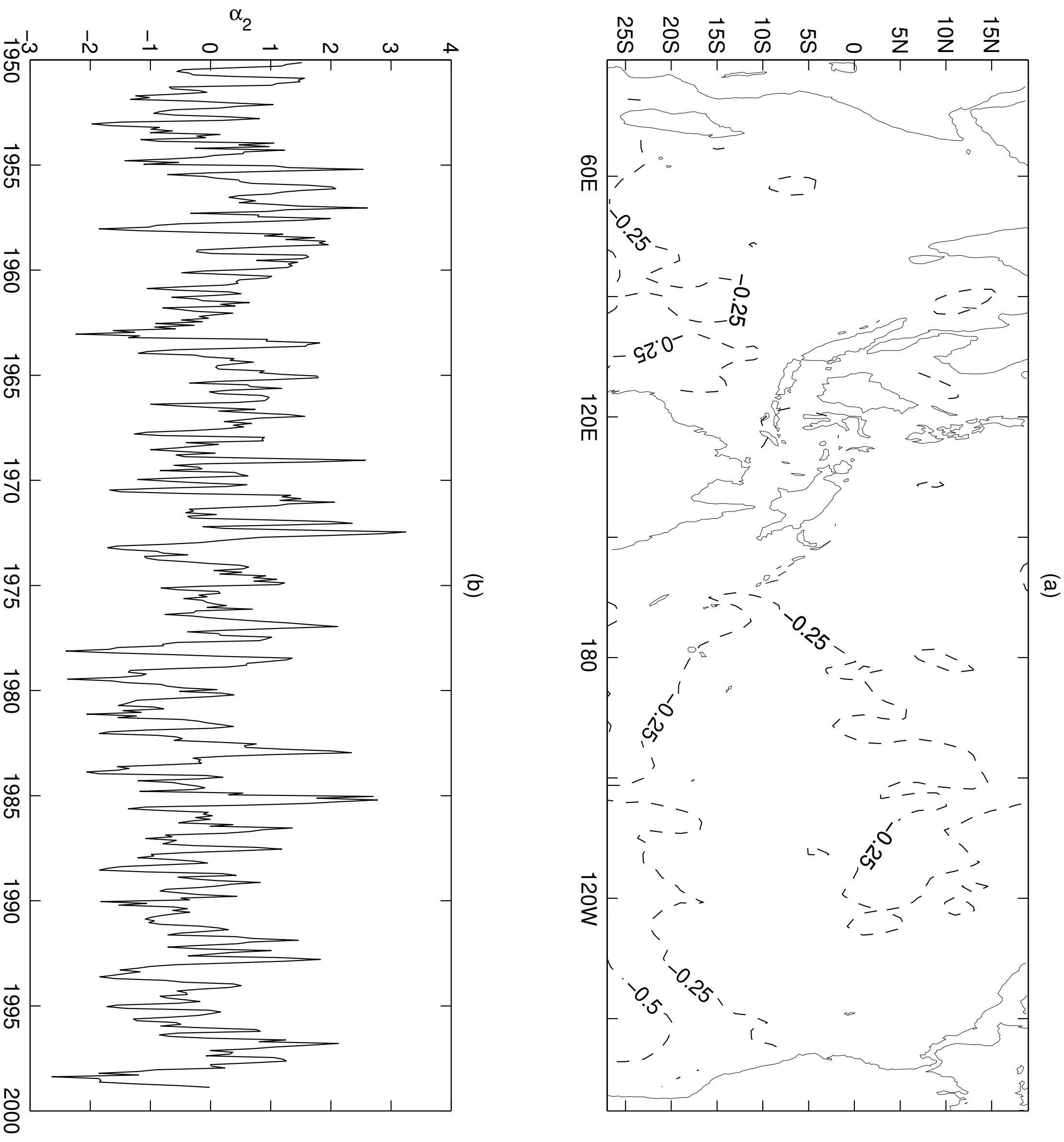


Figure 4.22: SLPA NLPCA mode 2 (a) Spatial pattern (*not* normalised, units are hPa) and (b) time series (normalised to unit variance).

4.5 Conclusions

Application of NLPCA to two data sets of climatic significance, namely tropical Pacific sea surface temperatures and tropical Indo-Pacific sea level pressure, has demonstrated that NLPCA is able to robustly produce one- and two-dimensional approximations that are superior to the corresponding approximations produced by PCA. The improvement is particularly striking in the case of SST: variability in this field is dominated by ENSO, the average manifestation of which is asymmetric between average El Niño and La Niña phases. As PCA is constrained to produce 1D approximations which are standing oscillations with fixed spatial pattern, it is unable to characterise this asymmetry. On the other hand, the 1D NLPCA approximation, by mixing PCA modes 1 and 2, is able to capture this difference in SST structure between average El Niño and La Niña episodes.

Figures 4.12 and 4.15 display, respectively, 2D nonmodal and modal NLPCA approximations to the SST data. These are seen to be rather similar in structure. Inspection of the time series $\beta_1(t_n)$ and $\beta_2(t_n)$ parameterising the 2D nonmodal NLPCA approximation highlights the importance of the fact, pointed out by Malthouse (1998), that NLPCA produces time series that are unique only up to an arbitrary homeomorphism. The time series $\beta_1(t_n)$ and $\beta_2(t_n)$ were strongly correlated, and therefore contain substantial overlap in the information they convey. A second feature analysis problem, solved using traditional PCA, was then used to untangle $\beta_1(t_n)$ and $\beta_2(t_n)$. The resulting time series bore strong similarities to the time series $\alpha_1(t_n)$ and $\alpha_2(t_n)$ corresponding to the first two modes of the *modal* NLPC analysis. This result further strengthens the interpretation that the modal and nonmodal analyses are producing essentially the same approximation. The fact that the use of nonmodal NLPCA required the solution of a subsidiary feature extraction problem to interpret the time series produced illustrates a deficiency of nonmodal NLPCA, as compared to modal.

The first modal NLPCA approximation to the SLP data describes average ENSO variability in this field, producing a somewhat better approximation to the original data than that produced by PCA and characterising the asymmetry in SLPA between average El Niño and La Niña episodes. The differences between the NLPCA and PCA approximations for SLP are less striking than was the case with SST, indicating that the low-dimensional structure of SLP is more linear than SST. Indeed, no nonlinear mode beyond the first could robustly be found in the data, indicating that either SLP variability in the tropical Indo-Pacific region is very nearly linear, or that any nonlinear structure is too subtle to detect within existing records.

Chapter 5

Nonlinear Principal Component Analysis of Northern Hemisphere Atmospheric Circulation Data

5.1 Introduction

Low-frequency, large-scale coherent variability in the Northern Hemisphere midlatitude circulation has been a subject of considerable interest in climate research over the last few decades. It is typically characterised in terms of spatially-fixed, temporally fluctuating anomaly patterns modifying the climatological mean circulation. Some of these patterns are zonally localised, such as the Pacific-North America (PNA) pattern (Wallace and Gutzler, 1981), a chain of alternating positive and negative geopotential height anomalies in the mid-troposphere extending from the subtropical North Pacific Ocean over North America, following a great circle route; and the North Atlantic Oscillation (NAO; van Loon and Rogers, 1978; Hurrell, 1995), a dipolar pattern with geopotential height anomalies of opposite sign over Iceland and the Azores. Other patterns are more zonal in structure, such as the Arctic Oscillation (AO; Thompson and Wallace, 1998) and the Antarctic Oscillation (AAO; Gong and Wang, 1999). These are approximately zonally-symmetric patterns of variability with anomalies of opposite sign over the polar region and the midlatitudes, for the Northern and Southern Hemisphere, respectively. The connections between these different patterns of variability (e.g. Deser, 1999), and their dynamical origin - in particular their maintenance by lower boundary forcing or internal dynamics (see, e.g., Feldstein and Lee, 1996; Corti et al, 1997; Trenberth et al, 1998)

- are areas of active research. These characteristic patterns of atmospheric variability have historically been diagnosed using correlation analysis (Wallace and Gutzler, 1981; Hsu and Lin, 1992), PCA (Kushnir and Wallace, 1989; Thompson and Wallace, 1998), combined PCA (Baldwin and Dunkerton, 1999; CPCA, also known as extended EOF analysis, is defined by Bretherton et al., 1992), canonical correlation analysis (Perlwitz and Graf, 1995), and rotated PCA (Barnston and Livezey, 1987). All of these methods are linear and produce spatial and temporal patterns that describe standing oscillations.

Recently, the Arctic Oscillation (AO) has been of particular interest. The AO was defined by Thompson and Wallace (1998) as the leading PCA mode of monthly averaged November through April Northern Hemisphere SLPA north of 20^{O}. The AO spatial pattern, derived from the Trenberth and Paolino SLP data set (1980), is displayed in Figure 5.1. The canonical AO structure is roughly zonally-symmetric, with anomalies of opposite sign in the polar region and the midlatitudes. Deviations from zonal symmetry are characterised by a wavenumber two pattern reflecting land-ocean contrasts. In particular, the dipole pattern in SLP over the North Atlantic strongly resembles the surface signature of the NAO. The AO diagnosed using PCA on the Trenberth and Paolino data set strongly resembles that obtained by Thompson and Wallace (1998), who found the AO and NAO time series to be highly correlated ($r = 0.69$). This point was considered further by Deser (1999), who considered the extent to which coherent variability in Northern Hemisphere circulation is really hemispheric in extent, and suggested that in fact the AO should be termed the "Arctic-Atlantic Oscillation".

The AO is strongly equivalent barotropic in structure, with coherent variability throughout the troposphere and into the lower stratosphere (Perlwitz and Graf, 1995; Thompson and Wallace, 1998, 1999a; Baldwin and Dunkerton, 1999). Interest in the AO has been particularly strong in recent years because of its potential as a sensitive

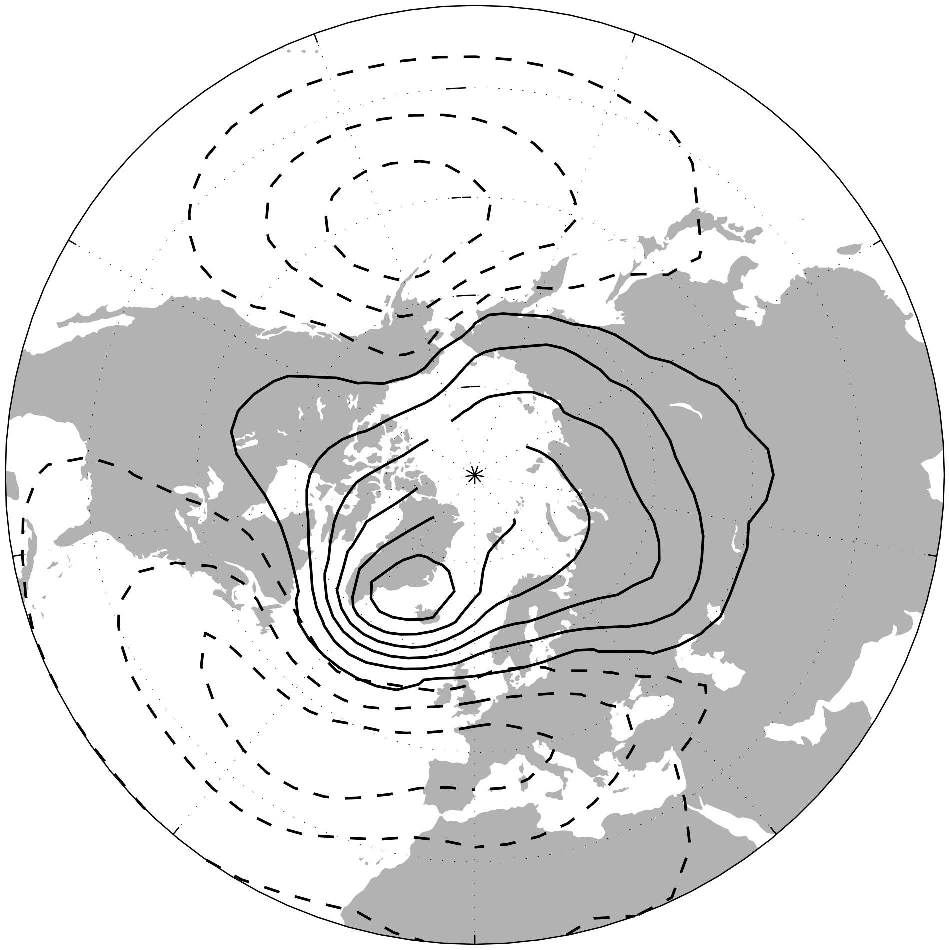


Figure 5.1: Spatial structure of the leading EOF pattern from observed SLP. Contour intervals are 1 hPa (...,-1.5,-0.5,0.5,1.5,...).

barometer of climate change. Thompson et al. (1999b) noted the similarity in structure between the spatial pattern of the AO and of recent trends in SLP in the Northern Hemisphere. They suggested that the climate change signal is manifesting itself as a secular shift toward the positive phase of the AO (strong polar vortex) superimposed upon monthly timescale AO fluctuations. Indeed, in modelling studies, Shindell et al. (1999, using the Goddard Institute for Space Sciences (GISS) atmospheric GCM) and Fyfe et al. (1999, using the Canadian Centre for Climate Modelling and Analysis (CCCma) coupled GCM) found a pronounced trend in the AO signal as greenhouse gases were increased. However, the physical mechanism for this behaviour remains controversial, as Shindell et al. found that only with a full stratosphere would their model produce an AO trend under increasing greenhouse forcing, while Fyfe et al. were able to produce this result using an atmospheric model with a poorly-resolved stratosphere.

In this chapter, the results of an NLPC analysis of fields characterising the Northern Hemisphere circulation will be considered, using data from the CCCma coupled GCM. The results will demonstrate that NLPCA is able to detect and characterise regime behaviour in multivariate data sets. Consideration of these regimes will throw some light on the controversies concerning the relationship between the AO and the NAO discussed above.

5.2 Data and Model Building

The data analysed in this chapter comes from two integrations of the CCCma coupled climate model (CGCM1): a 1001-year control integration with atmospheric carbon dioxide (CO_2) at pre-industrial levels and a 500-year stabilisation integration with atmospheric CO_2 concentrations at four times pre-industrial level, corresponding to predicted CO_2 levels in 2100. Use of model data rather than observations allows the investigation of potential changes in the atmospheric variability associated with increased atmospheric CO_2 concentrations. The fields considered are monthly-averaged Northern Hemisphere sea level pressure and 500mb geopotential height (Z_{500}) from 20^ON to 90^ON, over the extended winter period from November though April. Both fields are on a Gaussian grid with 3.75^O resolution in the zonal and meridional directions. The CCCma CGCM1 model and its climate are described in Flato et al. (1999). The atmospheric component is a T32 spectral primitive equation model with 9 unequally spaced levels (McFarlane et al., 1992). The ocean component is a global primitive equation grid-point model with 1.875^O resolution and 29 vertical levels. It is based on the Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model (MOM) 1.1 (Pacanowski et al., 1993). The leading EOF of this coupled model displays spatially and temporally realistic AO behaviour (Fyfe et al., 1999).

For both the SLP and Z_{500} fields, monthly anomalies (SLPA and Z_{500}A, respectively) were computed by subtracting the climatological annual cycle. Fields were weighted by the square root of the cosine of the latitude before calculation of the EOFs to account for the poleward concentration of gridpoints on the Gaussian grid. NLPC analysis of SLPA and Z_{500}A data was carried out using the early stopping algorithm described in Chapter 2. In all analyses, 20% of the data was held aside in a validation set, for which network performance was monitored as training proceeded. Training was stopped when this performance began to degrade, or after 20000 iterations, whichever came first. It was found that increasing the maximum number of iterations beyond 20000 did not affect the results of the analysis.

5.3 Analysis of GCM Sea Level Pressure

As was done with the tropical SLPA and SSTA data in the previous chapter, the northern hemisphere SLPA and Z_{500}A were projected onto the spaces of their first 10 EOF modes, in which 85% and 76% of the variance are respectively contained. The first four EOF patterns of SLPA and Z_{500}A are displayed in Figures 5.2 and 5.3, respectively. The map displayed in Figure 5.2(a) is the spatial pattern of the canonical Arctic Oscillation in the CCCma model (Fyfe et al., 1999).

Figure 5.4 displays a scatterplot of the SLPA data projected in the plane spanned by the leading two EOFs, overlaid with a histogram-based estimate of the marginal probability density function (PDF) of these data in this space. The PDF displays a marked deviation from Gaussian structure in the form of a pronounced lobe in the lower-right quadrant. Figure 5.5 displays the estimate of the PDF along with the 1D NLPCA approximation $\hat{\mathbf{X}}(t_n)$ to these data. The NLPCA approximation was found using a network with $L = 2$ neurons in the encoding and decoding layers. This approximation was obtained from an ensemble of 5 networks, of which 3 were candidate models differing from each other with an NMSD of at most 1%. The approximation, which explains 26.5% of the total variance (in contrast to 24% explained by the leading PCA approximation), is thus robust. Note that $\hat{\mathbf{X}}(t_n)$ has a piecewise-linear structure, composed of three branches. The associated standardised time series

$$\alpha(t_n) = \frac{s_f(\mathbf{X}(t_n)) - < s_f >}{(< (s_f - < s_f >)^2 >)^{1/2}}, \tag{5.1}$$

along with a histogram estimate of the PDF of $\alpha(t_n)$, are displayed in Figure 5.6. The distribution of $\alpha(t_n)$ is strongly bimodal, demonstrating the existence of two well-separated regimes in which the NLPCA approximation resides. Each of the three branches of $\hat{\mathbf{X}}(t_n)$ corresponds to a feature in the PDF of $\alpha(t_n)$: the upper branch (hereafter referred to as Branch 1) is associated with the larger of the two peaks, the lower branch (hereafter

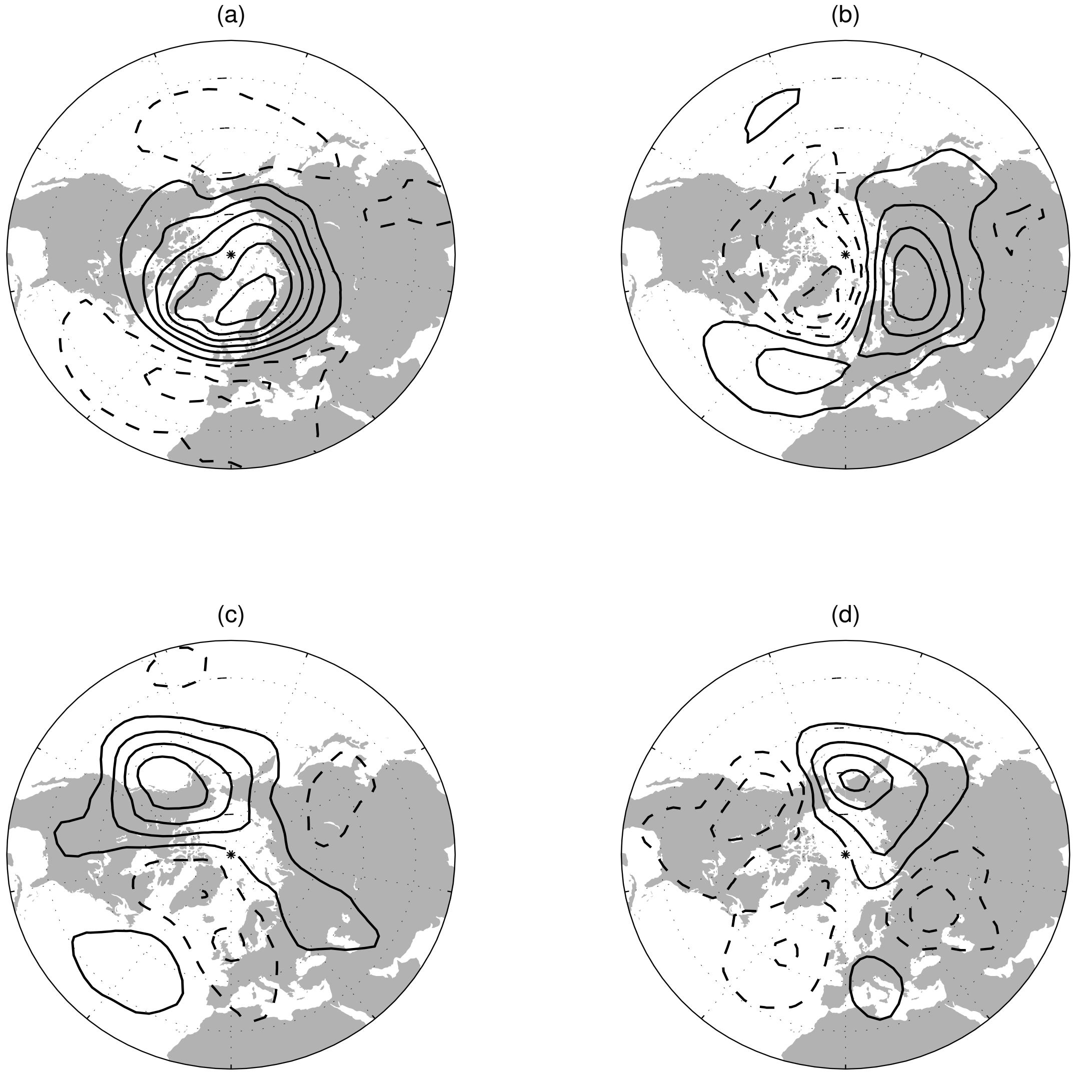


Figure 5.2: Spatial structure of the leading four EOF patterns from CCCma SLPA: (a) EOF 1, (b) EOF 2, (c) EOF 3, (d) EOF 4. These patterns explain 23.7%, 10.6%, 8.5%, and 6.5% of the variance in SLP, respectively. Negative contours are dashed. Contour intervals are 1 hPa (..., -1.5, -0.5, 0.5, ...).

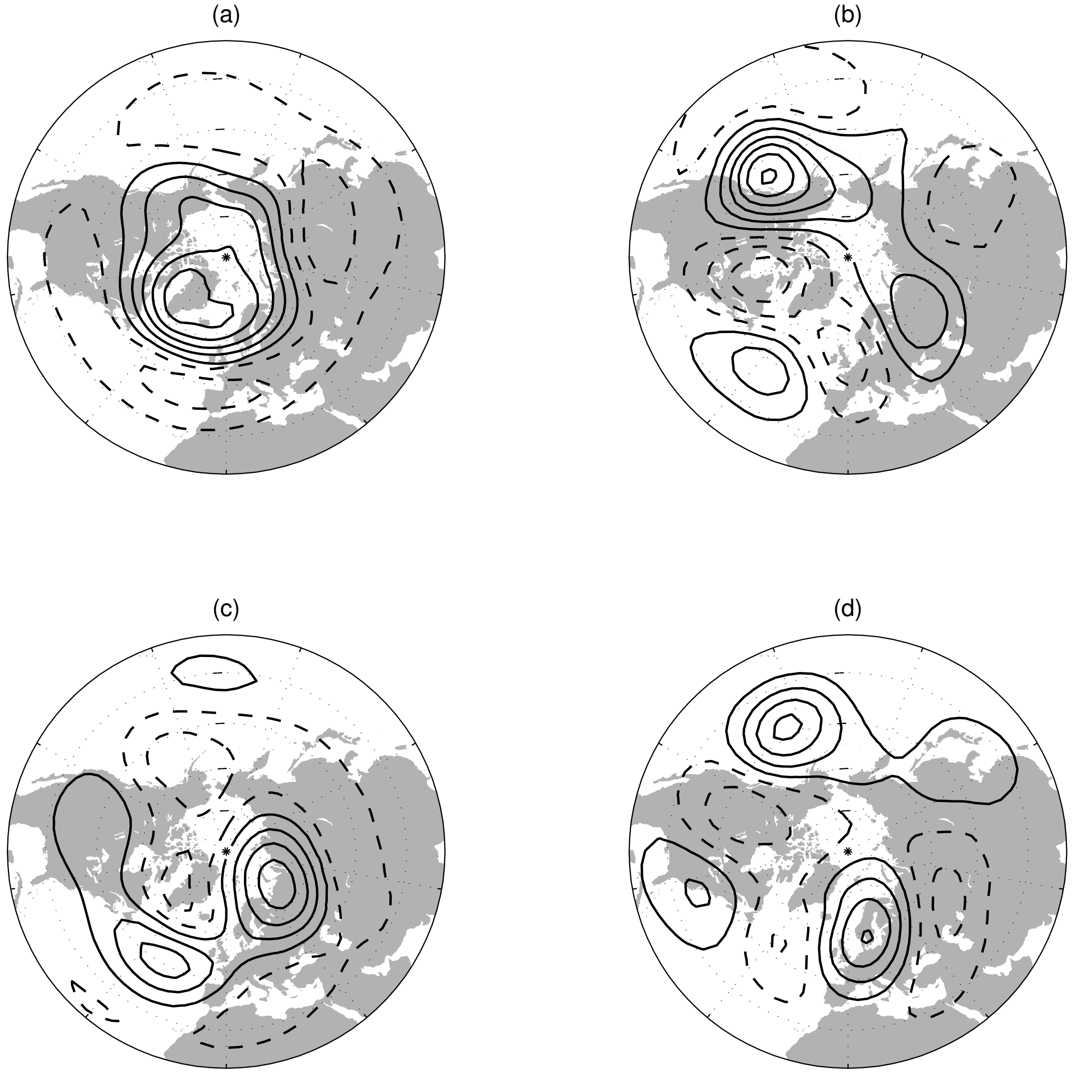


Figure 5.3: Spatial structure of the leading four EOF patterns from CCCma $Z_{500}A$: (a) EOF 1, (b) EOF 2, (c) EOF 3, (d) EOF 4. These patterns explain 19.6%, 12.5%, 9.3%, and 8.2% of the variance in Z_{500}, respectively. Negative contours are dashed. Contour intervals are 10 m (..., -15, -5, 5, ...).

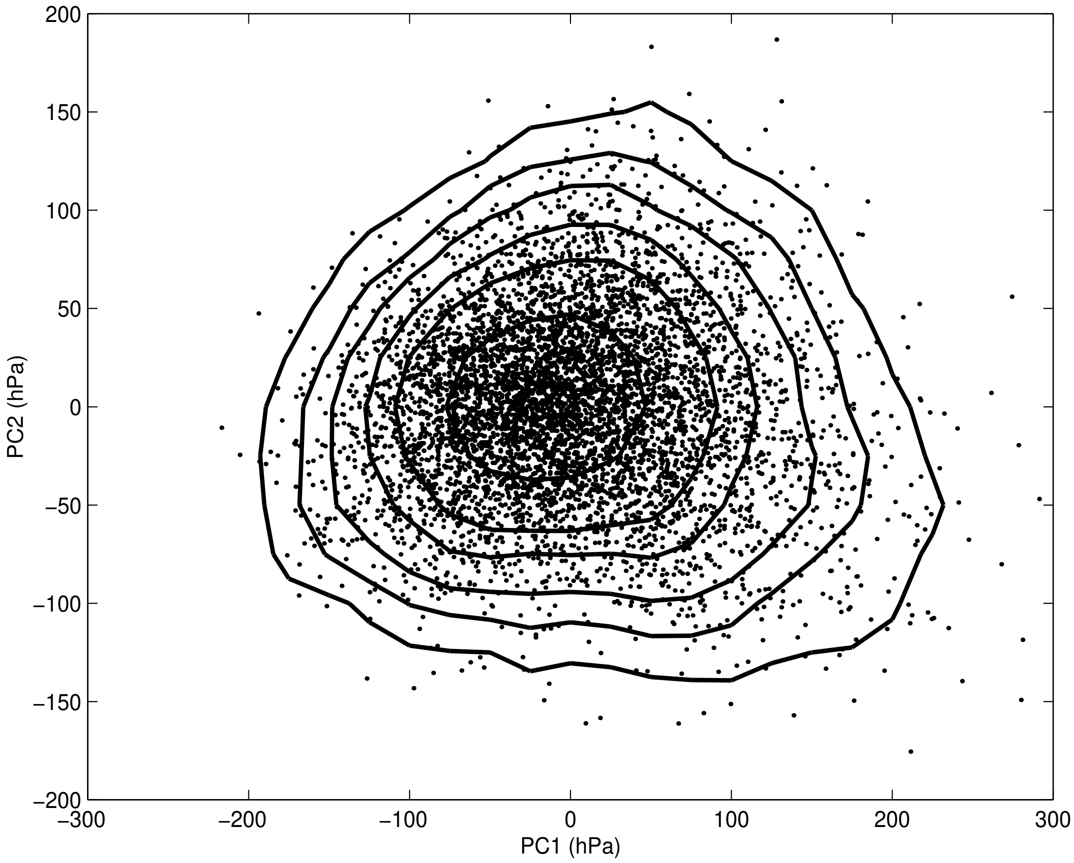


Figure 5.4: Scatterplot of the leading two SLPA PC time series, overlaid with a histogram estimate of the corresponding marginal probability density function. Contour intervals are $5 \times 10^{-4}, 1.5 \times 10^{-3}, 3 \times 10^{-3}, 6 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 3 \times 10^{-2}$. The histogram bin size is 25 hPa in both directions.

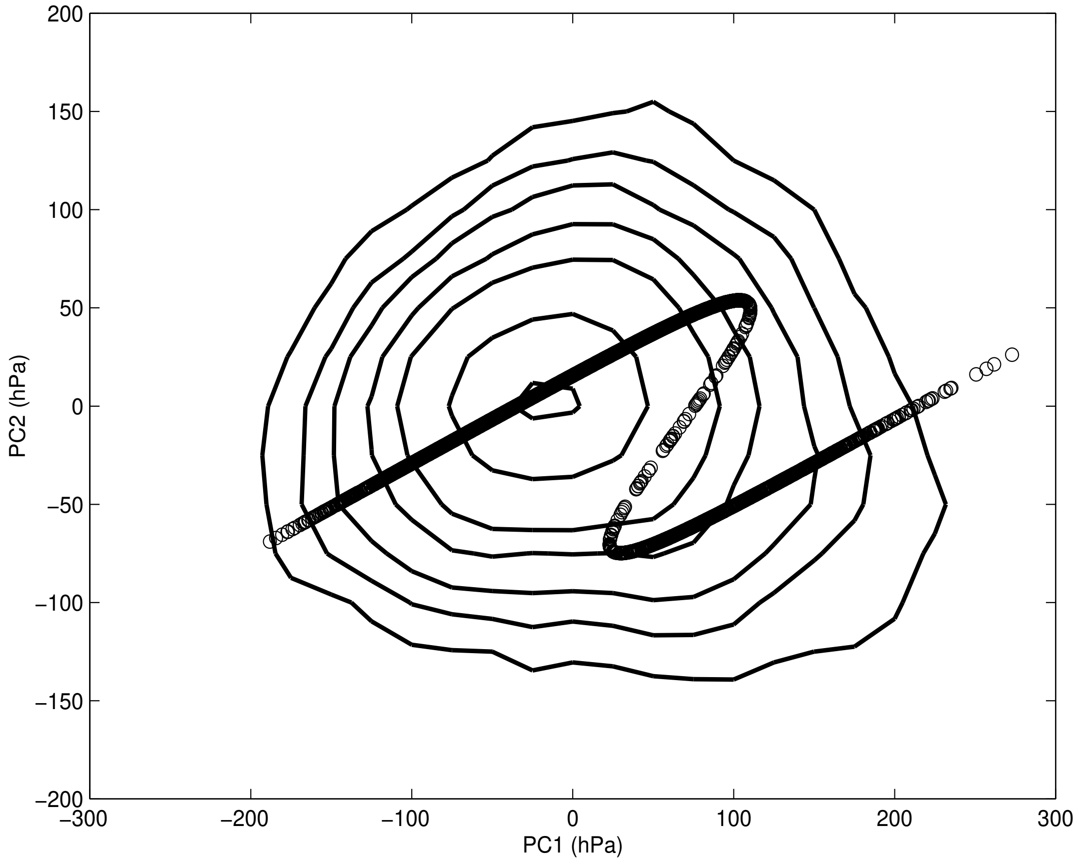


Figure 5.5: 1D NLPCA approximation $\hat{\mathbf{X}}$ of SLPA, projected in the space of the first two SLPA EOFs (open circles), overlaying histogram estimate of SLPA PDF as in Figure 5.4.

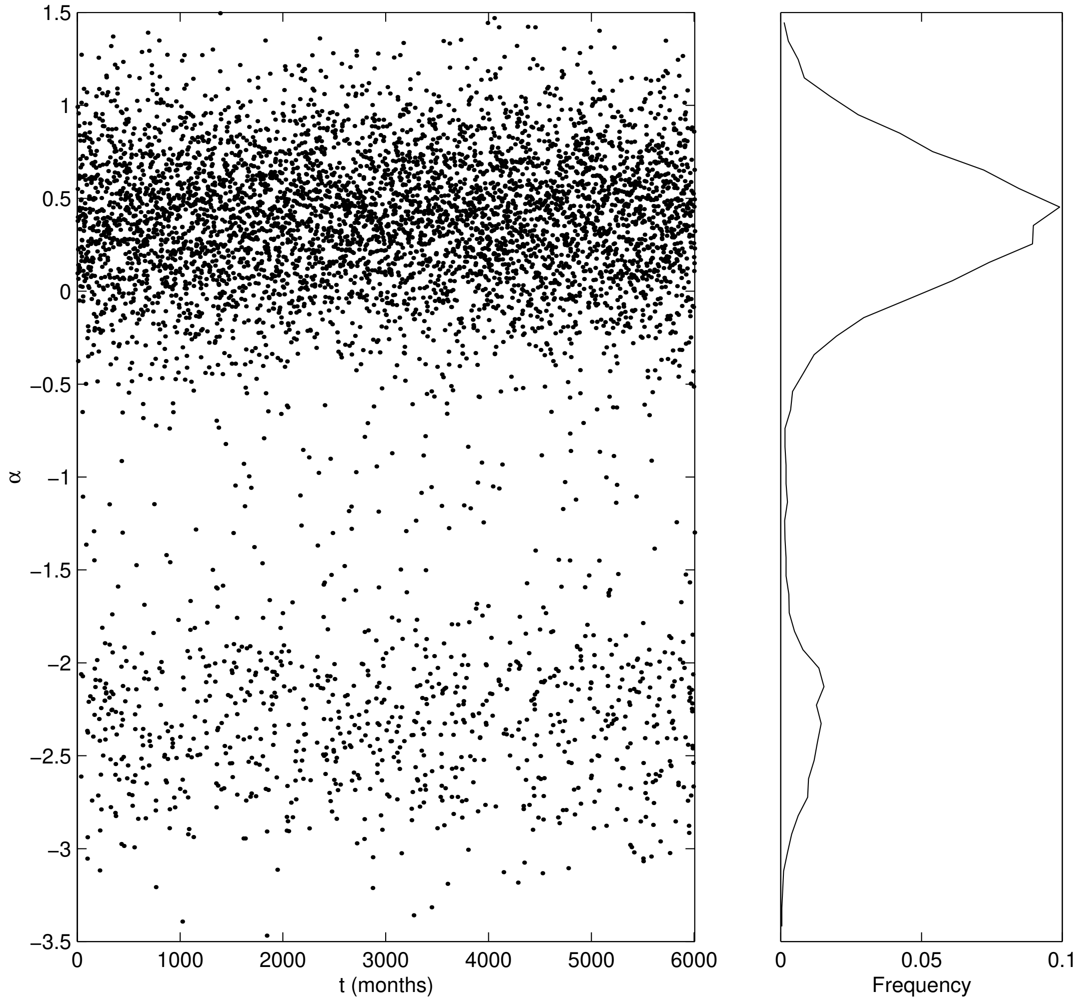


Figure 5.6: Plot of the 1D SLPA NLPCA time series $\alpha(t_n)$ (left) and the associated histogram estimate of the PDF (right).

referred to as Branch 2) is associated with the smaller peak, and the intermediate branch is associated with the minimum of the PDF of α between the two peaks, and is rarely visited. The approximation $\hat{\mathbf{X}}(t_n)$ is in Branch 1 for 84% of the months and in Branch 2 for 13%. An inspection of $\alpha(t_n)$ indicates that variability on $\hat{\mathbf{X}}(t_n)$ consists of oscillatory motion along Branch 1 with occasional episodic excursions to Branch 2, on which the approximation rarely resides longer than a month or two.

Figure 5.7 is the same as Figure 5.5 but with the estimated PDFs of the populations corresponding to the upper and lower branches plotted separately. The PDF of the population projecting onto Branch 1 is seen to be nearly Gaussian, with a major axis nearly parallel to Branch 1. Branch 2 also runs through the middle of its associated PDF. The overlap of the two PDFs is an artifact of the coarse binning used in their estimation.

As was discussed in the previous Chapter, NLPCA differs from PCA in that the 1D approximation produced by NLPCA does not correspond to a unique spatial pattern, but in fact to a sequence of maps. In Chapter 4, these maps were presented at representative points along the curve $\hat{\mathbf{X}}(t_n)$. In this chapter, a somewhat different methodology for producing maps corresponding to the NLPCA approximation is adopted. Because the features of interest involve atmospheric circulation data at different altitudes, based on an approximation calculated using data at a single altitude, instead representative points along the approximation $\hat{\mathbf{X}}(t_n)$ are selected and composite the original data over all times that the approximation resides in the neighbourhoods of these points. A comparison of the two methods for the field from which the NLPCA approximation was derived demonstrates that the resulting maps are essentially identical, as expected since the NLPCA approximation is constrained to run through the "middle" of the data.

Figure 5.8 displays composites of SLPA over neighbourhoods in α of width 0.4, centred at intervals of 0.4 from -3.1 to 1.3. Variability along Branch 1 is exemplified by

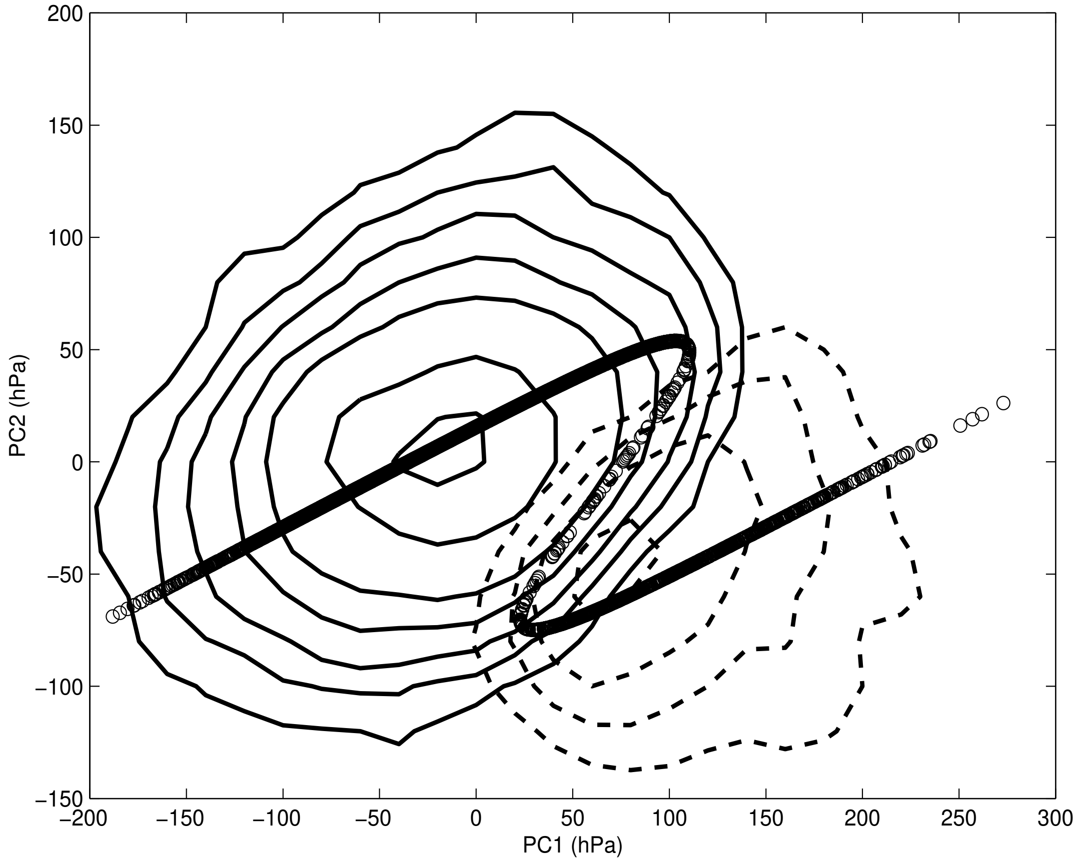


Figure 5.7: As in Figure 5.5, but with the PDFs of the populations corresponding to Branch 1 (solid contours) and Branch 2 (dashed contours) plotted separately, and with a bin size of 20 hPa.

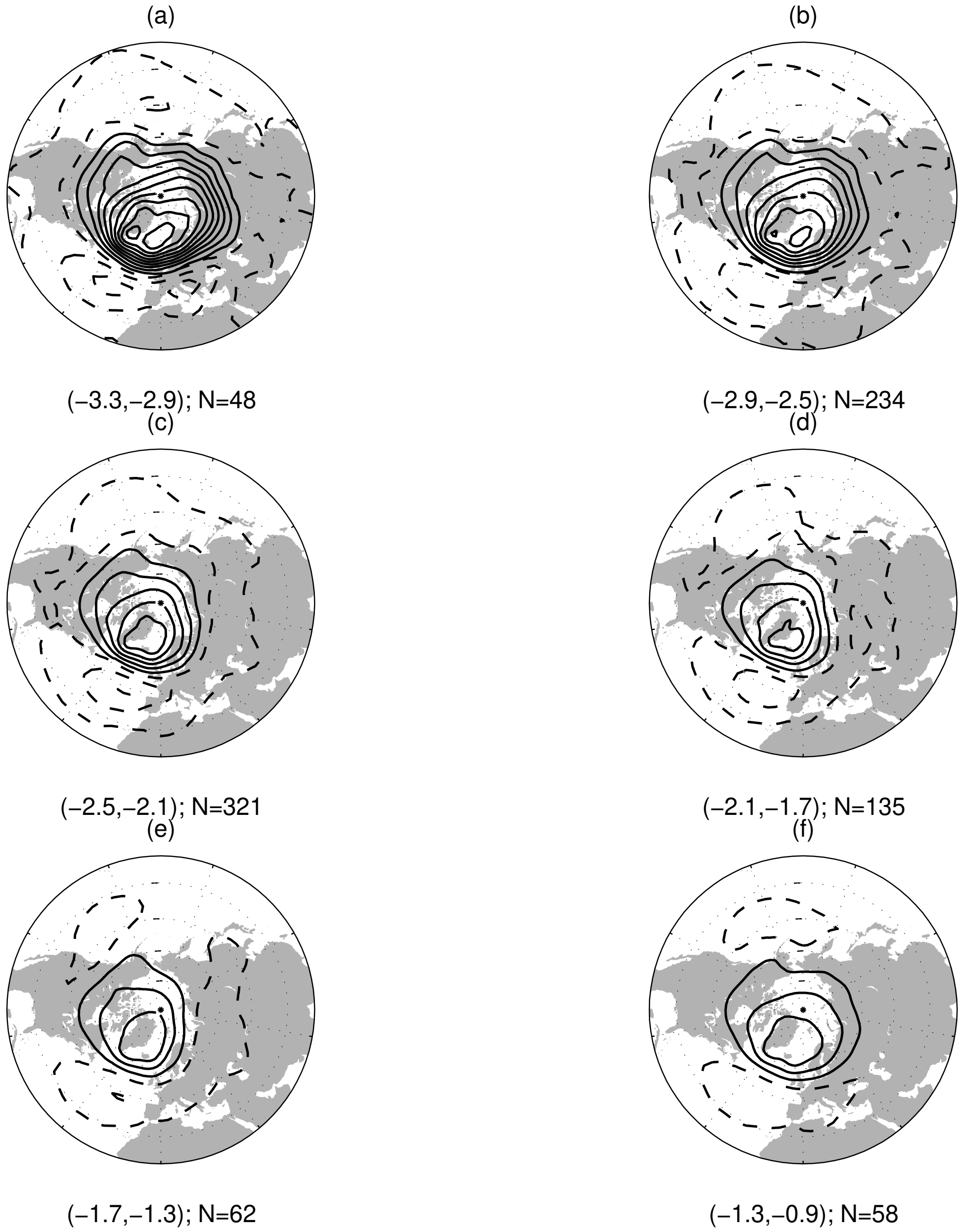


Figure 5.8: Composites of SLPA over characteristic ranges of α. These ranges are indicated in parentheses below the maps, along with the number N of maps used in the composite. Contour interval is 2 hPa (...,-3,-1,1,3,...). Continued on next page.

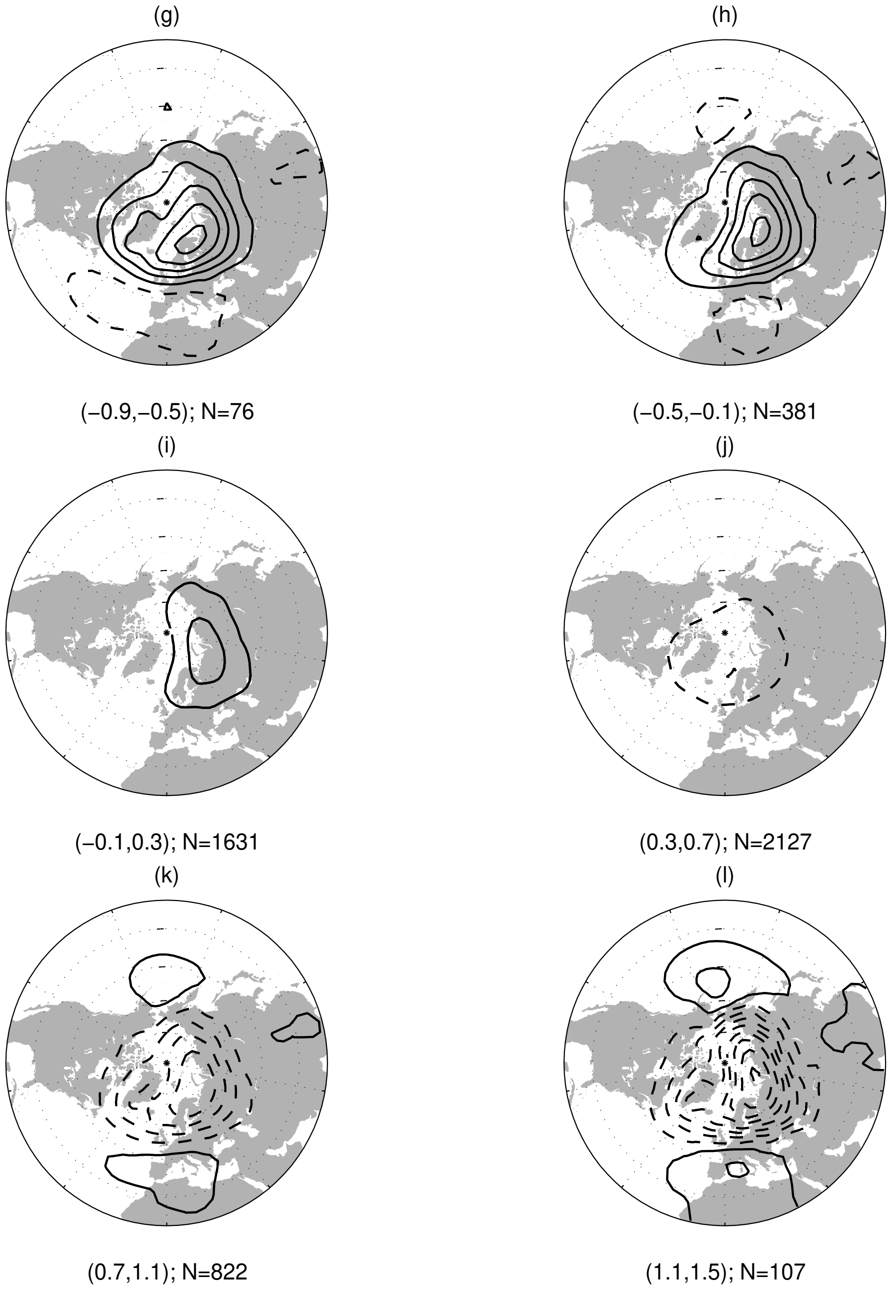


Figure 5.8: Continued.

composites (h) and (k) in Figure 5.8. These two maps display patterns of SLP anomalies that differ in sign and magnitude, but not in spatial pattern, consistent with the interpretation of Branch 1 variability as describing a standing oscillation. The anomalies associated with this oscillation are of opposite sign over the polar cap and the midlatitudes, with a polar local extremum over northern Eurasia and small local midlatitude extrema over the west Mediterranean and the north Pacific. This pattern resembles that of the canonical AO as diagnosed by EOF analysis, but with eastward shifted polar and Mediterranean centres of action. Indeed, the correlation between the AO time series (ie, the leading SLPA PC) and $\alpha(t_n)$, over those times when the approximation is on Branch 1, is -0.96.

Characteristic SLPA anomalies associated with Branch 2 are illustrated in Figure 5.8(c). This map shares certain features with the anomaly patterns displayed in Figures 5.8(h) and 5.8(k), in particular anomalies of opposite sign over the polar cap and the midlatitudes. However, the polar extremum in Figure 5.8(c) is shifted to a location over Iceland, the Atlantic midlatitude extremum is centred over the Azores, and anomalies over the midlatitude Pacific ocean are weak. This map resembles strongly the SLPA signature of the negative phase of the North Atlantic Oscillation. Note also that the anomaly pattern of opposite sign to that in Figure 5.8(c) does not appear on Branch 2, in contrast to what was observed along Branch 1. Variability along Branch 2 does not describe an oscillation, but episodic excursions to a single-phased, strongly anomalous circulation.

Figures 5.9 and 5.10 display respectively composites of the CCCma modelled Z_{500}A and Z_{500} fields over the same intervals in $\alpha(t_n)$ used to calculate the composites in Figure 5.8. The patterns displayed in Figure 5.9 demonstrate that the anomalous circulations are strongly equivalent barotropic, with a slight westward phase tilt with height. The midlatitude Z_{500} anomalies are slightly stronger relative to the polar anomalies than is

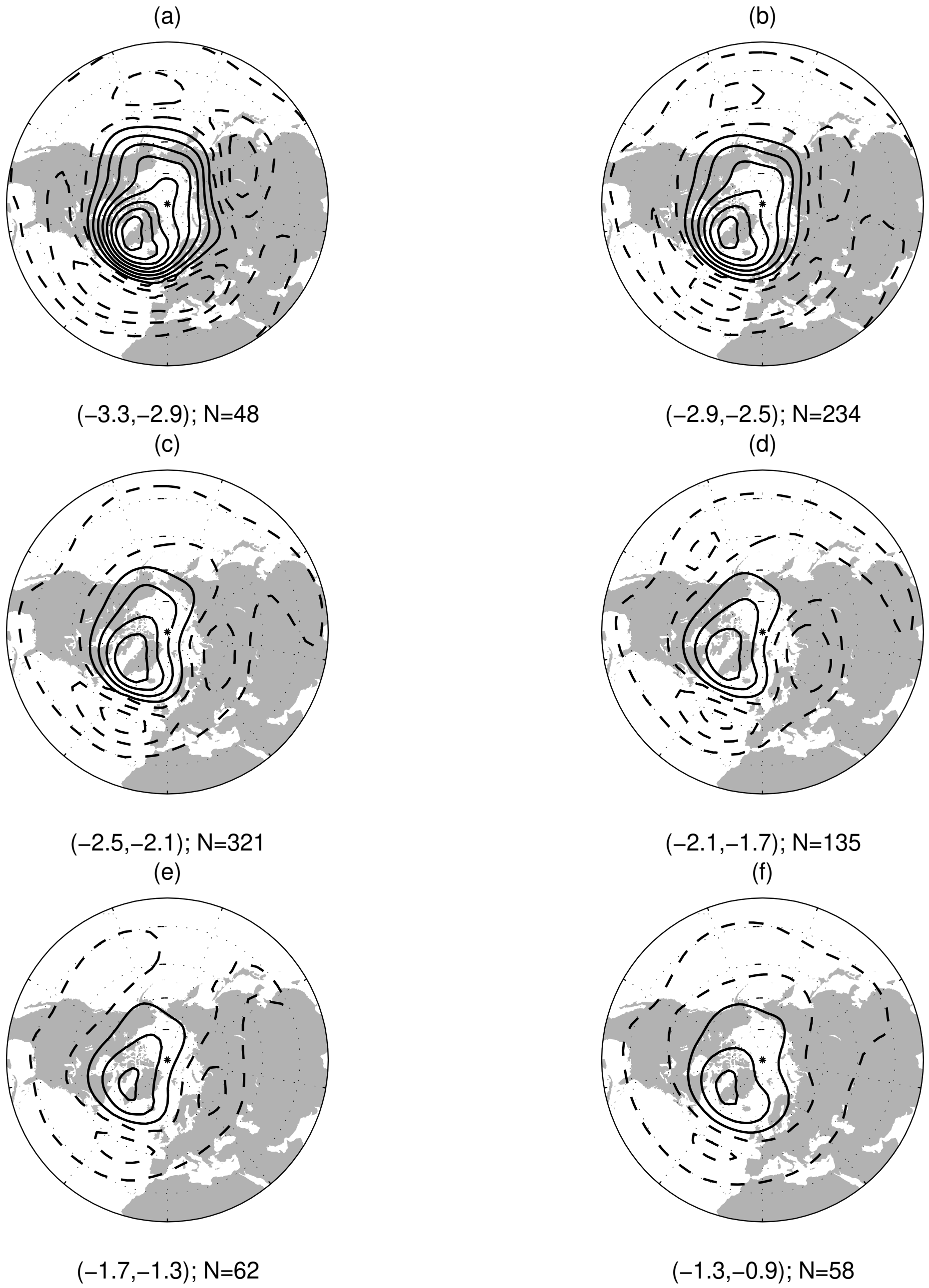


Figure 5.9: As with Figure 5.8, but for composites of Z_{500} anomalies. The contour interval is 20 m (...,-30,-10,10,30,...).

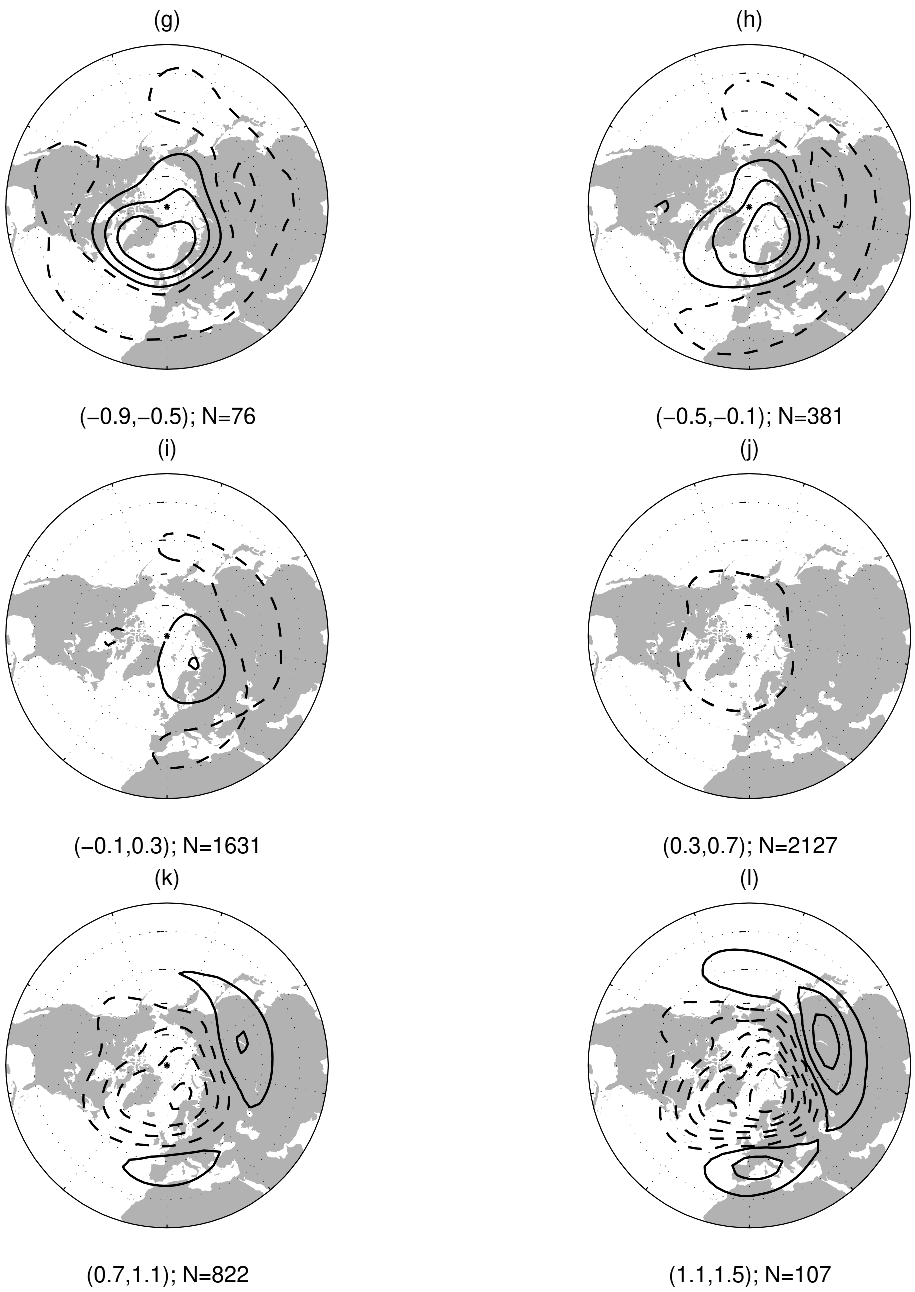


Figure 5.9: Continued.

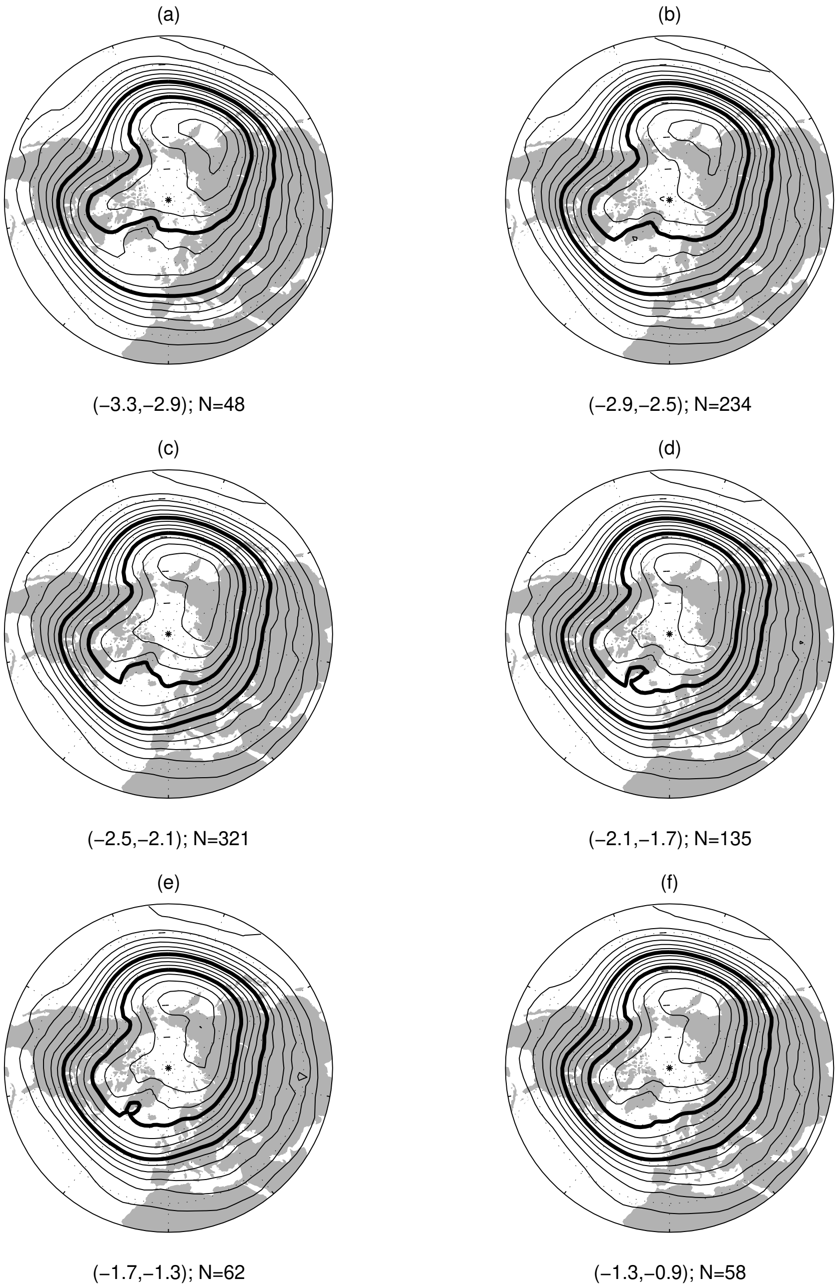


Figure 5.10: As with Figure 5.8, but for composites of Z_{500}. The 5300 and 5500 m contours are in bold. Contour interval is 50 m.

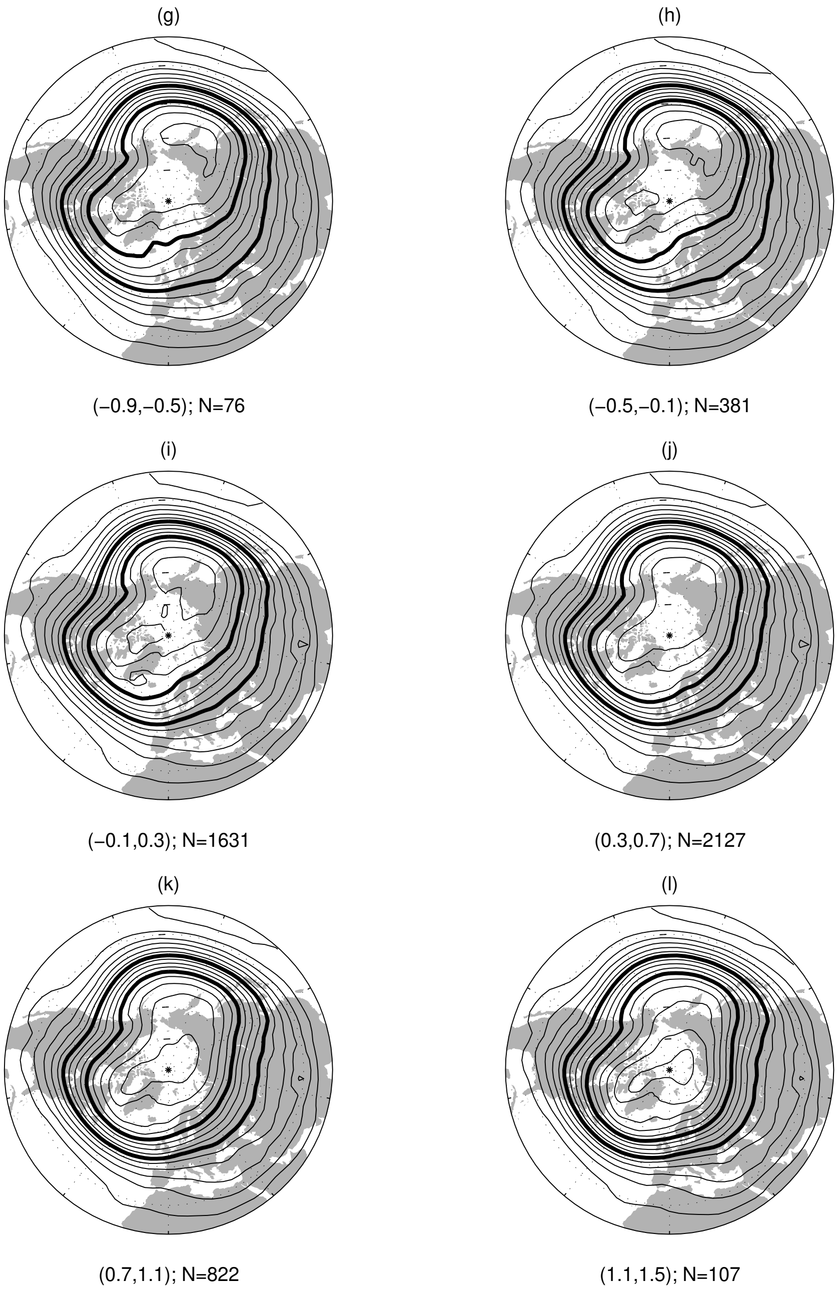


Figure 5.10: Continued.

the case with SLPA. Examination of the Z_{500} field is illuminating. Figures 5.10 (h)-(k) indicate that the mid-tropospheric manifestation of variability along Branch 1 consists of an alternating amplification and attenuation of the climatological ridge over Europe. Figure 5.10(c) demonstrates that Branch 2 describes, on average, an amplified climatological ridge over western North America and flow split around a local anticyclone over southern Greenland. Again, the mid-tropospheric composites associated with $\hat{\mathbf{X}}(t_n)$ describe two markedly different modes of atmospheric circulation: a standing oscillation in the strength of the climatological ridge over North Europe (Branch 1), and episodic split flow events over Greenland (Branch 2).

Figure 5.11 displays maps of the geographical distribution of variance and skewness of SLP, for both the observations and the CCCma model output. The skewness, s, of a random variable x is the ratio

$$s = \frac{< (x- < x >)^3 >}{< (x- < x >)^2 >^{3/2}} \tag{5.2}$$

and is a measure of the asymmetry of the distribution of x about its mean. For both the modelled and observed SLP, the variance is greater in the polar latitudes than in the midlatitudes, and displays a broad extremum from southern Greenland to northeastern Eurasia, with a second extremum over the North Pacific. The Pacific centre is substantially stronger in observations than in the model results. The broad local extremum in the variance of GCM SLP corresponds precisely to the polar centre of action in SLP EOF 1 (Figure 5.2(a)), and its west and east flanks correspond to the polar centres of action of Branches 1 and 2 of the NLPCA approximation $\hat{\mathbf{X}}(t_n)$. The leading EOF pattern of SLP is in a way a compromise between the circulation anomalies described by Branches 1 and 2. The broad extremum in variance results from the separate occurrence of two circulation regimes, and the PCA approximation, being linear, must attempt to characterise both modes of variability simultaneously.

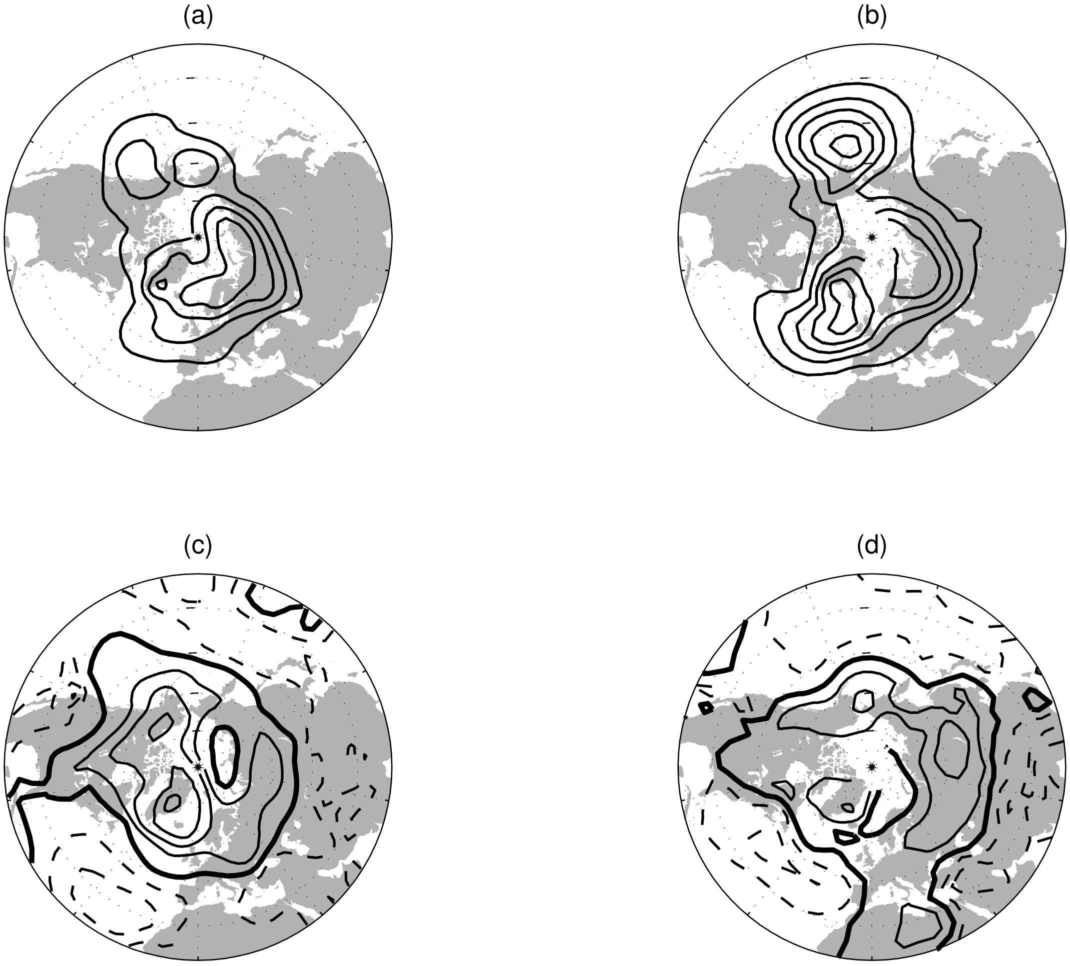


Figure 5.11: Maps of spatial distribution of variance of SLP from (a) CCCma GCM and (b) observations (contour interval 10 $(\mathrm{hPa})^2$), and skewness of SLP from (c) CCCma GCM and (d) observations (contour interval 0.2).

There is broad agreement between observed and modelled skewness; it is seen that both are generally characterised by negative values in the midlatitudes and positive values in polar latitudes, as was found by Nakamura and Wallace (1991) and by Holzer (1996). As well, both observed and modelled SLP skewness has a local maximum over Greenland and a local minimum over the midlatitude North Atlantic Ocean. In observations, the Greenland maximum is weaker than that of the modelled SLP, and the North Atlantic minimum is shifted northwestward of the corresponding minimum in the modelled SLP. As well, the modelled skewness over northern Eurasia is weaker than in observations, and the North Pacific extremum evident in observations is shifted to over Alaska. Note however that some of these apparent differences may in fact be an artifact of poor observational data coverage north of 70N, as is discussed further in section 5.6.

In their observational study of the geographical distribution of skewness in Northern Hemisphere tropospheric circulation, Nakamura and Wallace (1991) suggested that, "Large skewness at a particular location might be indicative of a juxtaposition of two flow regimes: one which prevails most of the time and another which occurs relatively infrequently and is thus characterised by large anomalies." This is in fact the situation seen to prevail in the NLPCA approximation $\hat{\mathbf{X}}(t_n)$ of SLPA. The dipole in skewness over the North Atlantic arises because of the regular occupation of Branch 1 and the episodic occupation of Branch 2. The negative extremum in skewness over the Atlantic Ocean west of Africa, which corresponds to the southern, negative centre of action of Branch 2 events, is displaced somewhat southward because skewness tends to be enhanced where the variance of the field is low, as is clear from equation (5.2).

5.4 Analysis of GCM 500mb Heights

In Figure 5.12 are displayed histogram estimates of the 2D marginal distributions of the Z_{500}A field PCs (PC1,PC2), (PC1,PC3), (PC2,PC3), and (PC1,PC4). The marginal distribution of PC1 and PC2 is closer to Gaussian than is the corresponding SLPA distribution, although PC1 is clearly skewed. In the Z_{500}A field, marked deviations from Gaussianity are most evident in the marginal distributions of PC1 with PC3 and of PC1 with PC4. Figures 5.12(b) and (d) demonstrate the existence of substantial non-Gaussian structure in the Z_{500}A data. All other 2D marginal distributions do not appear to differ strongly from being Gaussian.

The 1D NLPCA approximation to the Z_{500} data is displayed in Figure 5.13, projected in the spaces spanned by Z_{500}A EOFs $(\mathbf{e}_1, \mathbf{e}_2)$, $(\mathbf{e}_1, \mathbf{e}_3)$, $(\mathbf{e}_2, \mathbf{e}_3)$, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. This approximation explains 23.9% of the variance in the data, in contrast to 19.6% explained by the first PCA approximation. Note that unlike the corresponding SLPA approximation it projects strongly onto EOFs other than the leading two. The NLPCA approximation is constructed based on the joint distribution of the data in the 10-dimensional embedding space, and because it has non-negligible projections on EOFs beyond the first two, it is not particularly useful to compare the 2D marginal distributions presented in Figure 5.12 with the projections of the approximation in these planes. Thus, no equivalent of Figure 5.7 is presented. The 1D NLPCA Z_{500}A approximation is similar to that of the SLPA in that it is composed of three branches. As well, the corresponding standardised time series $\alpha(t_n)$ and its histogram (displayed in Figure 5.14) display a strikingly bimodal character. As was the case with the 1D SLPA NLPCA approximation, one branch of $\hat{\mathbf{X}}(t_n)$ corresponds to the larger peak of the distribution of $\alpha(t_n)$, a second to the smaller peak, and the third branch to the rarely-visited region in between. The system spends 78% of all months in the branch corresponding to the larger peak, 19% in the branch

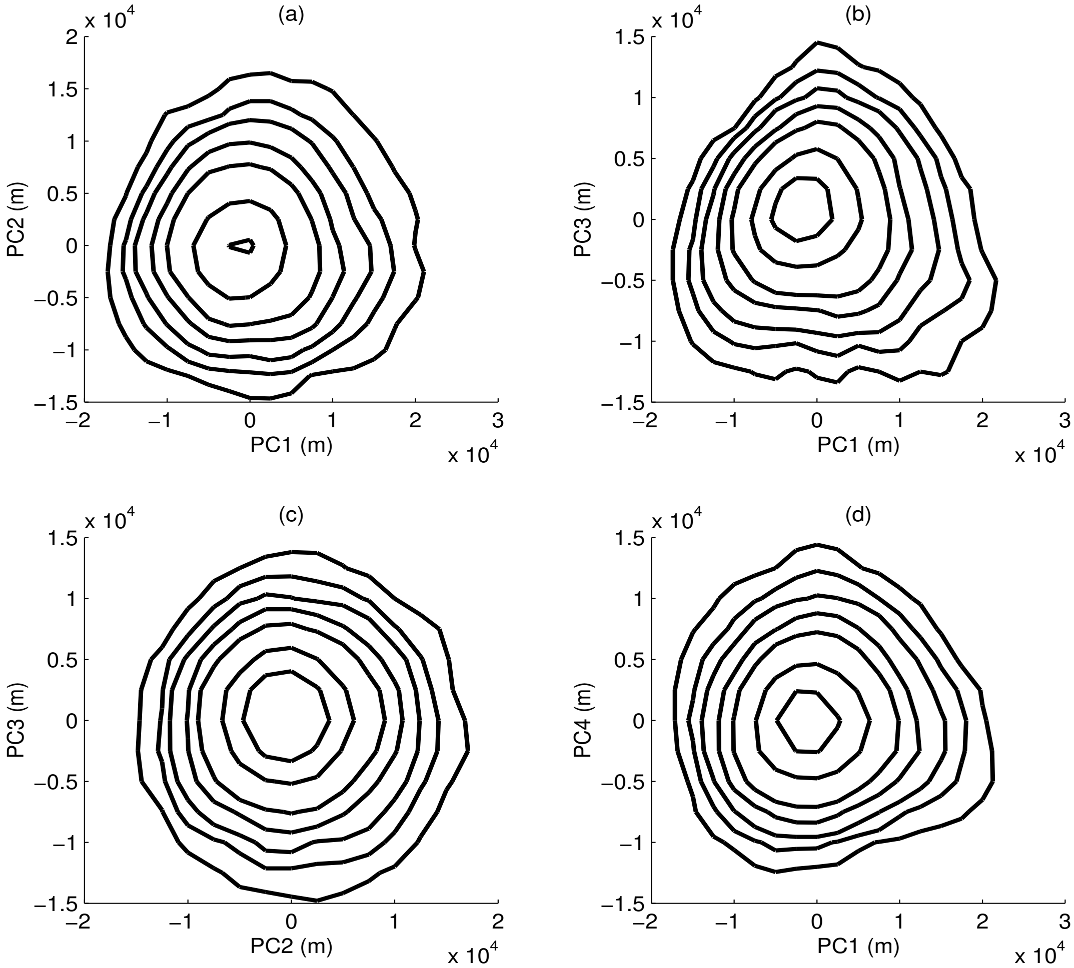


Figure 5.12: Histogram estimates of 2D marginal PDFs of Z_{500}A PCs (a)$(PC1, PC2)$, (b)$(PC1, PC3)$, (c)$(PC2, PC3)$, and (d)$(PC1, PC4)$. The contour intervals are as in Figure 5.4. Bin sizes are 2500 m.

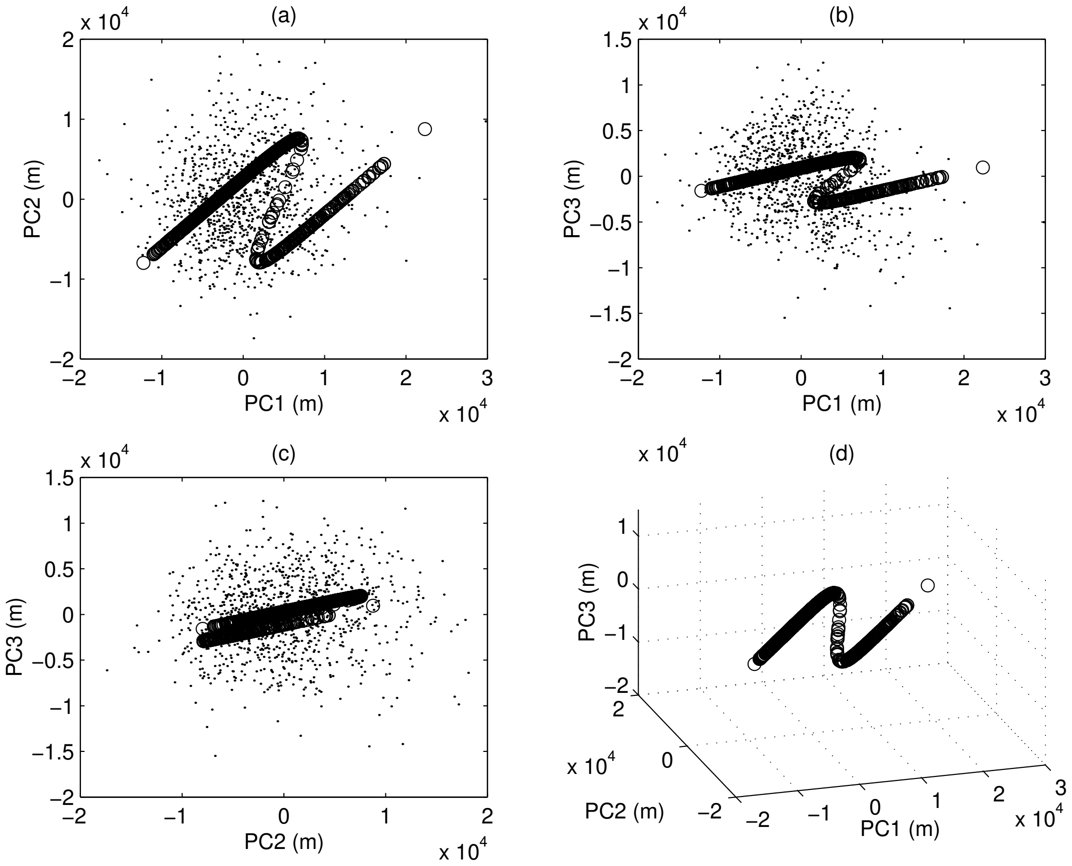


Figure 5.13: Projection of the Z_{500}A data (dots) and its 1D NLPCA approximation (open circles) onto the spaces spanned by EOFs (a) $(\mathbf{e}_1, \mathbf{e}_2)$, (b) $(\mathbf{e}_1, \mathbf{e}_3)$, (c) $(\mathbf{e}_2, \mathbf{e}_3)$, and (d) $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$.

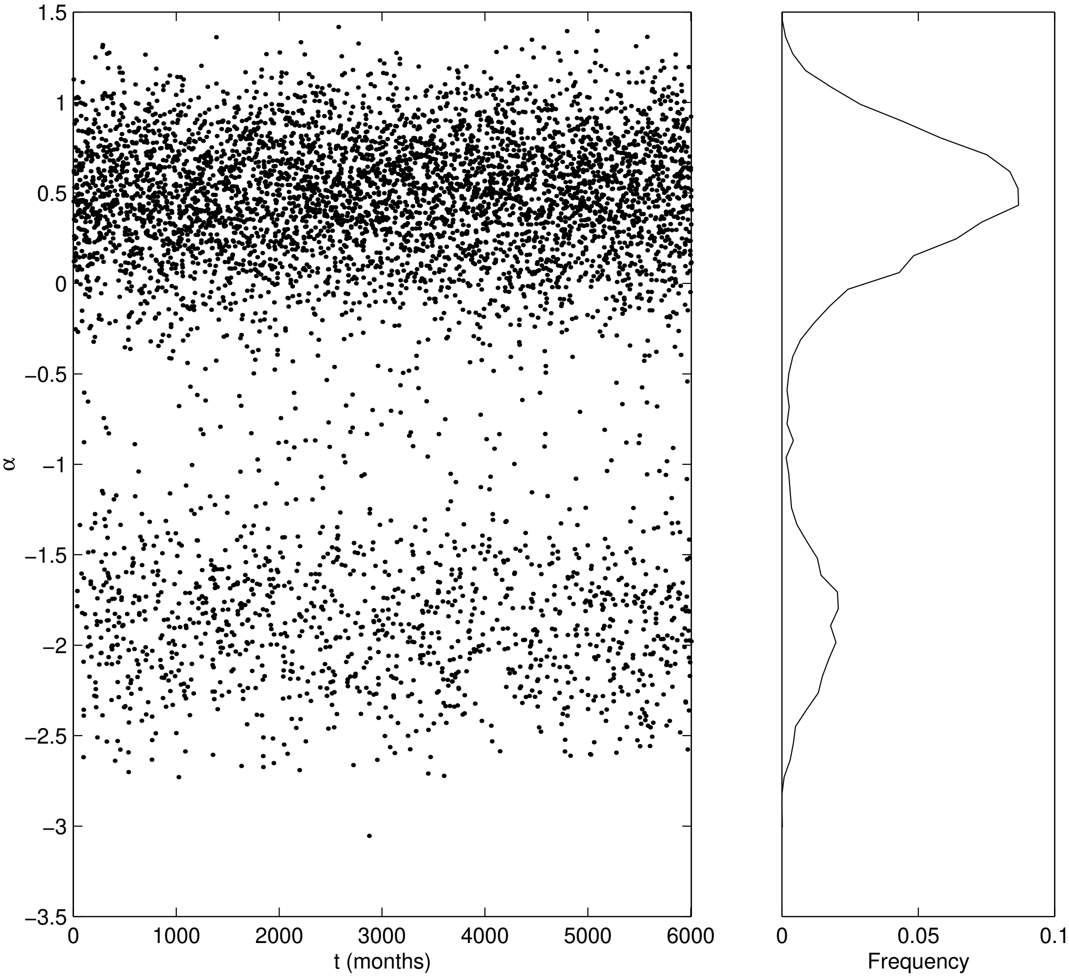


Figure 5.14: Plot of the 1D Z_{500}A time series $\alpha(t_n)$ (left) and the associated histogram estimate of the PDF (right).

corresponding to the smaller peak, and 3% in the intermediate branch. An inspection of the time series $\alpha(t_n)$ indicates that variability in the approximation is characterised by oscillatory motion along the more populated branch, with episodic excursions to the less populated branch.

Figure 5.15 displays composites of Z_{500}A over times associated with neighbourhoods in α of width 0.375, centred at intervals of 0.375 from -2.9 to 1.6. Inspection of these composites indicates that the branch of variability associated with the larger of the two peaks in the PDF of $\alpha(t_n)$ corresponds to essentially a standing oscillation (Figures 5.15 (h)-(l), especially (h) and (k)). This oscillation is characterised by anomalies with a strong wavenumber four signal in midlatitudes and a wave number two signal in higher latitudes. The circulation anomalies over the eastern North Pacific and North America display a PNA-like structure, and those over Eurasia resemble the Scandinavia pattern described in Bell and Halpert (1995). The structure of this oscillation in midlatitudes bears a striking similarity to that observed in the 500 mb field of the leading mode of the combined EOF analysis carried out by Baldwin and Dunkerton (1999), although the signal in the polar regions is rather different.

Variability along the branch associated with the smaller of the two peaks is illustrated in Figures 5.15 (a)-(d). Unlike the branch associated with the larger peak, anomalies along this branch are of a single phase, with highs over the polar region and lows over the midlatitudes. A local maximum in the anomaly field over the polar region is located over Southern Greenland, with a local negative extremum in midlatitudes over the Azores. This structure is strongly reminiscent of the negative phase of the NAO.

A comparison of the composites presented in Figure 5.15 with those given in Figure 5.9, and of the time series presented in Figures 5.14 and 5.6, demonstrates that the 1D NLPCA approximations of SLPA and Z_{500}A describe essentially the same mode of variability. The circulation anomalies plotted in Figure 5.9(b)-(d) bear a striking resemblance

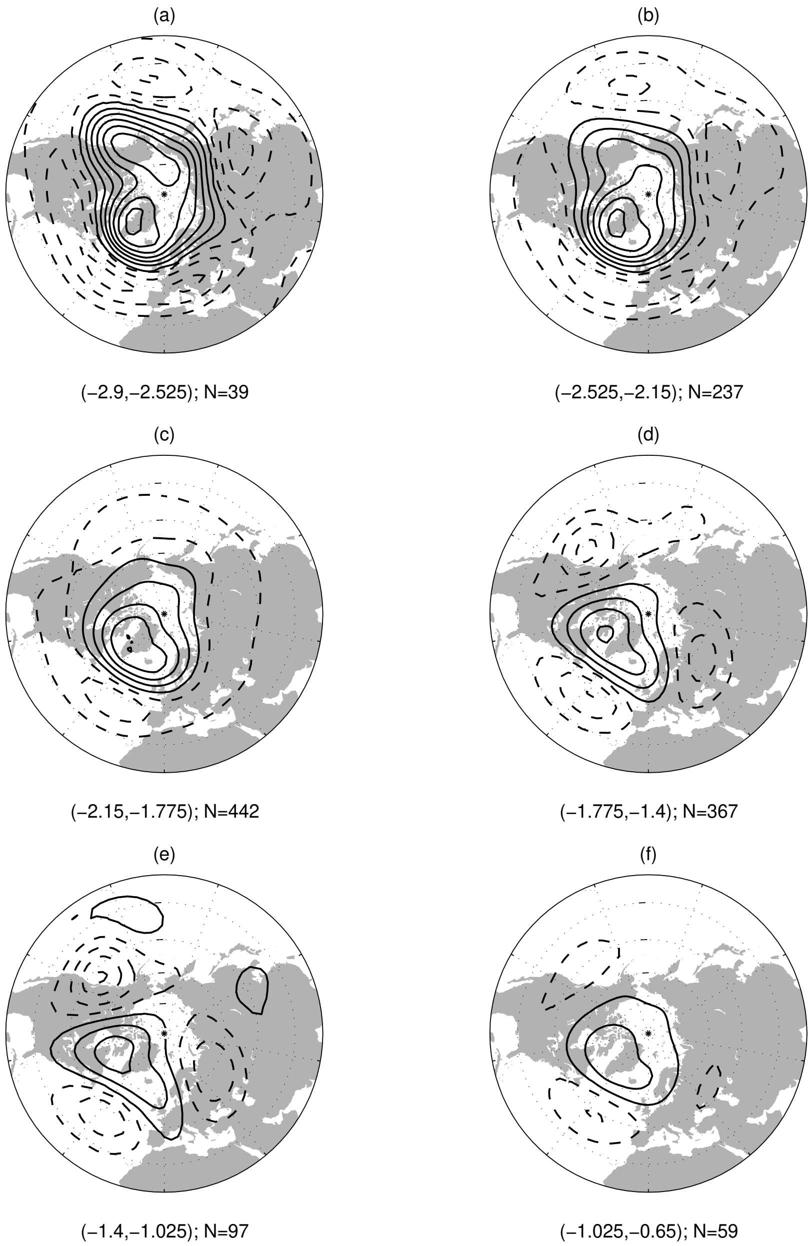


Figure 5.15: As in Figure 5.9, but for Z_{500}A composited over characteristic ranges of α associated with the 1D Z_{500}A NLPCA approximation. Contour interval is 20m (...,-30,-10,10,30,...). Continued on next page.

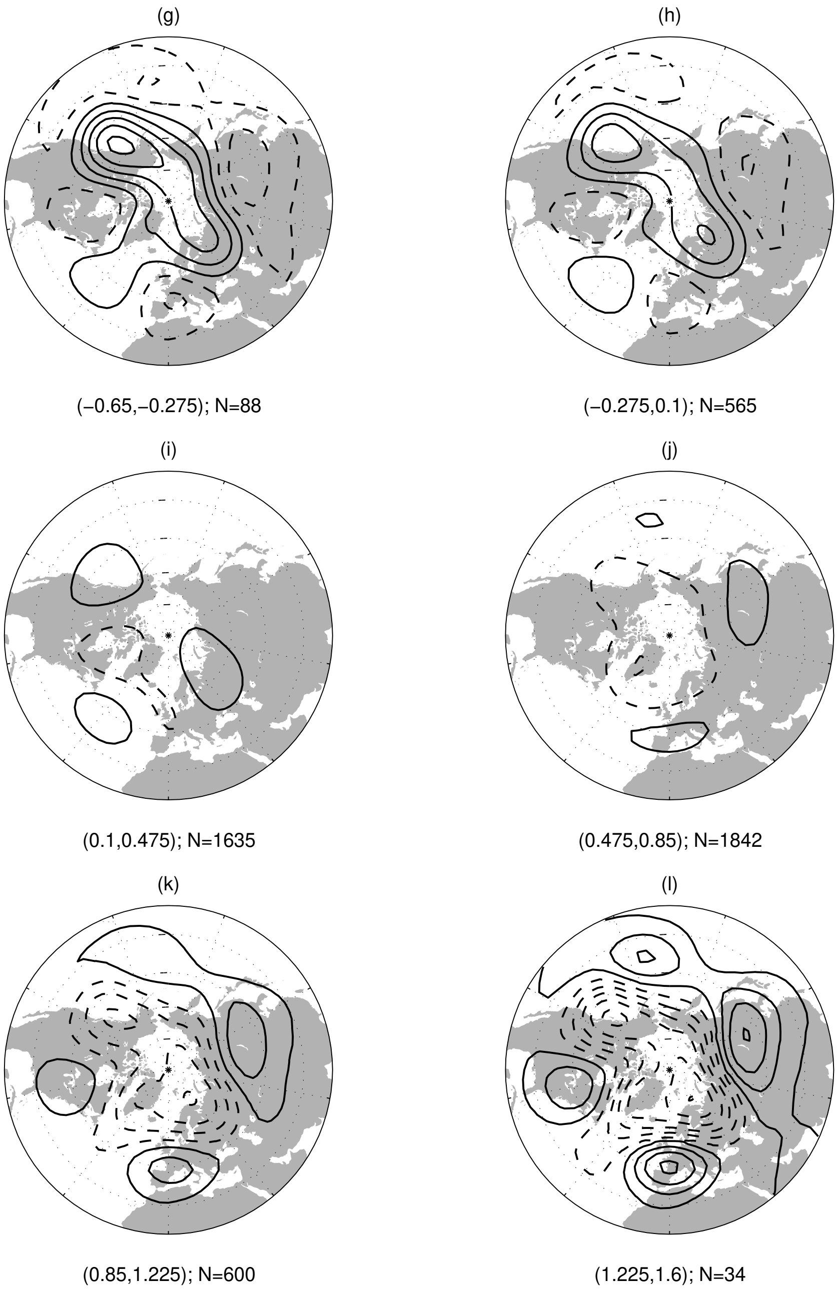


Figure 5.15: Continued.

to those plotted in Figure 5.15(b)-(d). Similarly, the polar and Eurasian sectors of the standing oscillation as diagnosed from the analysis of SLPA (Figure 5.9 (h) and (k)) are essentially the same as those diagnosed from the analysis of Z_{500}A (Figure 5.15(h) and (k)). The more and less populated regimes of circulation in the Z_{500}A field are referred to as the oscillatory and split-flow regimes, respectively. This characterisation will be borne out by the composites of Z_{500} to be presented later.

The two approximations do not correspond exactly. The Z_{500}A approximation is in the less-populated regime 19% of the time, in contrast to 13% for the SLPA approximation. The corresponding numbers for the branch describing the standing oscillation are 78% and 84%. The split-flow regime is more frequently occupied in the Z_{500}A approximation than in the SLPA. Figure 5.16 displays a scatterplot of the time series $\alpha(t_n)$ corresponding to the SLPA and Z_{500}A approximations. The correlation coefficient between these two time series is 0.81. Typically the approximations are simultaneously both in either the oscillatory or split-flow regimes. For a small number of months, the SLPA approximation is in the split-flow regime while the Z_{500}A approximation is in the oscillatory regime, but more commonly when the two approximations differ it is because the SLPA approximation is in the oscillatory regime when the Z_{500}A approximation is in the split-flow regime. This is a reflection of the fact that the split-flow regime is more frequently occupied in the Z_{500}A approximation than in the SLPA approximation.

This difference in occupation statistics has little effect on the composites characterising the split-flow branch. However, the standing oscillation diagnosed from the Z_{500}A data is more hemispheric in extent than is that from the SLPA data. Note in particular the presence in the Z_{500}A approximation of a local extremum in the height anomalies over western Canada and Alaska of the same sign as the anomalies over the pole, and local extrema of the opposite sign over the central North Pacific and eastern North America which together comprise a wavetrain reminiscent of the PNA pattern. These extrema do

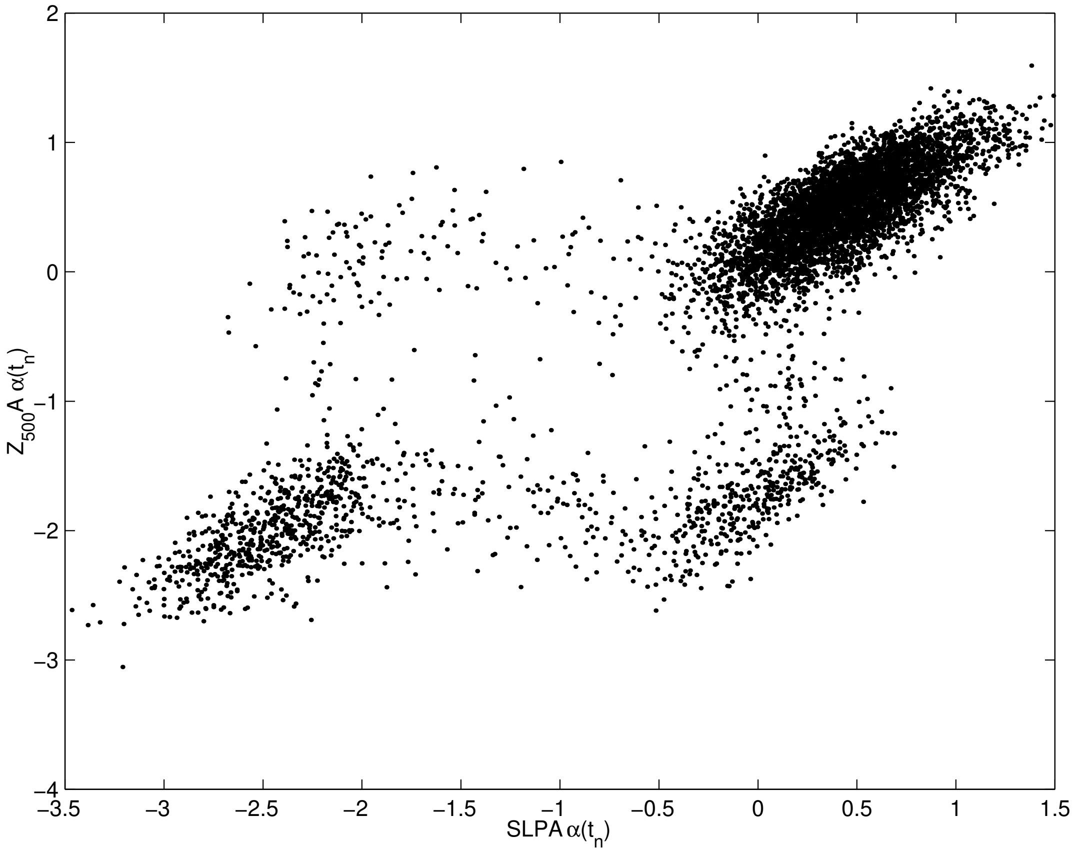


Figure 5.16: Scatterplot of the time series $\alpha(t_n)$ corresponding to the 1D SLPA and Z_{500}A NLPCA approximations.

not occur in the composites corresponding to the 1D SLPA NLPCA approximation.

Figure 5.17 presents the geographical distribution of variance and skewness of the CCCma modelled Z_{500}A field. These maps show the same sort of structure as those for the modelled SLP: the variance displays a general increase with latitude, and the skewness is generally negative in midlatitudes and positive in the polar region. As well, the standard deviation and skewness of Z_{500}A display local extrema in much the same locations as SLPA, but with a considerably stronger signal in the North Pacific. As was the case with SLPA, both the skewness and variance maps are consistent with the two-regime structure of the NLPCA approximation of Z_{500}A.

Composites of the total Z_{500} field are given in Figure 5.18. The composites on the split-flow branch of the Z_{500}A approximation do not differ substantially from those based on the SLPA analysis. The split-flow branch of the 1D Z_{500}A NLPCA approximation is seen indeed to describe split flow over Greenland and an enhanced ridge over Western Canada. Variability along the oscillatory branch of the Z_{500}A approximation describes alternating amplification and attenuation of the climatological ridges over both Europe and North America, in contrast to variability along the SLPA approximation, which was concentrated in the European sector.

Finally, Figure 5.19 displays composites of SLPA based on the 1D Z_{500}A NLPCA approximation. As was the case with the Z_{500}A and Z_{500} fields, the composites of SLPA over characteristic ranges of the time series $\alpha(t_n)$ corresponding to the Z_{500}A approximation are more hemispheric in extent than was the case with the corresponding composites based on the SLPA approximation. In particular, there is enhanced variability over western Canada, corresponding to an amplified wavenumber 2 contribution to the anomaly field.

Thus, a 1D NLPCA analysis of a second field from the CCCMA coupled GCM, viz Northern Hemisphere wintertime 500mb geopotential height anomalies, produces results

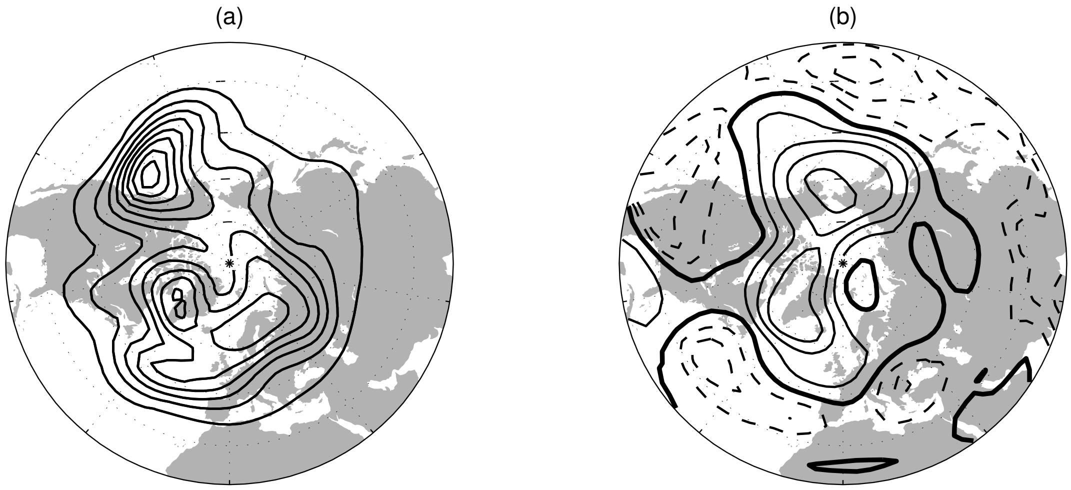


Figure 5.17: Maps of variance (a) and skewness (b) of CCCma modelled Z_{500}. Contour interval in (a) is 500 m^2 and in (b) is 0.2.

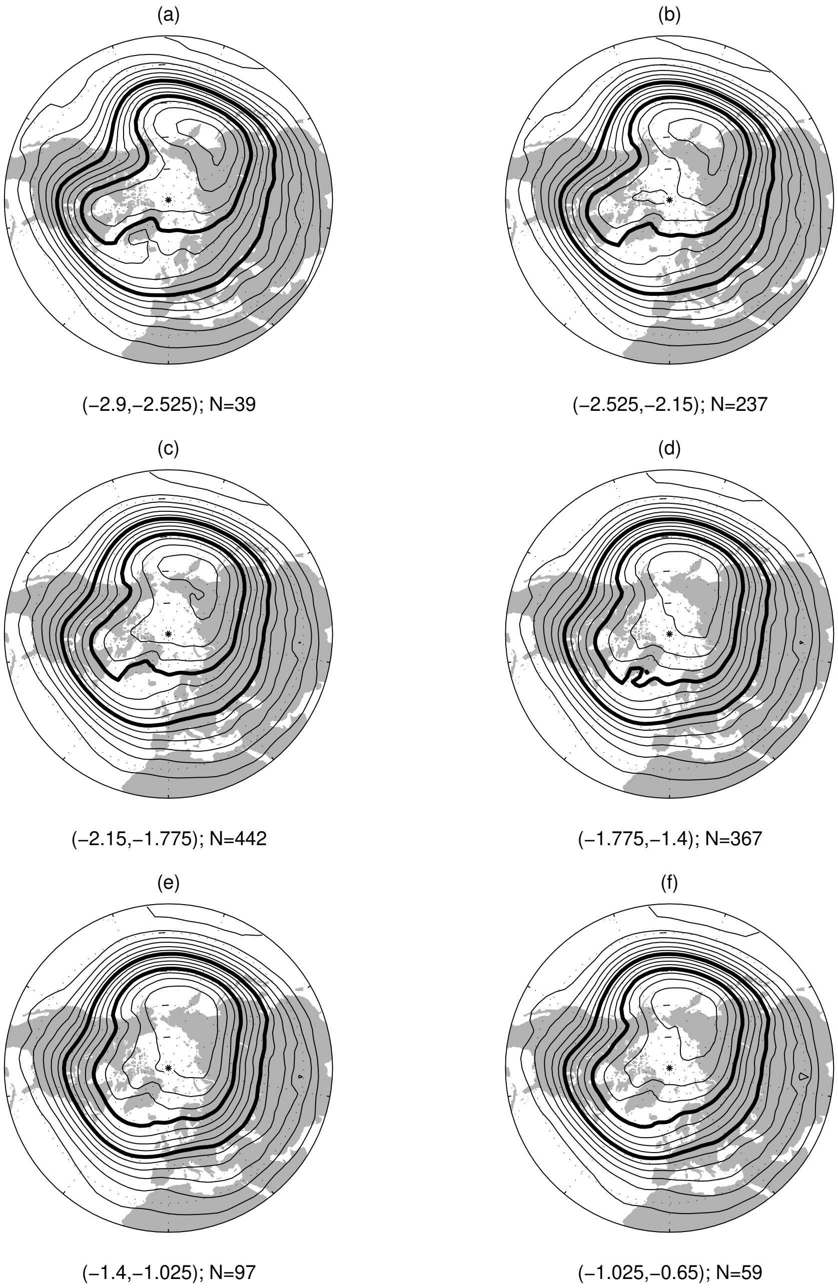


Figure 5.18: As with Figure 5.15, but for composites of Z_{500}. The 5300 and 5500 m contours are in bold. Contour interval is 50 m.

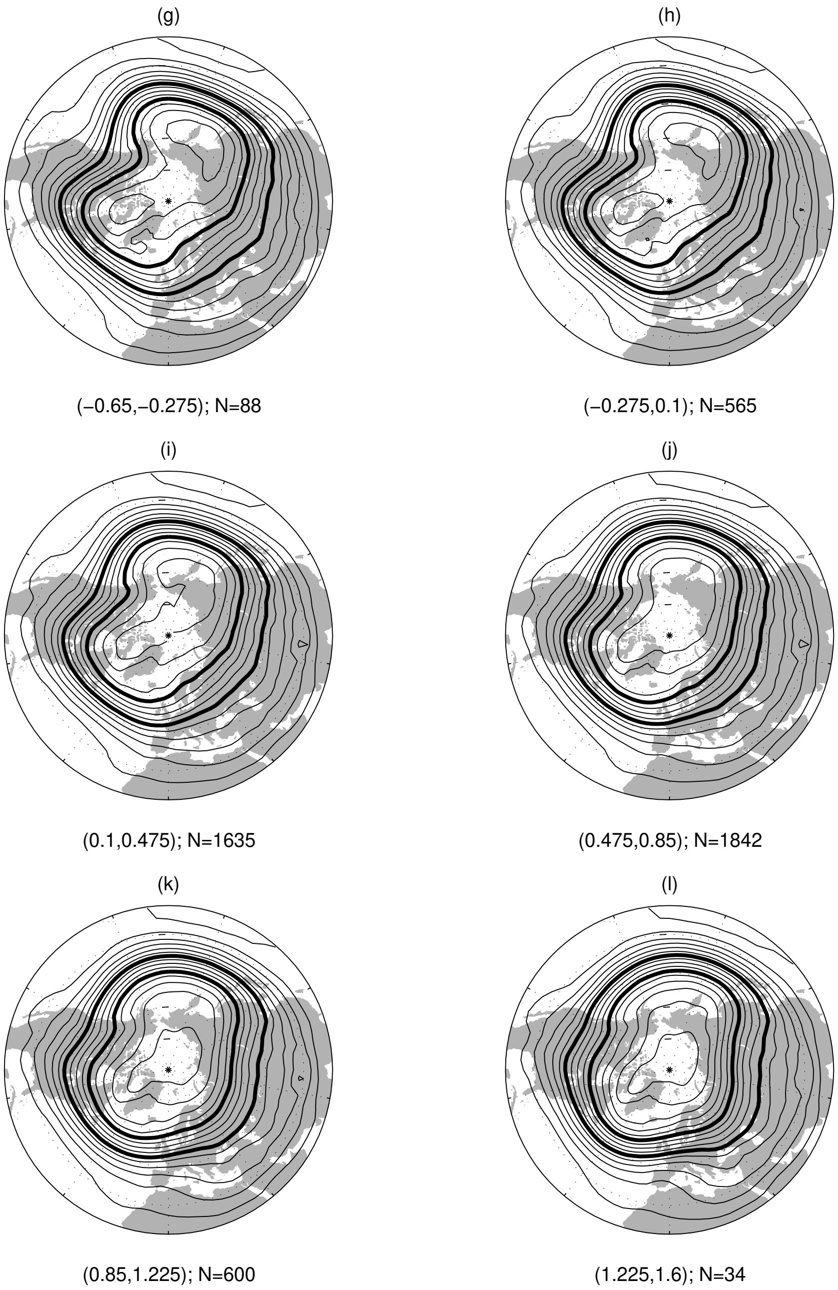


Figure 5.18: Continued.

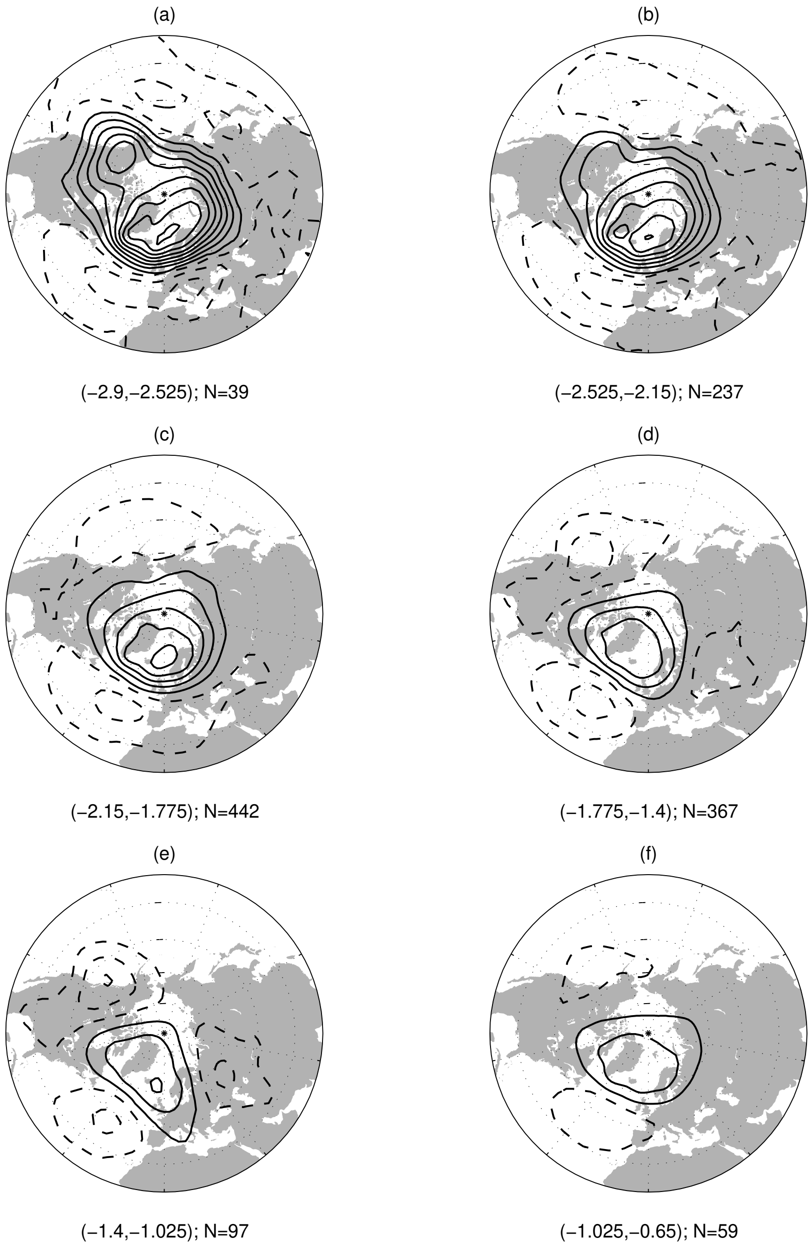


Figure 5.19: As in Figure 5.15, but for SLPA. Contour interval is 2 hPa (...,-3,-1,1,3,...). Continued on next page.

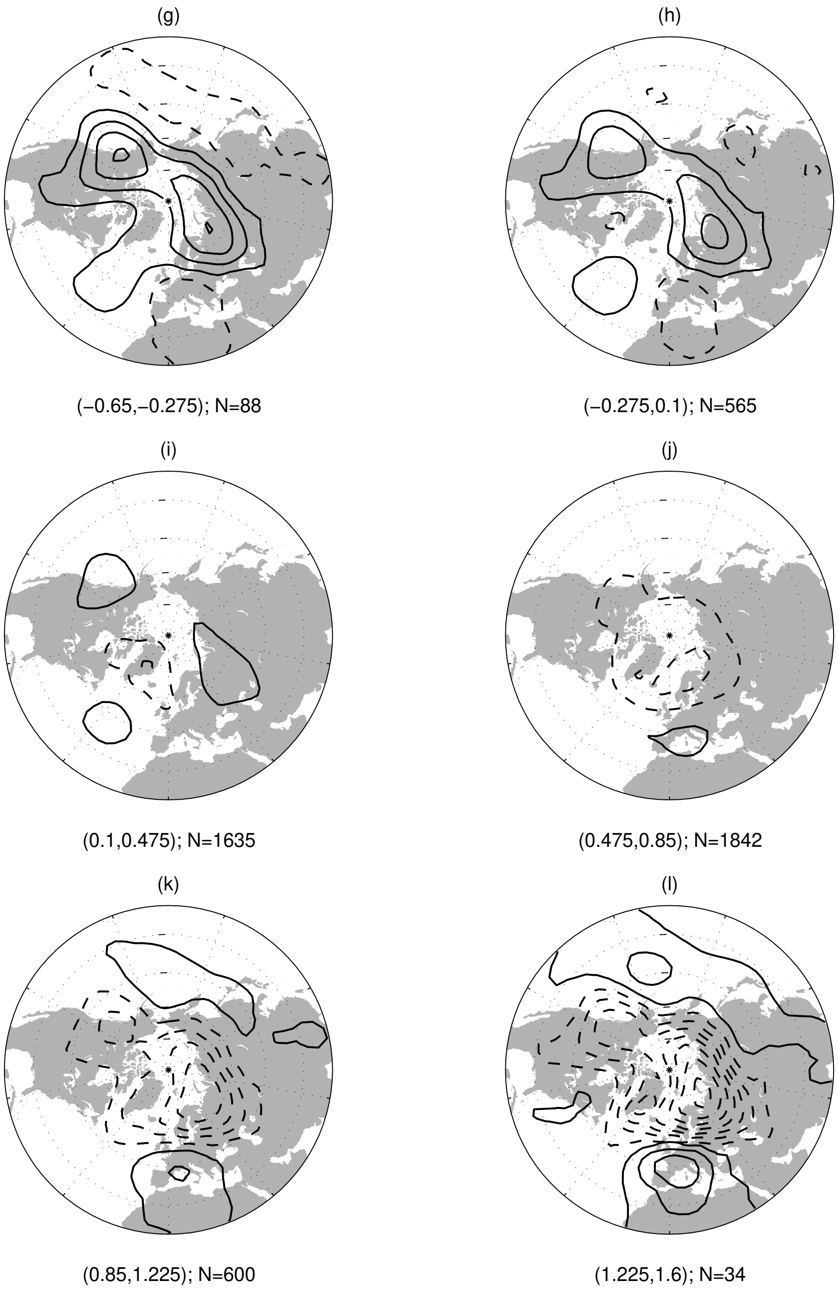


Figure 5.19: Continued.

that are very similar to those produced by the analysis of the SLPA field, providing strong evidence that this structure is a robust feature of the data. The structures diagnosed in the two fields are not exactly the same: the Z_{500}A approximation anomaly fields are more hemispheric in extent than those of the SLPA approximation, and the split-flow branch is occupied more frequently.

5.5 Analysis of GCM SLP in a $4\times CO_2$ Integration

A plot of the histogram estimate of the 2D marginal PDF of SLPA from the GCM integration with CO_2 concentrations at four times the pre-industrial level, in the space of the leading two control integration EOFs, is presented in Figure 5.20, along with the corresponding 1D NLPCA approximation. The NLPCA approximation explains 31.4% of the variance of the data, in contrast to 29.8% explained by the first PCA mode. The most striking aspect of this PDF is the fact that it appears to be much more Gaussian than the corresponding PDF for the control integration (Figure 5.4). In particular, the PDF of the $4\times CO_2$ integration lacks the bulge in the lower-right quadrant associated with the split-flow branch of the control integration 1D NLPCA approximation. That the structure of the data is now much more nearly Gaussian is reflected in the structure of the 1D NLPCA approximation displayed in Figure 5.20. This approximation is no longer branched, but is a slightly curved line lying nearly along the major axis of the marginal distribution. The corresponding time series (Figure 5.21) is unimodal. Note that the NLPCA approximation of the SLPA in the $4\times CO_2$ integration lies along what was the oscillatory branch of the SLPA NLPCA approximation from the control integration. Figure 5.22 displays composites of the 1D $4\times CO_2$ SLPA NLPCA approximation over characteristic ranges of the (standardised) time series $\alpha(t_n)$. The variability illustrated in Figure 5.22 is not precisely that of a standing oscillation; the composites over the

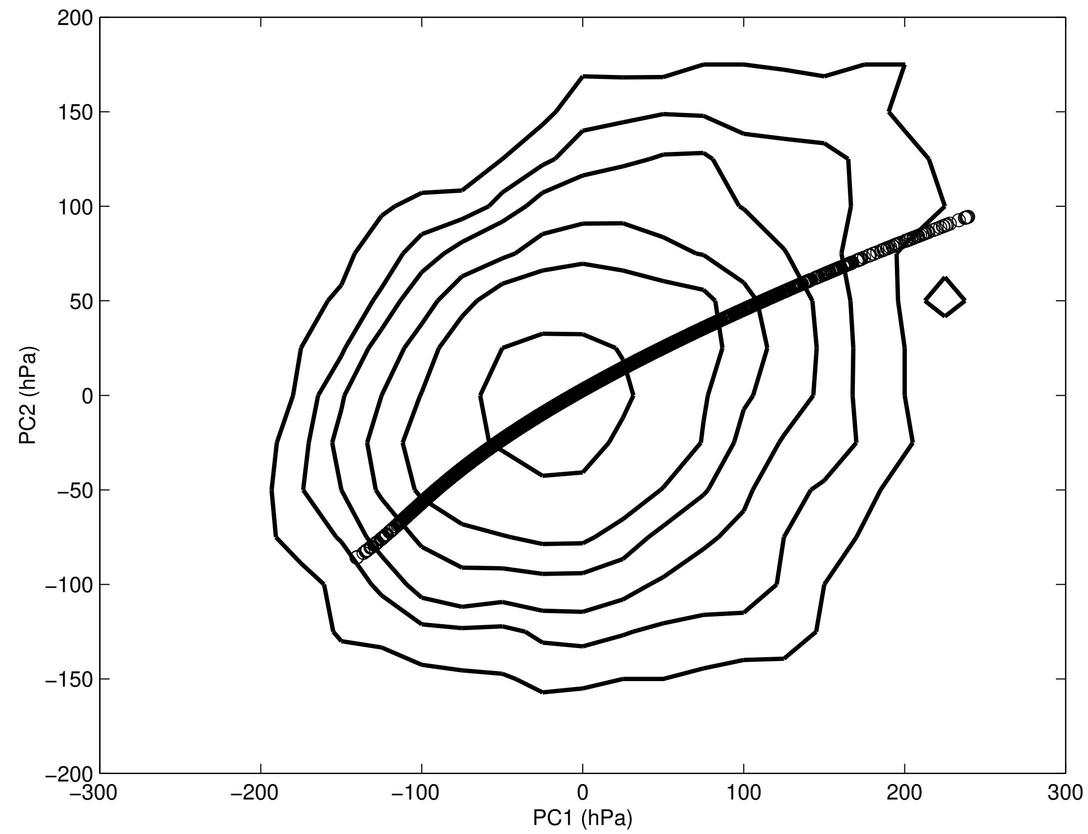


Figure 5.20: Histogram estimate of the marginal probability density function (contours) of SLPA from the GCM integration with CO_2 concentrations at four times the pre-industrial value, in the space of the leading two control integration SLPA PCA modes, overlaid with the corresponding 1D NLPCA approximation (open circles). Contour intervals are $5 \times 10^{-4}, 1.5 \times 10^{-3}, 3 \times 10^{-3}, 6 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 3 \times 10^{-2}$. The histogram bin size is 25 hPa in both directions.

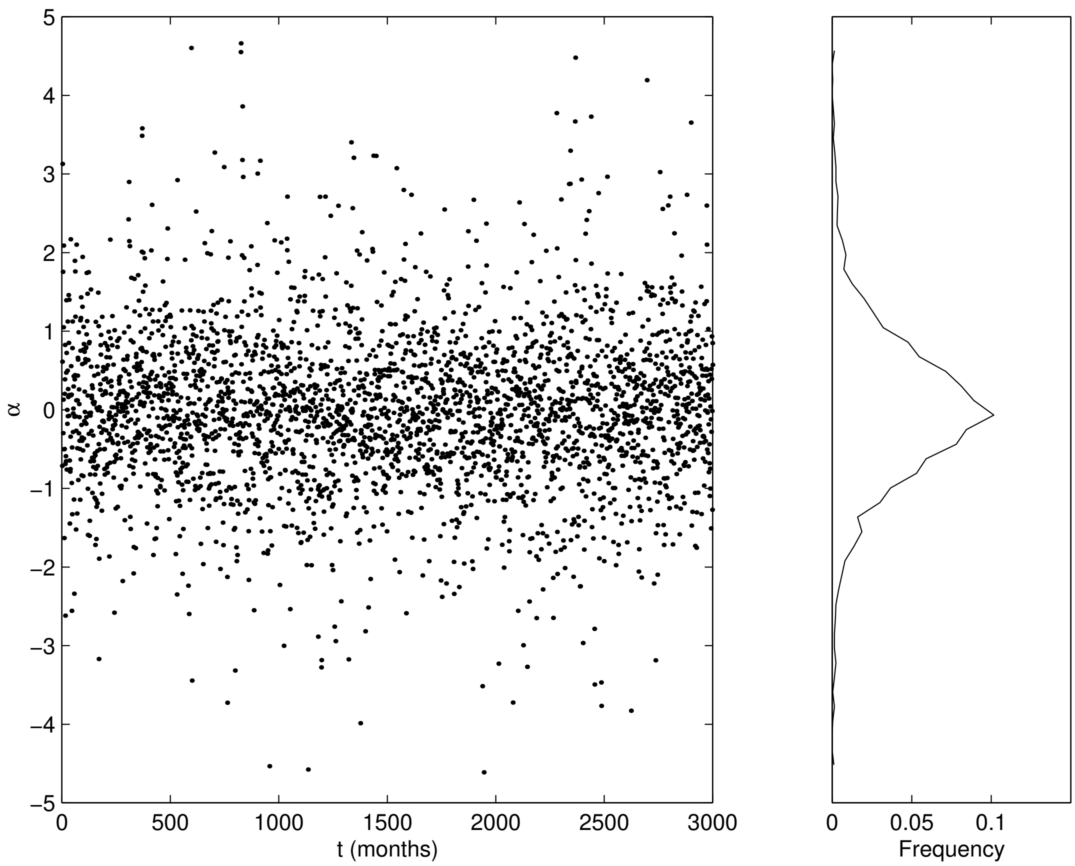


Figure 5.21: Plot of the 1D NLPCA SLPA time series $\alpha(t_n)$ (left) and the associated histogram estimate of the PDF (right) for the GCM integration with CO_2 concentration at four times the pre-industrial level.

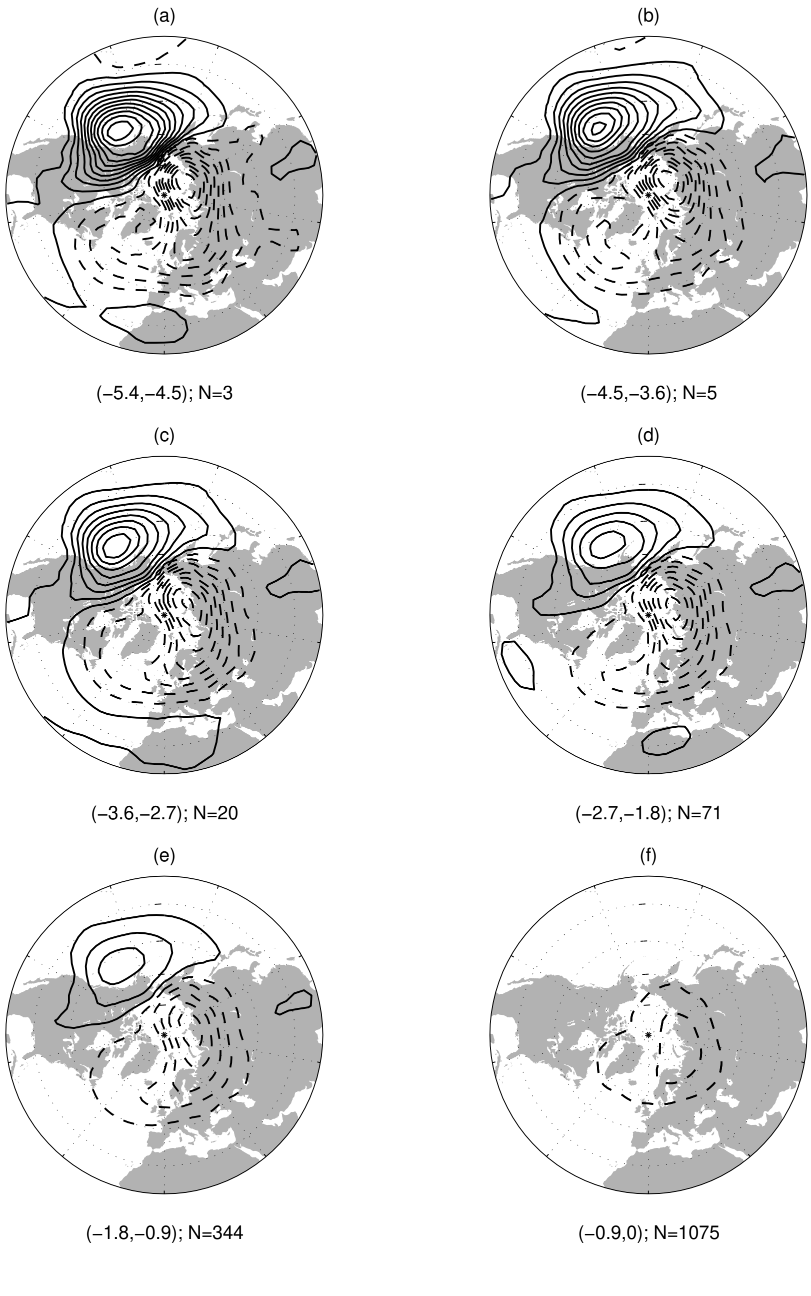


Figure 5.22: As in Figure 5.8, but for SLPA in the GCM integration at four times the pre-industrial CO_2 concentration. Contour interval is 2 hPa (...,-3,-1,1,3,...). Continued on next page.

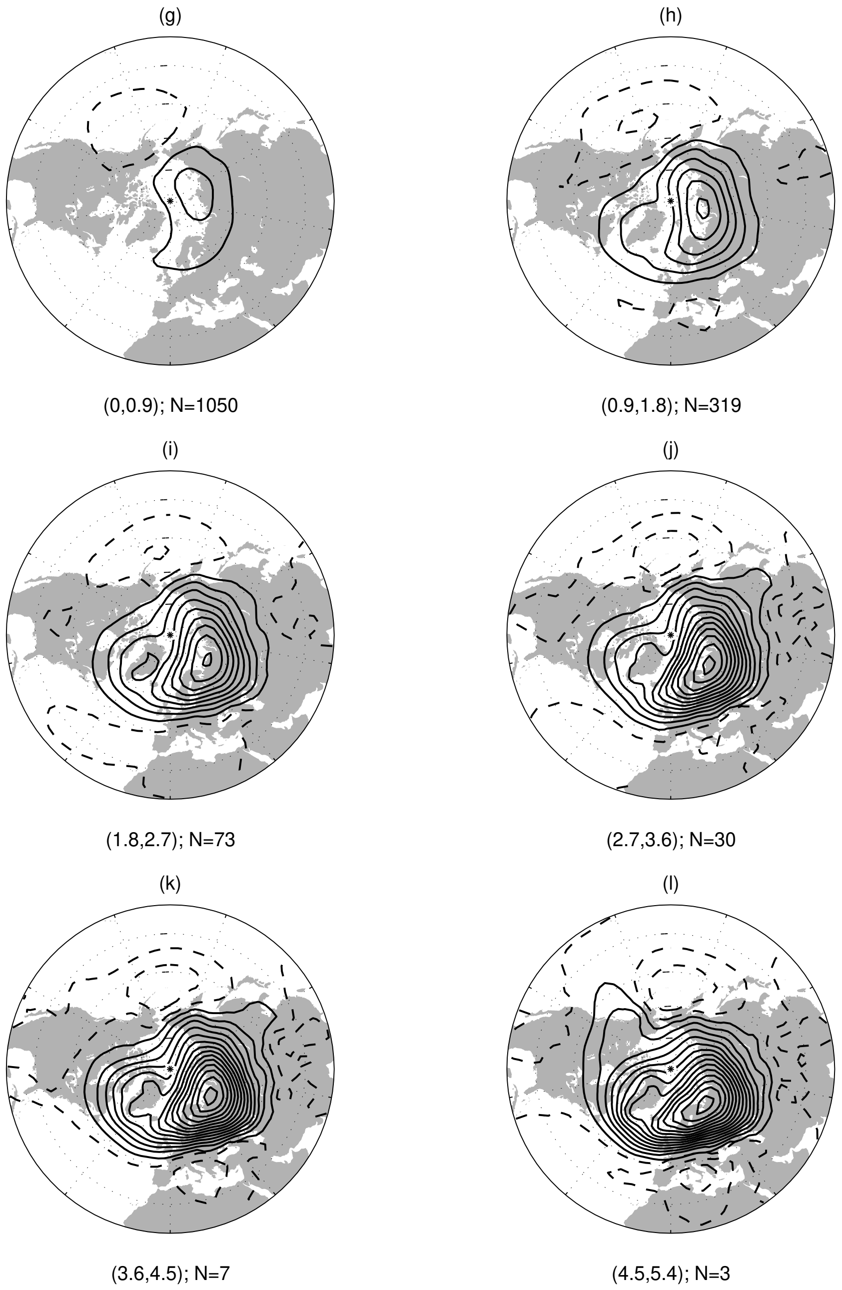


Figure 5.22: Continued.

tails of α accentuate strongly positive anomalies over the northeast Pacific and northern Siberia, respectively. Note, however, that the approximation is in a state between Figures 5.22(e) and (h) 93% of the time, over which range the structure is essentially that of a standing oscillation. Note also that the structure of this oscillation is very similar to that of the oscillatory branch of the 1D NLPCA approximation of SLPA in the control integration, with a slightly enhanced North Pacific centre of action. Thus, the dominant mode of nonlinear variability in SLPA under quadrupled CO_2 is only weakly nonlinear and strongly resembles the oscillatory regime of SLPA variability under pre-industrial CO_2 concentrations.

Palmer (1999) has suggested, based on experiments with simple low-dimensional nonlinear systems, that the response of the climate system to external perturbations (e.g. increased atmospheric CO_2) will not be a change in the structure of dominant circulation regimes, but rather in their occupation frequencies. Our results are broadly consistent with this hypothesis. Under quadrupled CO_2, the oscillatory regime of SLPA variability becomes more frequently occupied at the expense of the split-flow regime. In fact, the latter almost disappears entirely.

Our results are consistent as well with those of Ulbrich and Christoph (1999), who found that the ECHAM4+OPYC3 coupled GCM predicted a systematic northeastward shift of the centres of action of the NAO with an increase in atmospheric CO_2. The NAO as diagnosed using linear statistical tools is seen in the light of the 1D control SLPA NLPCA approximation as a compromise between Branch 1 and Branch 2 variability. As the split-flow branch becomes depopulated with increasing atmospheric CO_2, the compromise will increasingly favour the anomaly patterns characteristic of the oscillatory branch. This change would manifest itself in a regional analysis as a secular northeastward shift of the centres of action of the NAO.

Figure 5.23 displays maps of the variance and skewness of the SLPA field in the $4\times CO_2$

integration. The variance displays local extrema over Northern Russia and Southern Alaska; both of these locations correspond to centres of action in the range of α over which the 1D NLPCA approximation is effectively a standing oscillation. The skewness is strongly positive over the Northeast Pacific and over European Russia. Inspection of the composites displayed in Figure 5.22 indicates that the NLPCA approximation is consistent with this distribution of skewness. For strongly negative values of $\alpha(t_n)$, there are strong positive SLP anomalies centred south of Alaska that are not counterbalanced by similar negative anomalies for large positive $\alpha(t_n)$. Similarly, for strongly positive $\alpha(t_n)$, strong positive SLP anomalies occur over Northern Russia that do not have a negative counterpart for large negative values of α. Thus, the 1D NLPCA approximation characterises the gross features of both the skewness and the variance of the $4\times CO_2$ SLPA data. Note that while the tails in the distribution of $\alpha(t_n)$ do not contribute substantially to the variance, the structure of which is dominated by variability near the mean of $\alpha(t_n)$, they are manifest in the spatial structure of the skewness. Note as well that because skewness involves a compromise between the second and third order moments, its extrema are shifted somewhat southward of the centres of action in the composite towards latitudes of lower variance.

Thus, in the CCCMA coupled GCM, the quadrupling of the atmospheric concentration of CO_2 results not so much in a change in the structures of atmospheric variability, but in a change in their occupation frequencies, consistent with the hypothesis of Palmer (1999). The split-flow branch of the 1D control SLPA approximation is depopulated in the $4\times CO_2$ run, while the oscillatory regime remains largely unchanged.

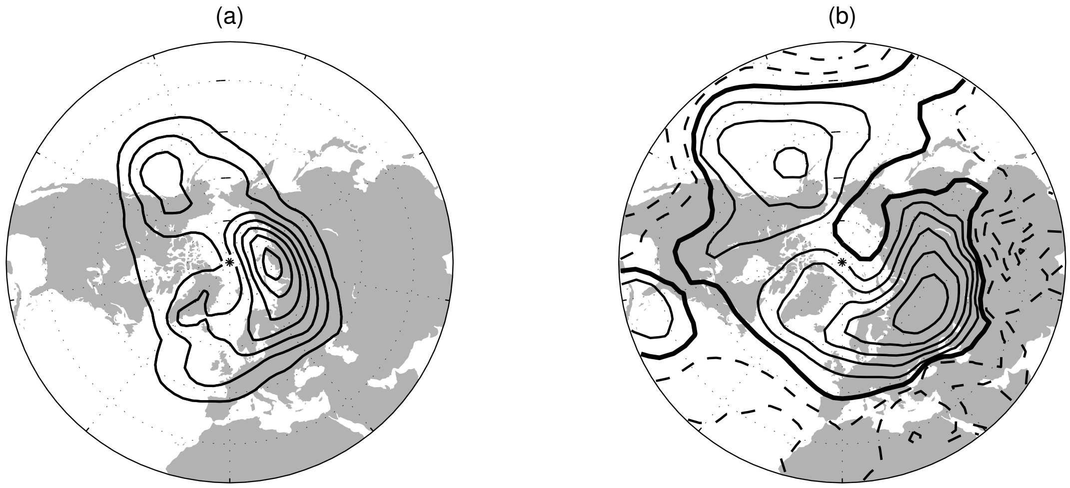


Figure 5.23: (a) Variance (contour interval 10 $(\mathrm{hPa})^2$) and (b) skewness (contour interval 0.2) of SLPA from GCM integration with quadrupled atmospheric CO_2.

5.6 Conclusions

In this chapter the NLPC analysis of sea level pressure anomaly data and 500mb geopotential height anomaly data from the Canadian Centre for Climate Modelling and Analysis coupled GCM have been considered. It was found that for data from a control integration at pre-industrial CO_2 concentrations, in both the SLPA and the Z_{500}A fields, the 1D NLPCA approximation was composed of three branches, and the corresponding time series was bimodal. The most frequently occupied branch describes the oscillatory amplification or attenuation of the climatological ridges over Europe, for the SLPA approximation, and over both Europe and North America for the Z_{500}A approximation. The less populated branch is occupied episodically, and strongly resembles the negative phase of the North Atlantic Oscillation. Mid-tropospheric circulation patterns in this regime are associated with split flow over southern Greenland. The fact that the independently-determined 1D NLPCA approximations found for SLPA and Z_{500}A are very similar provides strong evidence that these approximations represent actual structure in the data.

An analysis of SLPA data from a GCM integration with CO_2 levels at four times the pre-industrial concentrations indicated that the characteristic regimes of low-frequency variability did not change in structure, but in occupation frequency, as was suggested by Palmer (1999). The oscillatory regime became increasingly populated at the expense of the split-flow regime, which in fact became so depopulated that the branched structure of the NLPCA approximation disappeared.

In all cases, it was seen that the 1D NLPCA approximation describes coherent hemispheric-scale atmospheric variability that is consistent with both the geographical distribution of *variance* and of *skewness*. In the case of the control run integrations, the strong skewness in the region of the NAO is associated with the fact that variability is dominated by Gaussian oscillations with episodic excursions to a single strong anomaly pattern. This mechanism for the origin of skewness has been suggested in the literature by Nakamura and Wallace (1991).

NLPC analyses were also carried out on a Southern Hemisphere SLPA data set from the CCCma coupled GCM and on the Trenberth and Paolino Northern Hemisphere SLPA data set (1980). In neither case could robust 1D NLPCA approximations be found that differed from the 1D PCA approximation. In the case of the Southern Hemisphere GCM SLPA, this result is likely because of the strong zonal symmetry of the lower boundary, which is not as favourable to the formation of geographically-fixed circulation anomalies as is the Northern Hemisphere, with its strong topography and land-sea contrast. There are two possible reasons for the failure to detect strong nonlinear structure in the Trenberth and Paolino SLPA data. The first is simply that such structure is not there, and that the structure of the data is predominantly linear. The second is that the structure is there, but cannot be found because the records are short and data are of very poor quality in the polar regions, precisely the latitudes in which the nonlinear structure found in the GCM data is strongest. In fact, no data are reported on the latitude circles 75°N and 85°N or at the pole. A future study will consider the analysis of NCEP reanalysis data, which does not suffer from this deficiency in geographical coverage.

Chapter 6

Seven-Layer Networks for Discontinuous Projection and Expansion Functions

6.1 Introduction

Kramer's NLPCA allows the estimation of *continuous* functions s_f and $\mathbf{f}$ such that

$$\hat{\mathbf{X}}(t_n) = (\mathbf{f} \circ s_f)(\mathbf{X}(t_n)) \tag{6.1}$$

is the optimal (in the least-squares sense) approximation to $\mathbf{X}(t_n)$ by a *continuous* curve or surface. As was pointed out by Kirby and Miranda (1996) and by Malthouse (1998), and discussed in Chapter 2, the restriction to continuity of the projection and expansion maps precludes the detection and characterisation by NLPCA of P-dimensional structures not topologically equivalent to the unit cube in $\Re^P$. The simplest example illustrating this problem arises in the NLPC analysis of the unit circle embedded in $\Re^2$,

$$S = \{(\cos t, \sin t); t \in [0, 2\pi)\}. \tag{6.2}$$

The topology of the manifold parameterising this structure is S^1; the corresponding map $s_f : \Re^2 \mapsto \Re$ is discontinuous, because for any $0 < \epsilon << 1$ there will be a small open neighbourhood D on the circle about the point $(1, 0)$ such that $s_f(D) = [0, \epsilon) \cup (2\pi - \epsilon, 2\pi)$. There is no continuous map between the circle and the unit interval in $\Re$, because they are topologically inequivalent. Thus, NLPCA as formulated by Kramer cannot characterise the low-dimensional structure underlying a self-intersecting curve or surface. On the other hand, if the functions in (6.1) are allowed to be *discontinuous*, then there exists a class of parameter manifolds topologically inequivalent to the unit cube in $\Re^P$ such that the projection and expansion functions of an NLPCA approximation $\hat{\mathbf{X}}(t_n)$ can be approximated. This class includes topologies such as the circle (S^1) and torus $(S^1 \otimes S^1)$ but excludes the sphere (S^2). Because the parameterisation of the sphere is degenerate at the poles, it cannot be expressed as a simple discontinuous function from $\Re^2 \mapsto \Re^3$.

Given enough neurons in its single hidden layer, a three-layer feed-forward neural network with M input neurons and P output neurons can approximate to arbitrary accuracy any continuous function from $\Re^M$ to $\Re^P$. Such a network generally does a poor job approximating a discontinuous function between these spaces. If, however, the number of hidden layers is increased to two, each with nonlinear transfer functions, the network is much better able to approximate a discontinuous function. This fact suggests the generalisation of Kramer's 5-layer network to a 7-layer network, with two encoding and two decoding layers. In such a network, the first four layers approximate the (potentially discontinuous) function $\mathbf{s_f} : \Re^M \mapsto \Re^P$ and the last four layers the (potentially discontinuous) function $\mathbf{f} : \Re^P \mapsto \Re^M$; the 7-layer network is just the composition of these two functions. This chapter demonstrates that in fact a 4-layer neural network is better able to approximate a discontinuous function than is a 3-layer network, and investigates the application of the generalised 7-layer network to the NLPC analysis of an ellipse embedded in $\Re^2$.

6.2 Neural Network Approximations to Discontinuous Functions

To demonstrate the superior ability of a 4-layer neural network to approximate a discontinuous function relative to that of a 3-layer network, consider the N-point data sets:

$$\mathbf{X}(t_n) \quad = \quad (\cos 2\pi t_n, \sin 2\pi t_n)$$

$$Y(t_n) = \begin{cases} t_n + 0.5 & 1/N < t_n \leq 0.5 \\ t_n - 0.5 & 0.5 < t_n \leq 1 \end{cases} \tag{6.3}$$

where $t_n = 1/N, 2/N, ..., 1$ The map g relating $\mathbf{X}(t_n)$ and $Y(t_n)$ is discontinuous at $t_n = 0.5$; it is displayed in Figure 6.1. The functional relationship between $\mathbf{X}(t_n)$ and $Y(t_n)$ (with $N = 1000$) is modelled using two different feed-forward neural networks. The first, denoted NN1, has a single hidden layer with 13 neurons, and the second, NN2, has two hidden layers with 5 neurons in each. Hyperbolic tangents were used as the transfer functions in each of the hidden layers. The number of network parameters in NN1 and NN2 is 53 and 51 respectively. Each network was trained for 10000 iterations, starting from randomly-determined initial weights and biases. Because the data are noise-free, early-stopping was not employed in the training process.

Figures 6.2(a) and (b) display respectively the approximations to $Y(t_n)$ produced by neural networks NN1 and NN2. Note that NN2 is much better able to represent the discontinuity in the relationship between $\mathbf{X}(t_n)$ and $Y(t_n)$ than is NN1. The network NN1 spreads the discontinuity out over a number of points, and introduces Gibbs oscillations in its vicinity. The point of discontinuity is much better represented by NN2: the width over which the discontinuity is spread is substantially reduced, and the Gibbs oscillations are suppressed. This difference is not a result of differences in initial weights between the two cases, as an ensemble of independently trained networks (not shown) demonstrates that for both NN1 and NN2 the approximations are independent of initial parameter values. Nor is the improvement a reflection of the number of parameters in the model, as NN1 and NN2 have essentially the same number of parameters. Thus, NN2 is better able to approximate discontinuous functions than is NN1 because of differences in their *architecture*.

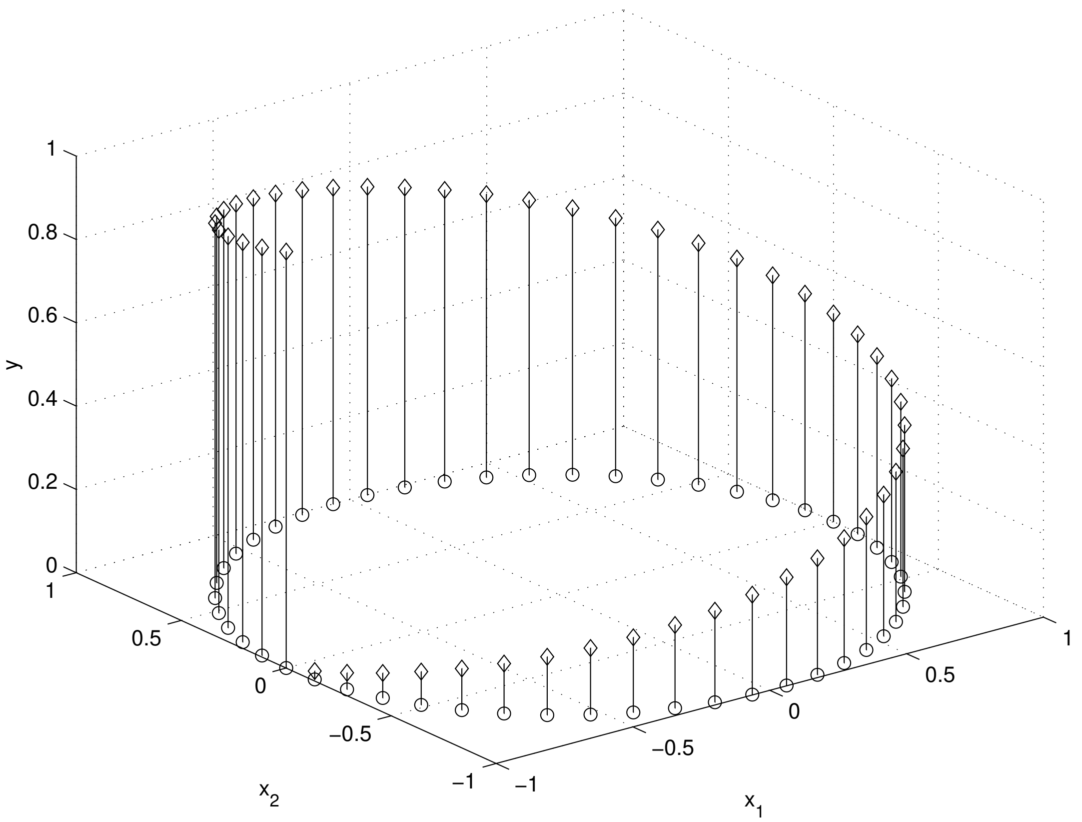


Figure 6.1: Plot of $Y(t_n)$ (diamonds) and $\mathbf{X}(t_n)$ (open circles) as defined by equation (6.3) with $N = 50$.

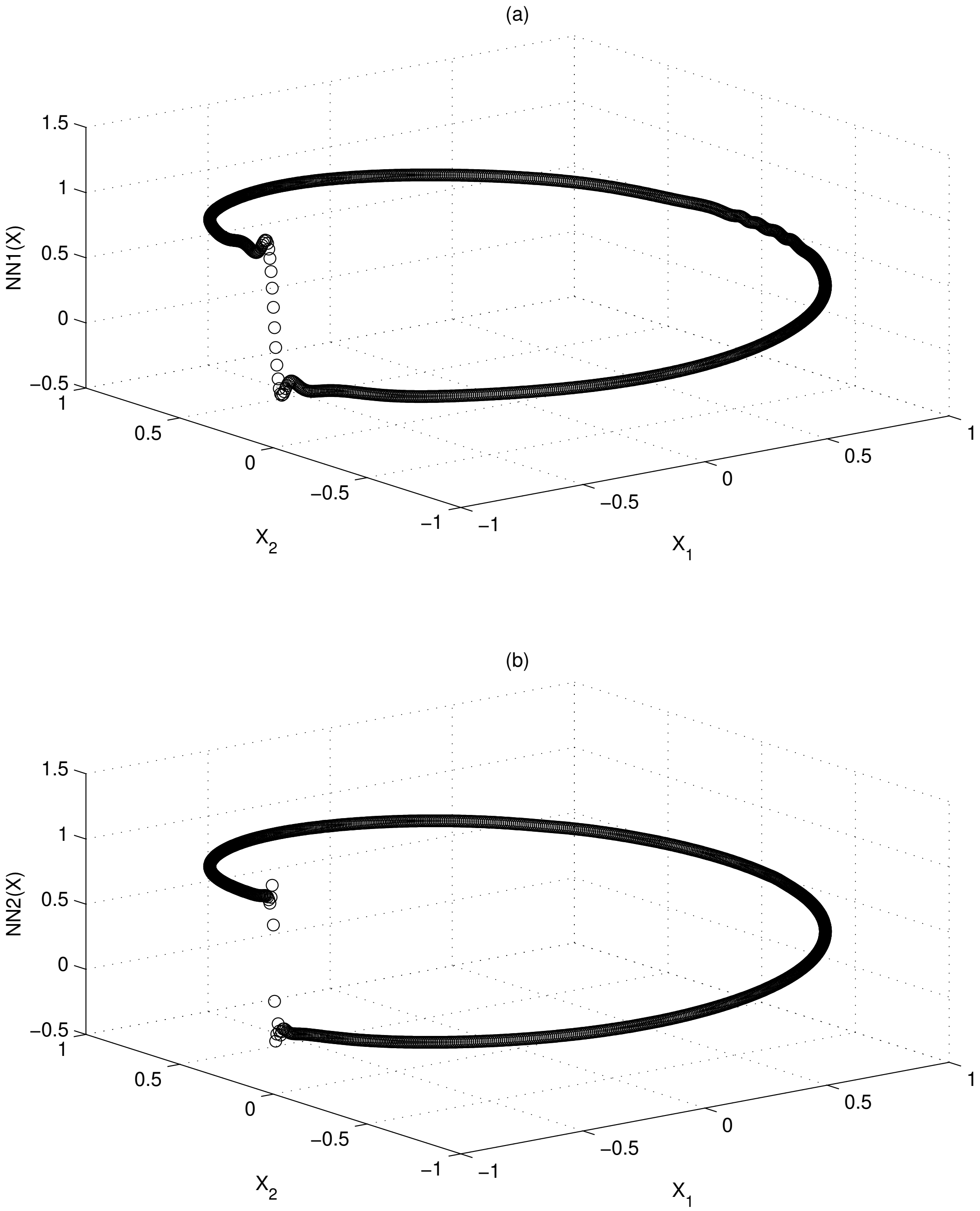


Figure 6.2: Neural network approximations of the functional relationship between data sets $\mathbf{X}(t_n)$ and $Y(t_n)$ defined in equation (6.3): (a) network with one hidden layer, (b) network with two hidden layers.

6.3 7-Layer NLPCA Network

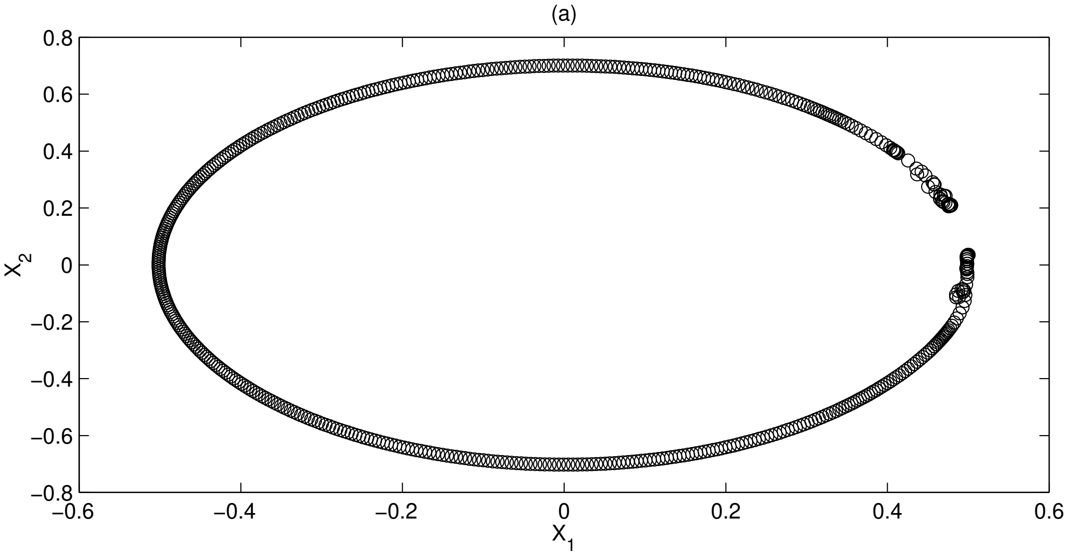
The improved ability of a 4-layer neural network over a 3-layer network to approximate discontinuous functions suggests that NLPCA carried out using a 7-layer network should be better able than that using a 5-layer network to approximate structures whose projection and expansion functions are discontinuous. In this section, this hypothesis is tested through consideration of the 1D NLPCA approximation of an ellipse in the plane:

$$\mathbf{X}(t_n) = (0.5\cos 2\pi t_n, 0.7\sin 2\pi t_n) \quad t_n = \frac{n-1}{N} \tag{6.4}$$

for $n = 1, ..., N-1$, where $N = 550$. As in the previous section, because these data are noise-free, no early stopping was used in constructing the model.

Figure 6.3 displays the NLPCA approximation $\hat{\mathbf{X}}(t_n)$ and the time series $\alpha(t_n) = s_f(\mathbf{X}(t_n))$ obtained using a 5-layer autoassociative neural network with $L = 13$ nodes in the mapping and demapping layer. Convergence to this approximation was extremely slow; the results shown are from a run in which 10^6 iterations were carried out in the training. As was discussed in Malthouse (1998), the approximation $\hat{\mathbf{X}}(t_n)$ displays high fidelity to the original data throughout the bulk of the domain (the FUV is 1.2×10^{-3}), but fails entirely over a range of points centred near (0.5,0.2). This set of points corresponds to the neighbourhood of the point at which the ideal projection and expansion functions are discontinuous because the topology of the manifold parameterising the curve is S^1. Note that the position of this point is arbitrary, and is determined randomly by the initial network parameters. The time series $\alpha(t_n)$ is unable to represent this discontinuity, displaying precisely the same overshoot as was observed in Figure 6.2(a). Because of the restriction of s_f and $\mathbf{f}$ to the space of continuous functions, the 5-layer autoassociative network cannot accurately approximate the ellipse.

On the other hand, Figure 6.4 displays the NLPCA approximation $\hat{\mathbf{X}}(t_n)$ and the corresponding time series $\alpha(t_n) = s_f(\mathbf{X}(t_n))$ obtained using a 7-layer autoassociative neural



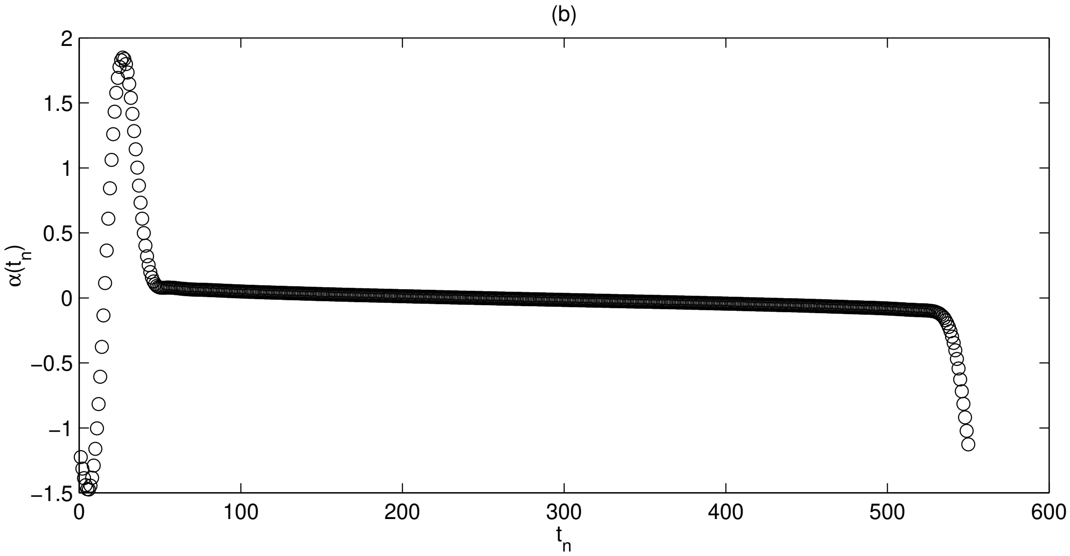


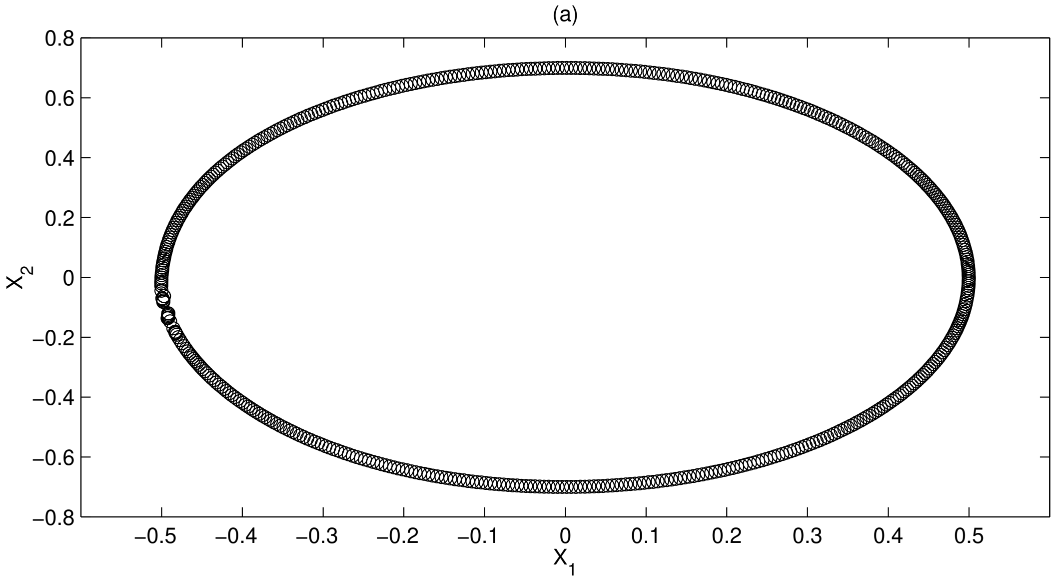
Figure 6.3: Results of 1D NLPC analysis of an ellipse using a 5-layer autoassociative neural network: (a) NLPCA approximation $\hat{\mathbf{X}}(t_n)$, (b) associated time series $\alpha(t_n) = s_f(\mathbf{X}(t_n))$ (note scale on y-axis is arbitrary).

network with 5 hidden neurons in each of the two mapping and demapping layers. As was the case with the 5-layer network, this network was trained for 10^6 iterations. This approximation displays great fidelity to the original data throughout the domain (the FUV is 2.4×10^{-5}), except for a rather small interval centred near (-0.5,-0.1). Again, this region corresponds to the neighbourhood of the point in which the ideal mapping and demapping functions are discontinuous; note that it is much smaller than the corresponding region obtained using a 5-layer network. As well, the FUV of the approximation from the 7-layer network is two orders of magnitude smaller than that of the approximation from the 5-layer network. The time series $\alpha(t_n)$ from the 7-layer network is a much better approximation of the discontinuous projection map than was that obtained from the 5-layer network.

The 5-layer autoassociative neural network considered above contained 107 parameters, while the 7-layer network contained 103. Thus, the superior performance of the 7-layer network is not due to its having a larger number of parameters. Both networks were trained for the same number of iterations. Other training runs (not shown) display the same superiority of the 7-layer network to the 5-layer network; this result is insensitive to the choice of initial model parameters. Thus, the 7-layer network is superior to the 5-layer network because of differences in architecture; the presence of two mapping and demapping layers allows the network to approximate a broader class of projection and expansion functions than that open to Kramer's original network.

6.4 Conclusions

It has been demonstrated that a 7-layer generalisation of Kramer's 5-layer network, in which two mapping and demapping layers are used, displays a marked superiority in its ability to model low-dimensional structure whose corresponding projection and expansion



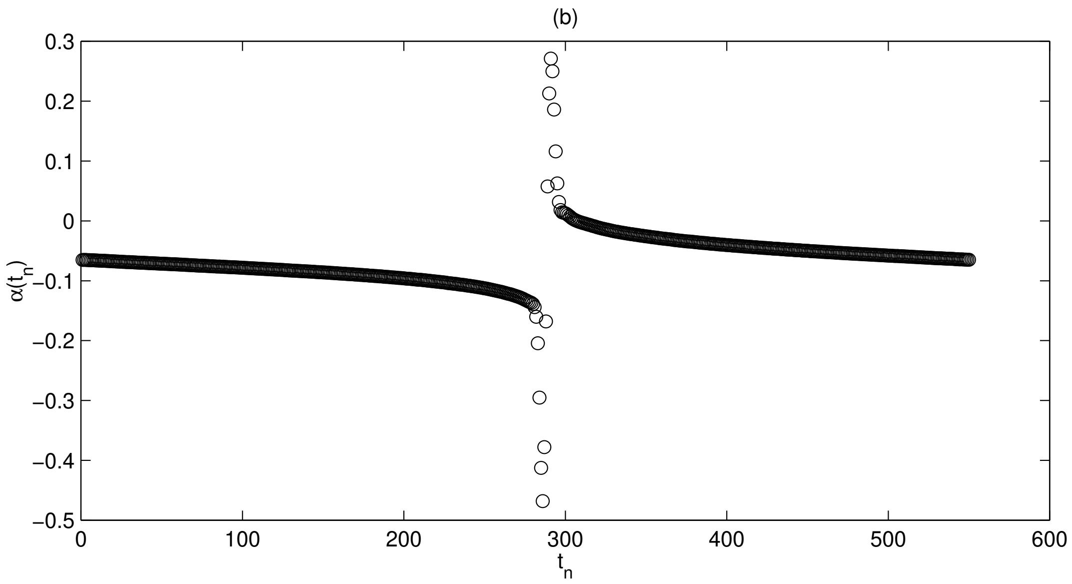


Figure 6.4: As in Figure 6.3, but for NLPCA performed using a 7-layer autoassociative neural network.

functions are discontinuous. This superiority was shown through the NLPC analysis of an ellipse in the plane, whose projection and expansion functions are discontinuous because of the S^1 topology of the manifold parameterising the ellipse.

Another solution to this problem was presented by Kirby and Miranda (1996), who introduced the concept of "circular nodes" in neural networks. Because standard nodes in neural networks map to the real line, Kirby and Miranda noted that such nodes are unable to encode "angular" information. In other words, they cannot map continuously to a space with S^1 topology. Kirby and Miranda introduced the idea of coupling pairs of nodes such that their output is constrained to fall on the unit circle. These coupled nodes can be treated as a single abstract node that can encode angular information. This coupling requires modification of the backpropagation algorithm (Appendix A); such a modification is presented by Kirby and Miranda.

There are two problems with Kirby and Miranda's approach. First, the method is somewhat difficult to implement, as it requires a modification of the backpropagation algorithm and thus cannot be carried out using a regular, commercial neural network package. Second, the use of circular nodes presumes the existence of periodic structure in the data. The 7-layer network presented here has neither of these difficulties. However, Kirby and Miranda's method allows an exact characterisation of maps to S^1, whereas the 7-layer generalisation of Kramer's network produces only approximations. It is not clear that this lack of exactness is a problem in practice, as neural networks can only ever produce approximate models to data.

Few large-scale climatic phenomena display the strict periodicity of the data set considered in this section. Climate variability is much more irregular, and examples of strictly periodic limit cycles are rare. Examples of periodic variability include the seasonal cycle (including annual and higher harmonics), the tides, and the diurnal cycle; all of these are periodic because they arise from external, astronomical forcings which are strongly periodic. These signals are of well-known frequency, as are their harmonics arising from through nonlinear rectification (Huang and Sardeshmukh, 1999). If the annual cycle and its harmonics are stationary in time, they can be removed from the data using harmonic analysis. If stationarity does not obtain, more complicated techniques such as wavelet analysis could be used instead. Thus, the inability of Kramer's 5-layer autoassociative neural network to model strictly periodic variability is not expected to be a significant liability in the analysis of climate data. The fact that this problem can be addressed by generalising Kramer's network is an interesting theoretical result. In practice, its application will probably be limited to data thought to have been generated by a highly periodic system.

Chapter 7

Summary and Conclusions

7.1 Summary

Principal component analysis is a tool of great utility in the analysis of climate data. The phase space of a typical climatic data set has a dimensionality in the range from hundreds to thousands, which makes straightforward visualisation impossible. PCA finds the ordered set of axes in the phase space which provides the most efficient linear description of the data, in that the projection of the data into the space spanned by the first P axes is the optimal P-dimensional linear approximation to the data (in a least-squares sense). It is thus a tremendously useful algorithm for the reduction of data dimensionality.

There is, however, no *a priori* reason to believe that any lower-dimensional structure underlying a multivariate dataset is linear, in that it is optimally described by a set of orthogonal axes. Indeed, it is a basic result of the theory of dynamical systems that if the dynamical system

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) \tag{7.1}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \tag{7.2}$$

possesses a K-dimensional stable attractor Γ (where K is not necessarily an integer), then in general the description of the manifold Γ by a set of Cartesian coordinates (ie, orthogonal coordinate axes) requires the dimension of the embedding space to be at least $2K + 1$. PCA can determine an appropriate embedding space for a general low-dimensional structure, but it cannot provide the most efficient description of this surface.

In an effort to circumvent this limitation of PCA, Kramer (1991) introduced a nonlinear generalisation of PCA, which he denoted Nonlinear Principal Component Analysis. Implemented using a 5-layer feed-forward neural network, NLPCA attempts to find functions $\mathbf{s_f} : \Re^M \mapsto \Re^P$ and $\mathbf{f} : \Re^P \mapsto \Re^M$ such that the approximation $\hat{\mathbf{X}}(t_n) = (\mathbf{f} \circ \mathbf{s_f})(\mathbf{X}(t_n))$ is the optimal P-dimensional nonlinear approximation to the M-dimensional data set $\mathbf{X}(t_n)$, in a least-squares sense. If the functions $\mathbf{s_f}$ and $\mathbf{f}$ are constrained to be linear, this approach reduces to PCA. This thesis presents the first systematic application of NLPCA to the analysis of climate data. A summary of the results follows.

1. The similarities and differences between PCA and NLPCA were discussed. Both PCA and NLPCA can be characterised as variational problems for detecting lower-dimensional structure in multivariate data sets: PCA finds linear structure, NLPCA can find more general nonlinear structure. The P-dimensional PCA approximation to a data set is the sum of its first P one-dimensional approximations; in NLPCA, the situation is more complicated. A single P-dimensional surface determined using an autoassociative neural network with P neurons in the bottleneck layer is said to be a nonmodal approximation, while the sum of the first P one-dimensional approximations is said to be a modal approximation; these two approaches are generally distinct. A degeneracy in the parameterisation of the surface determined by a P-dimensional nonmodal NLPCA complicates interpretation of the parameterisation for $P \geq 2$, and generally requires the consideration of a secondary feature extraction problem; thus, modal NLPCA for the analysis of real climate data is preferred. Both 1D PCA and NLPCA modes correspond to a single time series, however, unlike PCA, a 1D NLPCA mode does not have a unique spatial pattern. In fact, the approximation is characterised by a sequence of spatial patterns, which

may be visualised cinematographically. Both PCA and NLPCA partition variance in that the sum of the total variance of a P-dimensional approximation with the total variance of the residual equals the variance of the original data. This result can be proved for PCA, but so far remains an empirical result for NLPCA.

2. Implementation of NLPCA was considered in some detail. The NLPCA algorithm is carried out using a 5-layer autoassociative feed-forward neural network model. Because NLPCA involves the solution of a nonlinear variational problem, no analytic formula for the solution exists, and iterative function minimisation procedures must be used. This minimisation process is referred to as "training". An issue of primary importance is the avoidance of overfitting; an NLPCA approximation should robustly characterise lower-dimensional structure in the data. Overfitting was avoided through the use of an early stopping technique in which a randomly-selected portion of the original data is set aside and not used to fit the model parameters. The performance of the model over this withheld data set is monitored as training progresses; if the model performance over this validation set is inferior to the performance over the training data, the model is discarded. An ensemble of models is constructed, each of which satisfy the robustness criteria described above. These are referred to as the "candidate models". The number of neurons in the encoding and decoding layers of the autoassociative network and the number of iterations used to train the model are adjusted until a sufficient number of similar candidate models is obtained, at which point a representative member is extracted and said to be the NLPCA approximation to the data.

3. A preliminary investigation of NLPCA was carried out using data sampled from the Lorenz attractor, to which random noise was added. It was found that at low to moderate noise levels NLPCA was able to produce robust approximations

that were more characteristic of the data, and explained higher percentages of the variance, than were those produced by PCA. In the limit of no noise, the 1D NLPCA approximation explained 76% of the total variance while the corresponding PCA approximation explained 60%. The NLPCA mode was able to describe covariability between two uncorrelated but dependent variables, which PCA cannot do. A 2D nonmodal NLPCA approximation explained 99.5% of the variance, while the 2D PCA approximation explained 95%. As the noise level increased, the improvement of NLPCA over PCA decreased, in terms of the percentage of variance explained. At high noise levels, in which the structure of the Lorenz data was entirely obscured, NLPCA could not produce robust approximations that differed from those obtained by PCA.

4. A tropical Pacific Ocean sea surface temperature data set was analysed by NLPCA. The 1D NLPCA approximation to these data describes average ENSO variability. Unlike the 1D PCA approximation, which also describes average ENSO variability, the 1D NLPCA approximation is able to characterise the asymmetry in SST spatial patterns between average El Niño and La Niña events. The distribution of SST is skewed toward positive anomalies in the eastern Pacific and toward negative anomalies in the western Pacific, and the 1D NLPCA approximation is able to characterise this distribution of skewness. The second NLPCA mode is also related to ENSO, and seems to characterise differences between individual events. It is particularly active in the period after 1977, a time which has been noted in a number of studies as corresponding to a shift in ENSO variability. A 2D nonmodal NLPCA approximation to the SST was also determined. A secondary PCA analysis in the 2D space of variables parameterising this surface indicated that the

variability described by this approximation contains essentially the same information as the first two NLPCA modes. An NLPC analysis of tropical Indo-Pacific sea level pressure was also carried out. The first mode was found to correspond to the ENSO signal in SLP, and characterised asymmetries in the SLP between El Niño and La Niña events. No robust nonlinear structure could be detected in the residuals from the first NLPCA mode; within the constraints imposed by the quantity of data available, SLP variability is linear beyond the first NLPCA mode.

5. The first NLPCA mode of monthly-averaged Northern Hemisphere SLPA from the Canadian Centre for Climate Modelling and Analysis coupled GCM was found to partition the data into two distinct populations. The 1D NLPCA approximation had a three-branched structure, and the distribution associated time series was strongly bimodal. One branch (Branch 1), associated with the larger peak of the distribution of the time series, corresponded to a standing oscillation with anomalies of opposite sign over the polar region and the midlatitudes, strongly resembling the Arctic Oscillation. Most of the data projected onto this branch. A second branch (Branch 2), which corresponded to the smaller peak of the time series PDF, was only occupied episodically, and strongly resembled the negative phase of the North Atlantic Oscillation. The Branch 1 signal in 500 mb geopotential height composites (based on the SLPA analysis) described alternating amplification and attenuation of the climatological ridge over Europe, while Branch 2 described strongly split flow over Greenland. An analysis of the SLP skewness indicates that there are strong positive and negative local extrema in skewness in the same locations as the positive and negative extrema of the Branch 2 anomaly patterns. These extrema in skewness thus seem to arise because of the combination of a standing oscillation displaying Gaussian variability with episodic occurrences of a strongly anomalous

circulation, a possibility that has been suggested by other authors in the past.

An NLPC analysis of the 500mb height anomaly field itself resulted in a branched approximation that was very similar to that obtained from the SLPA field; the primary difference between the two is that the former corresponds to anomaly fields that are somewhat more hemispheric in extent. Thus, the two-regime structure identified using NLPCA appears to be equivalent barotropic in nature.

Finally, the results of an analysis of SLPA from a GCM run with CO_2 concentrations at four times the pre-industrial levels indicated that the oscillatory branch of the control NLPCA approximation was largely unchanged but that the split-flow branch was substantially depopulated. This behaviour is consistent with the suggestion by Palmer (1999) that the climate response to greenhouse forcing will not be changes in the structure of characteristic circulation regimes, but in their occupation frequencies.

6. Because Kramer's 5-layer autoassociative neural network can only find continuous projection and expansion functions, a 7-layer generalisation was suggested for the analysis of data sets where the expansion and projection functions are discontinuous. Such a data set is an ellipse, because the manifold parameterising the low-dimensional approximation is topologically different than the unit interval. It was found that the NLPCA approximation produced by a 7-layer autoassociative neural network was substantially better than that produced by a 5-layer network, because the former was much better able to approximate discontinuous projection and expansion functions. It was demonstrated that this improvement was due to the different architectures of the two networks, and not to a difference in the number of model parameters.

7.2 Conclusions

Nonlinear Principal Component Analysis has been demonstrated to be a useful tool for the analysis of climate data. Where PCA characterises the variance in a multivariate data set, NLPCA is able to also characterise higher order moments of variability. This thesis has introduced NLPCA to the study of climate data, but there remains work to be done. The model building procedures described in Chapter 2 retain an element of subjectivity. It would be useful to develop an automated, objective technique for building NLPCA approximations; such a methodology could perhaps use a sophisticated regularisation technique such as Generalised Cross Validation (Yuval, 1999). Useful generalisations of NLPCA might also be developed through modifications of the cost function to include constraints on the approximation. An example would be a simple structure constraint of the form used in rotated PCA analysis (Richman,1986). A second modification of the cost function to ensure self-consistency of the NLPCA model is suggested in Rico-Martinez et al. (1996). Kirby and Miranda (1999) suggest a number of other possible constraints. As well, NLPCA could be used in a nonlinear generalisation of Singular Systems Analysis (SSA), which is simply PCA applied to a time series expressed in delay coordinate space (Broomhead and King, 1986; von Storch and Zwiers,1999). Finally, analytic demonstrations of features of NLPCA discovered empirically, such as the partitioning of variance, are lacking.

The analysis of large multivariate datasets, either observations or GCM output, is an important activity for the understanding of climate variability. Nonlinear Principal Component Analysis will not replace traditional PCA, because it is more difficult to implement. However, I believe that NLPCA may well become an important addition to the geophysical statistician's toolbox, and will provide important insight into the variability of the climate system.

Appendix A

Neural Networks

As is described in detail by Bishop (1995) and by Hsieh and Tang (1998), a feed-forward neural network is a non-parametric statistical model used to estimate (generally nonlinear) functional relations between two data sets, $\mathbf{X}(t_n) \in \Re^S$ and $\mathbf{Z}(t_n) \in \Re^T$. The neural network is composed of a series of parallel layers, each of which contains a number of processing elements, or *neurons*, such that the output of the ith layer is used as input to the $(i+1)$th. If $y_j^{(i)}$ is the output of the jth neuron of the ith layer, then

$$y_k^{(i+1)} = \sigma^{(i+1)} \left(\sum_j w_{jk}^{(i+1)} y_j^{(i)} + b_k^{(i+1)} \right) \tag{A.1}$$

is the output of the kth neuron of the $(i+1)$th layer. The elements of the arrays $w_{jk}^{(i+1)}$ are referred to as the *weights*, and those of the vectors $b_k^{(i+1)}$ as the *biases*. The *transfer function* characterising the $(i+1)$th layer is denoted $\sigma^{(i+1)}$; it may be linear or nonlinear. The first, or input layer, receives the values of the data presented to the network; its transfer function is simply the identity map $\sigma_I : x \mapsto x$. The famous flexibility of neural networks comes from the use of nonlinear transfer functions (typically hyperbolic tangents) in some or all of the remaining layers. An important result due to Cybenko (1989) is that a 3-layer neural network with S input neurons, hyperbolic tangent transfer functions in the second layer and linear transfer functions in the third layer of T neurons can approximate to arbitrary accuracy any continuous function from $\Re^S$ to $\Re^T$, if the number of neurons in the second layer is sufficiently large.

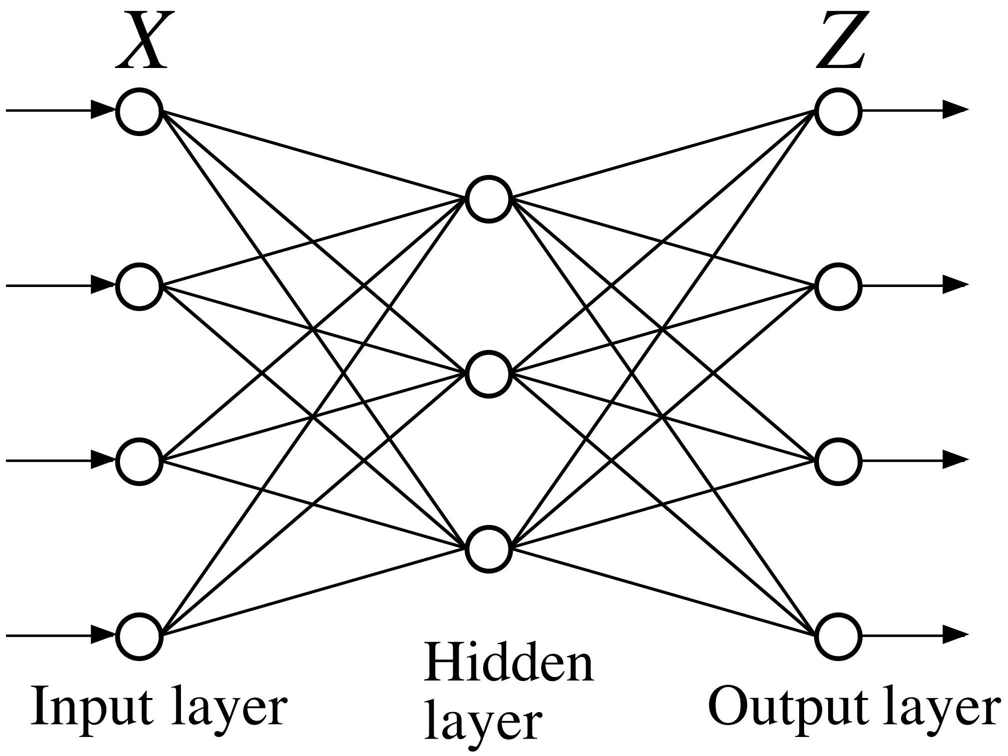


Figure A.1: Diagrammatic representation of neural network with input data $\mathbf{X}$ and output data $\mathbf{Z}$.

Neural networks are often represented diagrammatically, with open circles representing the neurons and straight lines the weights, as is illustrated in Figure A.1. These diagrams are meant to be suggestive of biological neuronal systems, reflecting the origin of neural network theory in the context of artificial intelligence research.

Feed-forward neural networks as described above are fit to data, or *trained*, as follows. Suppose it is desired to fit the data $\mathbf{X}(t_n) \in \Re^S$ to the data $\mathbf{Z}(t_n) \in \Re^T$. Denoting the network as $\mathcal{N}$, the weights and biases (referred to collectively as μ) are adjusted until the cost function

$$J = < ||\mathbf{Z} - \mathcal{N}(\mathbf{X};\mu)||^2 > \tag{A.2}$$

is minimised. That is, parameters μ_{min} are determined such that

$$0 = \left.\frac{\partial J}{\partial \mu_i}\right|_{\mu=\mu_{\text{min}}} \tag{A.3}$$

for each parameter μ_i. In the above, the angle brackets $< . >$ denote averaging over time and $||.||$ denotes the L^2-norm. For a network in which all transfer functions $\sigma^{(i)}$ are linear, equation (A.3) can be expressed as a simple matrix equation which admits an analytic solution, and the approach simply reduces to multivariate regression. When the transfer functions are nonlinear, equation (A.3) does not reduce to a simple matrix equation, and the cost function J must be minimised numerically.

The minimisation of the cost function was carried out using a conjugate-gradient algorithm (Press et al., 1992). At each step of this algorithm, the gradient of J with respect to the parameters over which it is being minimised must be evaluated. It is easy to show that for an I-layer neural network,

$$\frac{\partial J}{\partial w_{jk}^{(I)}} = -2 < \epsilon_j \sigma'^{(I)}(s_j^{(I)}) y_k^{(I-1)} > \tag{A.4}$$

where

$$s_j^{(i)} = \sum_l w_{lj}^{(i)} y_l^{(i-1)} + b_j^{(i)}, \tag{A.5}$$

$$\epsilon(t_n) = \mathbf{Z}(t_n) - \mathcal{N}(\mathbf{X}(t_n); \mu) \tag{A.6}$$

and the prime represents differentiation. Thus, the gradient of the cost function with respect to the weights of the output layer can be evaluated exactly at every step of the minimisation algorithm. The same is true for the gradient of J with respect to the biases of the output layer.

Furthermore, it can be shown that

$$\frac{\partial J}{\partial w_{jk}^{(I-1)}} = -2 \left\langle \left[\sum_l \epsilon_l \sigma'^{(I)}(s_l^{(I)}) w_{lj}^{I} \right] \sigma'^{(I-1)}(s_j^{(I-1)}) y_k^{(I-2)} \right\rangle \tag{A.7}$$

and

$$\frac{\partial J}{\partial w_{jk}^{(I-2)}} = -2\left\langle \sum_m \left[\sum_l \epsilon_l \sigma'^{(I)}(s_l^{(I)}) w_{lm}^{(I)} \sigma'^{(I-1)}(s_m^{(I-1)}) \right] w_{mj}^{(I-1)} \sigma'^{(I-2)}(s_j^{(I-2)}) y_k^{(I-3)} \right\rangle \quad \text{(A.8)}$$

Note that in the expression for the derivatives of J with respect to the weights at the $(i-1)$th layer, quantities used in the calculation of the derivative with respect to the weights of the ith layer appear (in the square brackets). An efficient algorithm for the evaluation of the gradient of J then is to calculate the quantities:

$$d_j^{(I)} = \epsilon_j \sigma'^{(I)}(s_j^{(I)}) \quad \text{(A.9)}$$

$$d_j^{(I-1)} = \sum_l d_l^{(I)} w_{lj}^{(I)} \sigma'^{(I-1)}(s_j^{(I-1)}) \quad \text{(A.10)}$$

$$d_j^{(I-2)} = \sum_l d_l^{(I-1)} w_{lj}^{(I-1)} \sigma'^{(I-2)}(s_j^{(I-2)}) \quad \text{(A.11)}$$

and so forth, up to $d_j^{(1)}$. The gradient of J with respect to the weights is simply then given by

$$\frac{\partial J}{\partial w_{jk}^{(i)}} = -2\left\langle d_j^{(i)} y_k^{(i-1)} \right\rangle \quad \text{(A.12)}$$

Similar equations hold for the gradient of J with respect to the biases. This algorithm for evaluating the gradient of the cost function with respect to the neural network parameters is referred to as *backpropagation*. Note that backpropagation allows the exact evaluation of the gradient at each step of the conjugate gradient algorithm.

Appendix B

Principal Curves and Surfaces

First consider Principal Curves, which are the 1D version of PCS. Following Hastie and Stuetzle (1989), let $\mathbf{X}$ be a random vector in $\Re^M$, the distribution of which, denoted by h, has finite second moments. As usual, assume $E(\mathbf{X}) = \mathbf{0}$, without loss of generality. Let the map $\mathbf{f} : \Re \to \Re^M$ be C^∞, unit speed (that is, $||\mathbf{f}'|| = 1$), and non-self-intersecting (that is, $\lambda_1 \neq \lambda_2 \Rightarrow \mathbf{f}(\lambda_1) \neq \mathbf{f}(\lambda_2)$). The projection function $s_f : \Re^M \to \Re$ is defined such that

$$s_f(\mathbf{x}) = \sup_{\lambda}\{\lambda : ||\mathbf{x} - \mathbf{f}(\lambda)|| = \inf_{\mu} ||\mathbf{x} - \mathbf{f}(\mu)||\}. \tag{B.1}$$

That is, of those λ such that $\mathbf{f}(\lambda)$ are the points closest to $\mathbf{x}$, $s_f(\mathbf{x})$ is the largest. Hastie and Stuetzle defined $\mathbf{f}$ to be *self-consistent* if the expectation value of all the points projecting onto a certain point on the curve $\mathbf{f}$ is that point itself, ie, if

$$E_h(\mathbf{X}|s_f(\mathbf{X}) = \lambda) = \mathbf{f}(\lambda). \tag{B.2}$$

where $E_h(.)$ denotes expectation over the distribution h. If $\mathbf{f}$ is self-consistent, then it is a *principal curve*. Hastie and Stuetzle proved that if $\mathbf{f}$ is constrained to be linear, and if it is self-consistent, then it is a principal component. Furthermore, in the space of curves through the data, a principal curve is a critical point of the distance function, in the following sense. Let $\mathbf{f}$ be a principal curve and $\mathbf{g}$ be an arbitrary smooth function from $\Re$ to $\Re^M$, and define $\mathbf{f}_t = \mathbf{f} + t\mathbf{g}$. Defining the distance function

$$d(\mathbf{x}, \mathbf{f}_t) = ||\mathbf{x} - \mathbf{f}_t(s_{f_t}(\mathbf{x}))|| \tag{B.3}$$

and $D^2(h, \mathbf{f}_t) = E_h d^2(\mathbf{X}, \mathbf{f}_t)$, then

$$\frac{d}{dt} D^2(h, \mathbf{f}_t)\Big|_{t=0} = 0 \tag{B.4}$$

Equation (B.4) is a formal expression of the idea that the principal curve passes through the "middle" of the data, and is the clearest point of connection between PCS and NLPCA.

A very useful fact pointed out by LeBlanc and Tibshirani (1994) is that principal curves partition variance such that

$$\sum_{j=1}^{M} \mathrm{var}(X_j) = \sum_{j=1}^{M} \mathrm{var}(f_j(s_f(\mathbf{X}))) + \sum_{j=1}^{M} \mathrm{var}(X_j - f_j(s_f(\mathbf{X}))) \tag{B.5}$$

It is therefore sensible to describe the principal curve $\mathbf{f}$ as explaining a certain fraction of the variance of the random vector $\mathbf{X}$.

The construction of principal curves presented above presupposes knowledge of the distribution h of the random vector $\mathbf{X}$; this is not usually known for real data sets. Hastie and Stuezle present an iterative algorithm, involving the use of locally-weighted running-lines smoothers, to determine the principal curve of a data set.

Hastie and Stuetzle denoted the generalisation of principal curves to two dimensions as *principal surfaces.* As with principal curves, given a two-dimensional surface $\mathbf{f} \in \Re^2$, a projection index $\mathbf{s}_f : \Re^M \to \Re^2$ is defined such that $\mathbf{s}_f(\mathbf{x})$ is the point on $\mathbf{f}$ closest to $\mathbf{x}$; $\mathbf{f}$ is a principal surface if

$$E(\mathbf{X}|\mathbf{s}_f(\mathbf{X}) = \lambda) = \mathbf{f}(\lambda) \tag{B.6}$$

Hastie and Stuetzle did not discuss principal surfaces in much detail; they did mention that preliminary numerical investigations indicated that principal surfaces share many properties with principal curves. Principal surfaces can be further generalised to surfaces of dimensionality higher than two; LeBlanc and Tibshirani (1994) constructed a piecewise linear generalisation they denoted *adaptive principal surfaces.*

Hastie and Stuetzle provided rigorous proofs of PCS results for only the 1D case, although LeBlanc and Tibshirani (1994) and Malthouse (1998) consider higher-dimensional generalisations.

A hybrid approach to NLPCA using both Kramer's autoassociative neural network and PCS has been proposed by Dong and McAvoy (1995). Their method involved a preliminary pre-processing of the data by PCS, followed by an NLPC analysis of the processed data. The logic behind this approach was that PCS possesses a better theoretical grounding than does Kramer's NLPCA, but does not produce a simple *model* of the data in that when presented with a new data point, there is no simple algorithm to determine its PCS approximation. Kramer's NLPCA, on the other hand, does produce such a model of the data. Dong and McAvoy recognised that by combining the two methods, the benefits of both can be realised. However, this approach is more cumbersome than Kramer's NLPCA alone, and was thus not implemented in this work.

Appendix C

Symmetric and Anti-symmetric Components of Composites

Consider a spatial field $\mathbf{Y}(t_n), n = 1, ..., N$, which is composited using a time series $X(t_n)$ as follows. Two subsets of time, $t_n^{(+)}$ and $t_n^{(-)}$, are defined by

$$\{t_n^{(+)}\} = \{t_n : X(t_n) > c\} \tag{C.1}$$

$$\{t_n^{(-)}\} = \{t_n : X(t_n) < -c\} \tag{C.2}$$

where c is some threshold: in our case, it is one standard deviation of $X(t_n)$. The positive and negative composites of $\mathbf{Y}(t_n)$, $\mathbf{Y}^{(+)}$ and $\mathbf{Y}^{(-)}$, are simply defined as the respective averages over $\{t_n^{(+)}\}$ and $\{t_n^{(-)}\}$:

$$\mathbf{Y}^{(+)} = <\mathbf{Y}>_+ \tag{C.3}$$

$$\mathbf{Y}^{(-)} = <\mathbf{Y}>_- \tag{C.4}$$

Maps of $\mathbf{Y}^{(+)}$ and $\mathbf{Y}^{(-)}$, where $\mathbf{Y}(t_n)$ is SSTA and $X(t_n)$ is the NDJ-averaged Niño 3.4 index are shown in Figures 4.6(a) and (b), respectively. In general, the spatial patterns of $\mathbf{Y}^{(+)}$ and $\mathbf{Y}^{(-)}$ differ by more than a sign.

It is desired to determine the symmetric and anti-symmetric (under a change of sign in $X(t_n)$) components of $\mathbf{Y}^{(+)}$ and $\mathbf{Y}^{(-)}$. To address this question, assume the minimal nonlinear model for the dependence of $\mathbf{Y}(t_n)$ on $X(t_n)$:

$$\mathbf{Y}(t_n) = \mathbf{a}^{(0)} + \mathbf{a}^{(1)}X(t_n) + \mathbf{a}^{(2)}X(t_n)^2 + \epsilon_n \tag{C.5}$$

where ϵ_n is a vector noise process, assumed to satisfy

$$<\epsilon>=<\epsilon>_+=<\epsilon>_-= \mathbf{0} \tag{C.6}$$

The validity of this approximation depends both on the validity of the model (C.5) and on the length of the records, $\{t_n\}$, $\{t_n^{(+)}\}$, and $\{t_n^{(-)}\}$. It can be assumed without loss of generality that both $\mathbf{Y}(t_n)$ and $X(t_n)$ are zero-centred in time. This implies that

$$\mathbf{0} = \mathbf{a}^{(0)} + \mathbf{a}^{(2)} < X^2 > \tag{C.7}$$

and so the model can be rewritten as

$$\mathbf{Y}(t_n) = \mathbf{a}^{(1)} X(t_n) + \mathbf{a}^{(2)}(X(t_n)^2 - < X^2 >) + \epsilon_n \tag{C.8}$$

The vector $\mathbf{a}^{(1)}$ is the field pattern anti-symmetric under a change of sign in X, while $\mathbf{a}^{(2)}$ is the field pattern symmetric under such a change of sign. They will be referred to, respectively, as the anti-symmetric and symmetric components of the composite.

Clearly, by the definition of the composite maps,

$$\mathbf{Y}^{(+)} = \mathbf{a}^{(1)} < X >_+ + \mathbf{a}^{(2)}(< X^2 >_+ - < X^2 >) \tag{C.9}$$

$$\mathbf{Y}^{(-)} = \mathbf{a}^{(1)} < X >_- + \mathbf{a}^{(2)}(< X^2 >_- - < X^2 >) \tag{C.10}$$

This is a linear equation which can easily be solved to yield:

$$\mathbf{a}^{(1)} = D^{-1}[(< X^2 >_- - < X^2 >)\mathbf{Y}^{(+)} - (< X^2 >_+ - < X^2 >)\mathbf{Y}^{(-)}] \tag{C.11}$$

$$\mathbf{a}^{(2)} = D^{-1}[- < X >_- \mathbf{Y}^{(+)} + < X >_+ \mathbf{Y}^{(-)}] \tag{C.12}$$

where

$$D = < X >_+ (< X^2 >_- - < X^2 >) - < X >_- (< X^2 >_+ - < X^2 >) \tag{C.13}$$

Figure 4.6(c) displays $\mathbf{a}^{(2)}$ for tropical Pacific SSTA composited according to the Niño 3.4 index. A map of $\mathbf{a}^{(1)}$ (not shown) looks very much like SSTA EOF mode 1 (Figure 4.6(a)); the spatial correlation between these is 0.975.

Hoerling et al. (1997) considered the linear combinations $\mathbf{Y}^{(+)} - \mathbf{Y}^{(-)}$ and $\mathbf{Y}^{(+)} + \mathbf{Y}^{(-)}$ and denoted them the linear and nonlinear responses of $\mathbf{Y}$ to X, respectively. The above analysis shows this identification is appropriate only in the special case that $< X >_- = - < X >_+$ and $< X^2 >_- = < X^2 >_+$. This is certainly not true in general, although for the case they considered, in which $X(t_n)$ was an SST index similar to Niño 3.4, it is a fairly good approximation.

In principle, one could use the technique described above to fit the more general model

$$\mathbf{Y}(t_n) = \sum_{k=0}^{K} \mathbf{a}^{(k)} X(t_n)^k + \epsilon_n \tag{C.14}$$

by stratifying the data into $K + 1$ subsets. Presumably, however, as K increases, so does the sampling variability associated with decreasing validity of approximations (C.6).

Bibliography

Baldwin, M. P. and Dunkerton, T. J. (1999). Propagation of the Arctic Oscillation from the stratosphere to the troposphere. *J. Geophys. Res.*, in review.

Barnston, A. G. (1994). Linear statistical short-term climate predictive skill in the northern hemisphere. *J. Climate*, 7:1513–1564.

Barnston, A. G., Glantz, M. H., and He, Y. X. (1999). Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997-98 El Niño episode and the 1998 La Niña onset. *Bull. Amer. Met. Soc.*, 80:217–243.

Barnston, A. G. and Livezey, R. E. (1987). Classification, seasonality, and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.*, 115:1083–1126.

Barnston, A. G. and Ropelewski, C. F. (1992). Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, 5:1316–1345.

Barnston, A. G., van den Dool, H. M., Zebiak, S. E., Barnett, T. P., Ji, M., Rodenhuis, D. R., Cane, M. A., Leetma, A., Graham, N. E., Repelewski, C. R., Kousky, V. E., O'Lenic, E. A., and Livezey, R. E. (1994). Long-lead seasonal forecasts - where do we stand? *Bull. Am. Met. Soc.*, 75:2097–2114.

Battisti, D. S. and Hirst, A. C. (1989). Interannual variability in a tropical atmosphere-ocean model: Influence of the basic state, ocean geometry and nonlinearity. *J. Atmos. Sci.*, 46:1687–1712.

Battisti, D. S. and Sarachik, E. S. (1995). Understanding and predicting ENSO. *Reviews of Geophysics*, Supplement:1367–1376. U.S. National Report to International Union of Geodesy and Geophysics 1991-1994.

Bell, G. D. and Halpert, M. S. (1995). *Interseasonal and Interannual Variability: 1986 to 1993*. NOAA Atlas No. 12. U.S. Department of Commerce.

Berliner, L. M. (1992). Statistics, probability, and chaos. *Statistical Science*, 7:69–122.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Bjerknes, J. (1969). Atmospheric teleconnections from the equatorial Pacific. *Mon. Weath. Rev.*, 97:163–172.

Bretherton, C. S., Smith, C., and Wallace, J. M. (1992). An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, 5:541–560.

Broomhead, D. and King, G. P. (1986). Extracting qualitiative dynamics from experimental data. *Physica*, 20D:217–236.

Buell, C. E. (1975). The topography of the empirical orthogonal functions. In *Fourth Conf. on Probability and Statistics in Atmospheric Sciences,* Tallahassee, FL, pages 188–193. Amer. Meteor. Soc.

Buell, C. E. (1979). On the physical interpretation of empirical orthogonal functions. In *Sixth Conf. on Probability and Statistics in Atmospheric Sciences,* Banff, AB, Canada, pages 112–117. Amer. Meteor. Soc.

Burgers, G. and Stephenson, D. B. (1999). The "normality" of El Niño. *Geophys. Res. Lett.*, 26:1027–1030.

Corti, S., Giannini, A., Tibaldi, S., and Molteni, F. (1997). Patterns of low-frequency variability in a three-level quasi-geostrophic model. *Climate Dynamics*, 13:883–904.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Contr. Signals Syst.*, 2:303–314.

Del Frate, F. and Schiavon, G. (1999). Nonlinear principal component analysis for the radiometric inversion of atmospheric profiles by using neural networks. *IEEE Trans. Geosci. Rem. Sensing*, 37:2335–2342.

DeMers, D. and Cottrell, G. (1993). Nonlinear dimensionality reduction. *Neural Inform. Processing Syst.*, 5:580–587.

Deser, C. (1999). A note on the annularity of the "Arctic Oscillation". *Geophys. Res. Lett.*, in review.

Dong, D. and McAvoy, T. J. (1996). Nonlinear principal component analysis - based on principal curves and neural networks. *Computers Chem. Engng.*, 20:65–78.

Farrell, B. F. and Ioannou, P. I. (1996). Generalised stability theory. Part I: autonomous operators. *J. Atmos. Sci.*, 53:2025–2040.

Feldstein, S. and Lee, S. (1996). Mechanisms of zonal index variability in an aquaplanet GCM. *J. Atmos. Sci.*, 53:3541–3555.

Finnoff, W., Hergert, F., and Zimmermann, H. G. (1993). Improving model selection by nonconvergent methods. *Neural Networks*, 6:771–783.

Flato, G. and et al. (1999). The Canadian Centre for Climate Modelling and Analysis global coupled model and its climate. *Climate Dynamics*, in review.

Fotheringhame, D. and Baddeley, R. (1997). Nonlinear principal components analysis of neuronal spike train data. *Biological Cybernetics*, 77:282–288.

Fyfe, J. C., Boer, G. J., and Flato, G. M. (1999). The Arctic and Antarctic Oscillations and their projected changes under global warming. *Geophys. Res. Lett.*, 26:1601–1604.

Ghil, M. and Childress, S. (1987). *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics.* Springer-Verlag.

Gong, D. and Wang, S. (1999). Definition of Antarctic oscillation index. *Geophys. Res. Lett.*, 26:459–462.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.*, 84:502–516.

Hastie, T. and Tibshirani, R. J. (1990). *Generalised Additive Models.* Chapman and Hall, London.

Hoerling, M. P., Kumar, A., and Zhong, M. (1997). El Niño, La Niña, and the nonlinearity of their teleconnections. *J. Climate*, 10:1769–1786.

Holzer, M. (1996). Asymmetric geopotential height fluctuations from symmetric winds. *J. Atmos. Sci.*, 53:1361–1379.

Hsieh, W. W. and Tang, B. (1998). Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Amer. Met. Soc.*, 79:1855–1870.

Hsu, H.-H. and Lin, S.-H. (1992). Global teleconnections in the 250-mb streamfunction field during the northern hemisphere winter. *Mon. Wea. Rev.*, 120:1169–1190.

Huang, H.-P. and Sardeshmukh, P. D. (1999). Another look at the annual and semiannual cycles of atmospheric angular momentum. *J. Climate*, in review.

Hurrell, J. W. (1995). Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, 269:676–679.

Jin, F.-F., Neelin, J. D., and Ghil, M. (1996). El Niño/Southern Oscillation and the annual cycle: subharmonic frequency-locking and aperiodicity. *Physica D*, 98:442–465.

Kirby, M. J. and Miranda, R. (1994). Nonlinear reduction of high-dimensional dynamical systems via neural networks. *Phys. Rev. Lett.*, 72:1822–1825.

Kirby, M. J. and Miranda, R. (1996). Circular nodes in neural networks. *Neural Comp.*, 8:390–402.

Kirby, M. J. and Miranda, R. (1999). Empirical dynamical system reduction I: Global nonlinear transformations. In *Semi-Analytic Methods for the Navier-Stokes Equations*, volume 20 of *CRM Proc. Lecture Notes*, pages 41–63. Amer. Math. Soc.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.*, 37:233–243.

Kushnir, Y. and Wallace, J. M. (1989). Low-frequency variability in the northern hemisphere winter: geographical distribution, structure and time-scale dependence. *J. Atmos. Sci.*, 46:3122–3142.

LeBlanc, M. and Tibshirani, R. (1994). Adaptive principal surfaces. *J. Amer. Statist. Assoc.*, 89:53–64.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather prediction. Technical report, Department of Meteorology, MIT. Science Report 1.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20:130–141.

Lücke, M. (1976). Statistical dynamics of the lorenz model. *J. Stat. Phys.*, 15:455–475.

Malthouse, E. C. (1998). Limitations of nonlinear PCA as performed with generic neural networks. *IEEE Trans. Neural Nets.*, 9:165–173.

McFarlane, N. A., Boer, G. J., Blanchet, J.-P., and Lazare, M. (1992). The Canadian Climate Centre second-generation general circulation model and its equilibrium climate. *J. Climate*, 5:1013–1044.

Miller, A. J., White, W. B., and Cayan, D. R. (1997). North Pacific thermocline variations on ENSO timescales. *J. Phys. Ocean.*, 27:2023–2039.

Mo, K. C. and Ghil, M. (1987). Statistics and dynamics of persistent anomalies. *J. Atmos. Sci.*, pages 877–901.

Monahan, A. H. (1999). Nonlinear principal compnent analysis by neural networks: Theory and application to the Lorenz system. *J. Climate*, in press.

Nakamura, H. and Wallace, J. M. (1991). Skewness of low-frequency fluctuations in the tropospheric circulation during the northern hemisphere winter. *J. Atmos. Sci.*, 48:1441–1448.

North, G. R. (1984). Empirical orthogonal functions and normal modes. *J. Atmos. Sci.*, 41:879–887.

Oja, E. (1997). The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17:25–45.

Oja, E. and Karhunen, J. (1993). Nonlinear PCA: Algorithms and applications. Technical Report A18, Helsinki University of Technology.

Ott, E. (1993). *Chaos in Dynamical Systems*. Cambridge University Press.

Pacanowski, R. C., Dixon, K., and Rosati, A. (1993). The GFDL modular ocean model users guide. Technical Report 2, Geophysical Fluid Dynamics Laboratory, Princeton, USA.

Palmer, T. N. (1999). A nonlinear dynamical perspective on climate prediction. *J. Climate*, 12:575–591.

Penland, C. (1996). A stochastic model of IndoPacific sea surface temperature anomalies. *Physica D*, 98:534–558.

Penland, C., Flügel, M., and Chang, P. (1999). On the identification of dynamical regimes in an intermediate coupled ocean-atmospheric model. *J. Climate*, page in press.

Penland, C. and Sardeshmukh, P. D. (1995). The optimal growth of sea surface temperature anomalies. *J. Climate*, 8:1999–2024.

Perlwitz, J. and Graf, H.-F. (1995). The statistical connection between tropospheric and stratospheric circulation of the northern hemisphere in winter. *J. Climate*, 8:2281–2295.

Philander, S. G. (1990). *El Niño, La Niña, and the Southern Oscillation*. Academic Press, San Diego.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge.

Priesendorfer, R. W. (1988). *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, Amsterdam.

Richman, M. B. (1986). Rotation of principal components. *Int. J. Climatology*, 6:293–335.

Rico-Martinez, R., Anderson, J., and Kevrekidis, I. (1996). Self-consistency in neural network-based NLPC analysis with applications to time-series processing. *Computers chem. Engng*, 20,Suppl.:S1089–S1094.

Sanger, T. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2:459–473.

Sengupta, S. K. and Boyle, J. S. (1995). Nonlinear principal component analysis of climate data. Technical Report 29, PCMDI.

Shindell, D. T., Miller, R. L., Schmidt, G. A., and Pandolfo, L. (1999). Simulation of recent northern winter climate trends by greenhouse-gas forcing. *Nature*, 399:452–455.

Smith, T. M., Reynolds, R. W., Livezey, R. E., and Stokes, D. C. (1996). Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, 9:1403–1420.

Stamkopoulos, T., Diamantaras, K., Maglaveras, N., and Strintzis, M. (1998). ECG analysis using nonlinear PCA neural networks for ischemia detection. *IEEE Trans. Sig. Proc.*, 46:3058–3067.

Suarez, M. J. and Schopf, P. S. (1988). A delayed action oscillator for ENSO. *J. Atmos. Sci.*, 45:3283–3287.

Takane, Y. (1998). Nonlinear multivariate analysis by neural network models. In *Studies in Classification, Data Analysis, and Knowledge Organisation: Data Science, Classification, and Related Methods*. Springer.

Tangang, F. T., Tang, B., Monahan, A. H., and Hsieh, W. W. (1998). Forecasting ENSO events: A neural network-extended EOF approach. *J. Climate*, 11:29–41.

Thompson, D. W. and Wallace, J. M. (1998). The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, 25:1297–1300.

Thompson, D. W. and Wallace, J. M. (1999). Annular modes in the extratropical circulation part I: Month-to-month variability. *J. Climate*, in review.

Thompson, D. W., Wallace, J. M., and Hegerl, G. C. (1999). Annular modes in the extratropical circulation part II: Trends. *J. Climate*, in review.

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., and Ropelewski, C. (1998). Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.*, 103:14291–14324.

Trenberth, K. E. and Paolino, D. A. (1980). The northern hemisphere sea level pressure data set: trends, errors, and discontinuities. *Mon. Wea. Rev.*, 108:855–872.

Ulbrich, U. and Christoph, M. (1999). A shift of the NAO and increasing storm track activity over Europe due to anthropogenic greenhouse gas forcing. *Climate Dynamics*, 15:551–559.

van Loon, H. and Rogers, J. C. (1978). The seesaw in winter temperatures between Greenland and northern Europe. part I: General description. *Mon. Wea. Rev.*, 106:296–310.

von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research.* Cambridge University Press, Cambridge.

Wallace, J. M. and Gutzler, D. S. (1981). Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weath. Rev.*, 109:784–812.

Wang, B. (1995). Interdecadal changes in El Niño onset in the last four decades. *J. Climate*, 8:267–285.

Whitaker, J. S. and Sardeshmukh, P. D. (1998). A linear theory of extratropical synoptic eddy statistics. *J. Atmos. Sci.*, 55:237–258.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences.* Academic Press, San Diego.

Woodruff, S., Slutz, R., Jenne, R., and Steurer, P. (1987). A comprehensive ocean-atmosphere data set. *Bull. Am. Met. Soc.*, 66:1239–1250.

Yuval (1999). Neural network training for prediction of climatological time series; regularized by minimization of the generalized cross validation function. *Mon. Weath. Rev.*, in press.