OCEAN DATA MINING WITH APPLICATION TO CLIMATE INDICES OF THE NORTH ATLANTIC

MADLENA HAKOBYAN

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-80852-8
Our file Notre référence
ISBN: 978-0-494-80852-8

Ocean data mining with application to climate indices of the North Atlantic

by

© Madlena Hakobyan

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of Science

Department of Computational Science

Memorial University of Newfoundland

March 2010

St. John's Newfoundland

To the memory of my dear father Haikaz Zohrabi Hakobyan

Abstract

This study is a part of the research project on development of a database and methods for data mining of ocean data. The first part of the project describes the implementation of the relational database management system (RDBMS) for ocean data. The second part of the project introduces a clustering method for identification of regions with homogeneous behavior of ocean parameters. Three algorithms K-means, Expectation Maximization(EM), and Farthest-First(FF) were implemented and evaluated in applications to the sea surface temperature data (SST). The clustering method was applied in analysis of two climate indices of the North Atlantic Ocean derived from the past observations of SST. The first one is associated with the North Atlantic Oscillation (NAO) and the second one with the variability of the Meridional Overturning Circulation (MOC). The two climate indices capture the most important long term variability of MOC and NAO.

Acknowledgements

This work is supported by the Canadian Foundation for Climate and Atmospheric Sciences and Memorial University of Newfoundland.

I would like to thank thesis supervisor Dr.Entcho Demirov for his guidance, suggestions, and everything that is done to me. The project would not have been possible without his help.

I would like to acknowledge and extend my gratitude to the Head of the Physics and Physical Oceanography Department, brilliant scientist and oceanographer, Professor Brad deYoung for his knowledge and understanding.

I would like to thank the Chair of the Computational Science Program, Professor Martin Plumer, and Ex-Chair, Professor George Miminis for their time, assistance and support.

I would like to thank Research Technician Mr.Jack Foley for his invaluable help with ocean instruments and field oceanography class.

I would like to thank my mother and my sisters for their love.

Many thanks to my friends from the Engineering, Physics and Physical Oceanography, and Geography Departments.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The classical methods for oceanic data analysis were based mainly on using manual logging of data and computation. Over the years, with technological advances and new electronic instruments, the methods of observations of oceanic data have greatly advanced and "new ocean instruments operate at data rates not possible with earlier mechanical devices and produce large amount of data that can only be handled by high-speed computers"[21]. The volume of data acquired is growing exponentially, and it is a challenging task to store, extract, manipulate and analyze the data[21]. This chapter presents an overview of present day methods for data storing and data analysis.

1.1 Database

The database research over the past 40 years has led to the database system analysis becoming a key research area in the field of the software engineering [11]. An electronic database helps to organize the information and provides the tools necessary to access specific bits of it quickly and efficiently. A *database* is a shared collection of logically related data. A database represents the entities, the attributes, and the logical relationships between the entities. An *entity* is a distinct object that is to be represented in the database. An *attribute* is a property that describes some aspect of the object that we wish to record, and a *relationship* is an association between entities. A *database management system*(DBMS) is the software that manages and controls access to the database. A *database application* is a program that interacts with the database at some point in its execution. A *database system* includes a collection of application programs that interact with the database. It is important to distinguish between the description of the database and the database itself. The description of the database is the *database schema.* The schema is specified during the database design process and is not expected to change frequently. However, the actual data in the database may change frequently; for example, it changes every time we insert a new entity in a table or add a new attribute. The data in the database at any particular point in time is called a *database instance.*

1.1.1 The Database Management System(DBMS)

The DBMS is software that manages and controls access to the database. It interacts with the users' application programs and the database[11]. The predecessor to the DBMS was the file-based system, which was the collection of application programs. Typically, a DBMS provides the following facilities:

(1) It allows users to define the database and to specify the data types, structures, and the constraints on the data that are to be stored in the database.

(2) It allows users to insert, update, delete, and retrieve data from the database, usually through a *Data Manipulation Language* (DML). DML provides a general inquiry facility to the data, called a *query language.* The most common query language is the *Structured Query Language* (SQL), which described in Section 1.1.5.

(3) It provides controlled access to the database, such as:

   (a) a security system to prevent unauthorized users accessing the database;

   (b) an integrity system, which maintains the consistency of stored data;

   (c) a recovery control system, which restores the database to a previous consistent state following a hardware or software failure.

The DBMS comes with tools that assist in organizing the data and simplifying information retrieval and modification. Built-in indexing makes it possible to locate information rapidly and efficiently, while automated processes ensure that data is always stored and cross-referenced in a consistent, error-free manner[55]. As a tool, DBMS was widely employed in managing databases. Among the major components of DBMS are: software, procedures, and data. The software component includes the DBMS software itself and the application programs, together with the operating system[13]. Typically, application programs are written in a third-generation programming language such as C, C++, Java, Visual Basic, FORTRAN, or using fourth generation language such as SQL, embedded in a third-generation language[13]. SQL statements give the power to manipulate all aspects of the database using code instead of visual tools, thus the use of fourth generation tools can improve the productivity significantly and produce programs that are easier to maintain[11]. The SQL statements are mostly coded instructions to perform operations such as extracting data, inserting new data, modifying existing data, and deleting data. *Procedures* refer to the instructions and rules that govern the design and use of the database. These may consist of instructions on how to log on to the DBMS, make backup copies of the database, change the structure of a table, or archive data to secondary storage.

The most important component of the DBMS environment is data. The data acts as a bridge between the machine components and the human components. The database may contain both the operational data and the meta-data. The effectiveness of database design and further data analysis depends on the quality of the data. As Hunter[30] pointed out, “data of poor quality is a pollutant to clear thinking”. There is no measurement procedure that is without the risk of error and the sources of errors can range from human errors to instrumentation failure.The main advantages of DBMS(Table 1.1) assure the quality of the data in a database.

Summary **Advantages of DBMS**
Control of data redundancy
Data consistency
Improved data integrity
Improved security

Table 1.1: Main advantages of DBMS

Database approach may control or completely eliminate the redundancy of the data by integrating the files in a way that multiple copies of the same data are not stored. This feature is especially important for the oceanographic data that comes from different data sources that overlap spatially and temporarily.

By eliminating or controlling redundancy, we are reducing the risk of inconsistencies that may occur during the data manipulation. If a data item is stored only once in a database, any update to its value will be performed only once and the new, updated value will be available immediately. If a data item is stored more than once and the system is aware of this, the system can ensure that all copies of the item are consistent[13].

Database integrity refers to the validity, consistency, and correctness of the stored data. Database integrity can be considered as another type of database protection as it concerned with the quality of data itself. Integrity is expressed in terms of *constraints*, which are rules that the database is not permitted to violate. Constraints apply to data items within a single record, as well as they apply to relationships between records.

Integration of data makes the data more vulnerable without appropriate security measures. Database security is the protection of the database from unauthorized users and the security takes the form of passwords to identify people authorized to use the database. The access that an authorized user is allowed on the data may also be restricted by the operation type(retrieve, update, or delete). Security concerns are becoming crucial when there is more than one user has an access to the database.

At the same time, the disadvantages of DBMS (Table 1.2) can expose the data in the database to possible loss,making the entire data vulnerable.

Summary **Disadvantages of DBMS**
Higher impact of failure
Complexity
Size

Table 1.2: Main disadvantages of DBMS

The centralization of resources increases the vulnerability of the system and the failure of any component can be disastrous[13]. The entire database may disappear if appropriate steps to secure the database will not be taken.

The functionality that DBMS provides makes it at the same time a complex piece of software [11].

The DBMS is also a large piece of software, and it occupies many megabytes of disk space and requires substantial amounts of memory to run efficiently[11].

1.1.2 Database Design and Data Model

The structure of the database is determined during database design. The database design has to produce a system that will satisfy current and future requirements of the end-users[13].

Data model is an integrated collection of concepts for describing and manipulating data, relationships between data, and constraints on the data[11]. It is a representation of 'real world' objects and events, and their associations. Major data models fall into three broad categories: object-based, record-based, and physical data models[13]. In a record-based model, the database consists of a number of records and they maybe of different types. Each record type defines a fixed number of fields, each typically of a fixed length. There are three principal types of record-based logical data model: *relational data model*, *network data model*, and *hierarchical data model*[55]. The relational data model is based on the concept of mathematical relations. In the *relational model*, data and relationships are represented as tables, each of which has a number of columns with a unique name. In the *network data model*, data is represented as collections of records, and relationships are represented by sets. Compared with the relational model, relationships are explicitly modeled by the sets, which become pointers in the implementation. The records are organized as generalized graph structures with records appearing as nodes and sets as edges in the graph. The *hierarchical data model* is a restricted type of network model. Data is represented as collections of records and relationships are represented by sets. A hierarchical model can be represented as a tree graph, with records appearing as nodes, also called segments, and sets as edges. The majority of modern database systems are based on the relational data model, whereas the early database systems were based on either the network or hierarchical data models.

1.1.3 Relational Database Management System(RDBMS)

The Relational Database Management System(RDBMS) represents the second generation of DBMSs and it is based on the relational data model. RDBMS is a "traditional" DBMS that was enhanced in the 1970s with the publishing of E.F.Codd's papers on Relational Databases [11, 9]. Since then, the RDBMS has become the dominant data-processing software in use today[11]. It is more robust and helpful than its DBMS predecessor. The theoretical foundation of the simple logical structure of the relational data model is a great strength of RDBMSs that was absent in the first generations of DBMSs with only network and hierarchical data models. RDBMS requires only that the database be perceived by a user as tables, where the mathematical concept of the a relation is physically represented as a table[13]. Many commercial vendors supply relational databases(Section 1.1.3), including Microsoft(SQL Server and Access), Oracle Corporation(Oracle), and IBM(DB2). There are also freely available open-source products such as MySQL and PostgresSQL. In our work, we are currently using MySQL platform that is described in Section 1.1.4.

1.1.4 MySQL

The RDBMSs are powerful, flexible, feature-rich software systems that are designed specifically for high-volume and mission critical applications[55]. They perform hundreds of transactions every second without batting an eyelid[55]. MySQL is a high-performance and multi-user RDBMS built around a client-server architecture[13]. MySQL designed specifically for speed and stability and currently it is one of the most popular RDBMSs. MySQL has two main fundamental features: performance and reliability. The time that takes on MySQL to execute a query and return the results is sometimes orders of magnitude faster than its competitors' [10, 55]. MySQL includes a unique new feature called a *query cache.* If a query made by a user returns a set of given records, repeating the same query should return the same set of records unless the underlying data has been changed. The query cache takes this principle further by storing the result in memory, thus bypassing the need to do the database search at all when a similar query is issued. Query cache enhances response times for queries that are called upon to retrieve the exact same data as a previous query[55]. As a relational database management system (RDBMS), MySQL is compatible with SQL. Because the data is stored in a similar structure on different RDBMS platforms, similar techniques are used to access and manipulate the data. Each of these platforms, including MySQL, uses Structural Query Language(SQL) as the universal language to implement these techniques.

1.1.5 Structural Query Language(SQL)

SQL is a standard database language that gained wide acceptance[11]. SQL is the language that is compatible with every relational database and is used to communicate and to administer the database. SQL allows user to:

- create the database and relation structures;
- perform basic data management tasks, such as the insertion, modification, and deletion of data from the relations;
- perform both simple and complex queries.

SQL is a portable language which allows use of the same command structure and syntax when we move from one RDBMS to another. SQL is a declarative language, not a procedural language, such as Java, Visual Basic, and so on. A user specifies what needs to be done in SQL and the Database Management System decides the best way to do it[13]. For instance, SQL statements that define what data needs to be retrieved, modified, updated or deleted do not specify *how* the database should do to that[13]. An SQL statement consists of **reserved words** and **user-defined words**. Reserved words have a fixed meaning, and they must be spelled exactly as required. User-defined words are made up by the user and represent the names of various database objects such as tables, columns, views, indexes, and so on. SQL is also an example of a **transform-oriented language**, or a language designed to use relations to transform inputs into required outputs. As a language, SQL has two major components such as:

- Data Defition Language(DDL) and Data Manipulation Language(DML) for defining the database structure and controlling access to the data;
- Data Manipulation Language(DML) for retrieving and updating data.

DDL and DML are discussed further in details in Sections 3.3.1 and 3.3.2.

1.2 Data Mining

Many methods of processing oceanographic data are recently developed for studying ocean circulation, currents, sea surface height, sea surface pressure, temperature, salinity, and and their variabilities. One group of the methods of processing and analyzing oceanographic data includes statistical analysis.

Data mining is one of the most vaguely defined fields. Fayyad et al.[22] define data mining as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. Friedman[35] points out that data mining sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. Data mining technology helps to explore large and complex data in order to discover useful, previously unknown patterns and relationships in a data. It is closely connected with developments in computer technology, particularly with the advancement and organization of database development. Unlike data mining though, the primary purpose of databases, as described in Sections 1.1 and 3.2, is *not* to analyze the data. Database query tools only retrieve information from a database using Structural Query Language (SQL) as described in Section 1.1.5 and present it to the user. This stage of data retrieval, when the data and information are extracted from databases is very similar to data mining, but the difference is that data mining looks for relations and associations between phenomena that are not known beforehand [23]. A query implemented on some random data set will have the lowest information capacity compare with data mining implemented on the same data set, and consequently, data mining will have the highest information capacity. Despite its great potential and effectiveness in analyzing various types of data sets, there is still no solid theoretical background behind data mining, which makes it hard to make definite statements about data mining techniques[23],[53]. It largely relies on trial and error when the decision needs to be made what technique is best for a particular data set and for solving a particular problem. Therefore, the representative steps that are highlighted in gray on Figure 1.2 may be repeated many times to get the desired data mining outcomes.

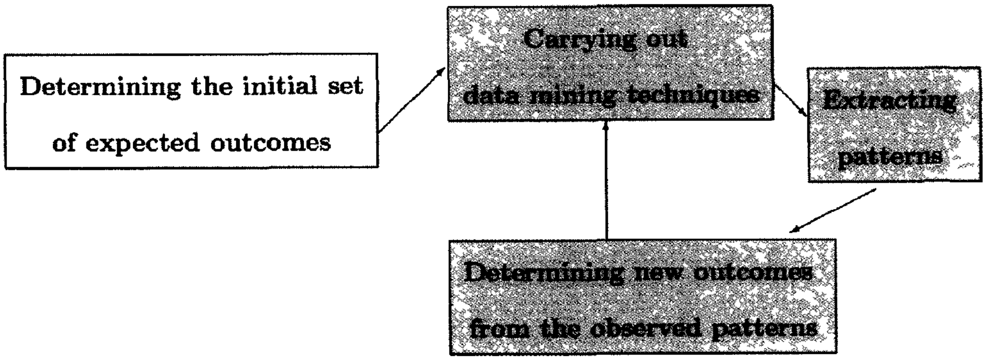

Determining the initial set of expected outcomes
Carrying out data mining techniques
Extracting patterns
Determining new outcomes from the observed patterns

Figure 1.1: Data Mining Outcomes

Current data mining packages provide statistical analysis procedures that include clustering methods, association rules, nearest neighbors, feature extraction, neural networks, genetic algorithms, and they do not include hypothesis testing, linear regression, logistic regression, canonical correlation, principal components, single factor analysis of variance(ANOVA), and factor analysis that are the "main-stay" in standard statistical packages[35].

Data mining, especially when it applies to great amount of data, can discover patterns and relations in the data that were not known beforehand, therefore, they cannot be used in developing a research hypothesis[23], and this is one of the main features that distinguishes data mining from statistics. Thus, data mining methods are concerned about selecting hypothesis from competing hypotheses rather than testing one. Among the other important features that distinguish data mining and statistics is the ability of data mining to analyze entire data set, as in statistical analysis it is necessary to have a sample from the data set that is being analyzed.

Berry and Linoff [3] marked two different analytical methodologies for data mining such as top-down and bottom-up. Top-down analysis starts with some idea, pattern, or hypothesis and uses statistical methods to test the hypothesis by confirming or rejecting it. There is no hypothesis to test in a bottom-up approach. This analysis examines the data and looks for the useful information and patterns to create hypothesis. The information that is obtained from a bottom-up approach shows patterns and tendencies in the data, but it cannot explain why the tendencies are useful and why and to what extend they are valid. And this is where the top-down analysis can be used to confirm the findings. Thus, these approaches are complementary.

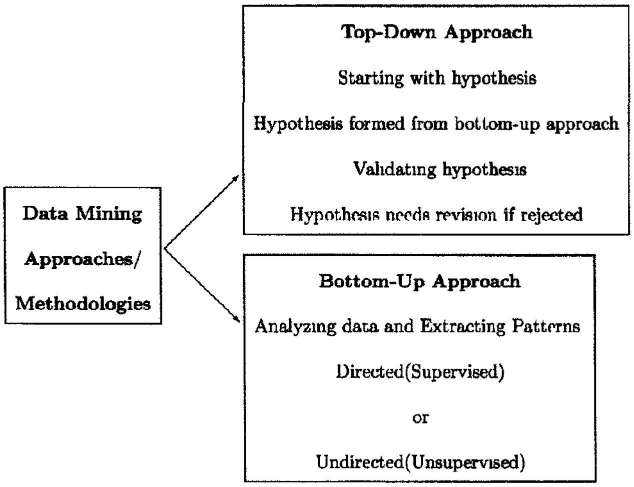


Figure 1.2: Data Mining Approaches

1.2.1 Processing Phases of Data Mining

The process of data mining must be automatic or semi-automatic [56], and the patterns discovered must be meaningful so they will enable to make nontrivial predictions on new data. The relationships and summaries or patterns that we derived from the data through the data mining process can be divided onto several phases, such as:

- Definition of the objectives for the analysis
  - Defining the objectives is one of the most important prerequisites to do entire analysis correctly; the objectives have to be stated as clear as possible without any room for ambiguity.
- Selection and organization of the data
  - Identifying the data source(s); checking possible presence of missing, or incorrect data; representation of the data in vector or matrix form.
- Data analysis
  - Choosing and implementing the best and most appropriate methods of analysis that fit to the data set
- Evaluation of the methods
  - Evaluation based on time constraints, data quality, data availability using the results from the previous phase.

1.2.2 Data Mining Techniques and Methods

A result of implementing a data mining algorithm on a data set can be expressed as a function $y(x)$, which takes x as an input and generates the output y. The exact form the function y takes is determined during the *training* or it is also called *learning* phase. The ability of a data mining to categorize correctly new examples of a data set that differ from those used for the training phase is called *generalization.* Usually, and especially with large data sets, the training data constitutes only very small fraction of all possible input vectors, therefore generalization is one of the central goals in data mining.

In many of the data mining applications, the input variables are also *preprocessed* to transform them into some new space of variables, where the particular data mining problem will be easier to solve[4]. This preprocessing stage is called *feature extraction.* It helps to find features that are faster to compute by preserving the useful discriminatory information[23]. Then, these features are used as the inputs in data mining process.

Data mining applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as *supervised classification*[4]. The goal of supervised classification is to assign each input vector to one of the finite number of discrete categories or *classes.* In other data mining problems, the set of input vectors x do not have corresponding target values, and the classification goal is aimed to find similar groups or *clusters* within the data, therefore the technique is called *clustering.*

Many data mining techniques such as *nearest neighbor* classification methods(Appendix VII), *cluster analysis*, (Section 2.1), and *multidimensional scaling methods* are based on similarity measures. There are two ways to obtain measures of similarity. One way is they can be obtained from the objects themselves, and second way is when it is necessary to give a precise definition of "similar". The formal definition of "similar" allows definition of "dissimilar" by applying a suitable monotonically decreasing transformation[26]. Therefore, if $s(i,\ j)$ denotes the similarity and $d(i,\ j)$ denotes the dissimilarity between objects i and j, then possible transformations include $d(i,j) = 1 - s(i,j)$. The definitions such as *distance* and *metric* are used to denote a measure of dissimilarity. Appendix III describes in more details the notion of distance measures, metric and Euclidean space.

1.2.3 The Form of the Data

There are different types of data that can be fed into a data mining application and the data may come in different forms. These forms are known as *schemas.* In all data mining applications it is important to be aware of the schema of the data, without which it is easy to miss patterns in the data[26]. The simplest form of data is a set of vector measurements on objects $o(1), ..., o(n)$. For each object we have measurements of p variables $X_1, ..., X_p$. Thus, the data can be viewed as a matrix with n rows and p columns, which is a data matrix, or it also can be referred as a *table.*

1.2.4 WEKA

The WEKA software that is used in this project is a collection of machine learning algorithms and data processing tools. WEKA was developed at the University of Waikato, New Zealand, and the WEKA name stands for *Waikato Environment for Knowledge Analysis.* It has tools to implement all standard data mining procedures such as classification, clustering, association rule mining, attribute selection. WEKA also allows to preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier without writing a code, as it includes:

- ⋆ point and click icons and menus
- ⋆ input dialog boxes
- ⋆ diagrams to describe analysis
- ⋆ a variety of data plots
- ⋆ graphical representations such as trees, networks, etc.

WEKA is written in JAVA and it has four interactive interfaces such as *Explorer*, *Knowledge Flow*, *SimpleCLI*, and *Experimenter.*

Explorer is a graphical user interface that gives access to all of WEKA facilities. It reads the database from ARFF file (or spreadsheet) and builds a decision tree from it. Choices and options are presented as menus and forms.

Knowledge Flow interface allows to drag boxes around the screen representing data sources and learning algorithms and join them together in a configuration that is necessary for the analysis. The data can be loaded incrementally.

Experimenter is helping to choose a method that will work best for a given problem. It also can distribute computing load among multiple machines. In this way large scale experiments can be implemented.

SimpleCLI allows access the WEKA features by entering textual commands through command line interface.

Data Mining is a useful mehtod which can be used in ocean, atmospheric and climate studies. The information which recently become available in these areas includes large data sets with complex temporal and spatial structure. Analysis of these data and pattern recognition requires advanced computer methods like data mining.

1.3 Data Mining Applications in Ocean and Climate Studies

Huang et al.[29] used cluster analysis to define the relationships between ocean salinity, temperature structures and climate variability. Huang et al.[29] proposed a quantitative inter-transaction association rules mining algorithm and introduced a technique for analyzing ARGO ocean data to extract information about salinity/temperature patterns. They also increased the data mining efficiency by adopting FITI and PrefixSpan algorithms. Their study is applied to ocean salinity measurements obtained from the waters surrounding Taiwan.

Steinbach et al.[38] developed an approach in clustering methodology for the discovery of Ocean Climate Indices (OCIs) and divided the cluster centroids into several categories: those that correspond to known OCIs, those that are variants of known OCIs, and those that represent potentially new OCIs. They developed alternative methodology for the discovery of OCIs using clusters that represent ocean regions with relatively homogeneous behavior. The centroids of those clusters are time series that summarize the behavior of these ocean areas. Their goal was to use climate variables, such as long term sea level pressure(SLP) and sea surface temperature(SST) to discover interesting patterns that relate changes in Net Primary Production(NPP) to land surface climatology and global climate, where NPP is the key variable for understanding the global carbon cycle and the ecological dynamics of the Earth.

Wooley et al.[6] described their preliminary findings in unsupervised classification(clustering) of a database with very large acoustic images (about 30 megabytes each) of the ocean floor. The authors used an approach of parallelizing unsupervised learning algorithm to partition data with the number of instances to be classified that is more than 10,000 and distribute it to multiple processors. The learning algorithm used is AutoClass, an unsupervised Bayesian classification system. Classification results when 2 or 3 processors are used are similar to the classification results achieved when one processor is used[6].

Cheng and Wallace[8] have used clustering techniques to analyze the long-term climate variability in the upper atmosphere on the Earth's Northern hemisphere. This variability is dominated by three recurring spatial pressure patterns (clusters) identified from data recorded daily by National Meteorological Center from 1948 to 1985. The authors used hierarchical cluster analysis based on the Ward method, which is performed on the Northern Hemisphere wintertime 10-day low-pass-filtered 500-hPa height field. It was one of the first studies of this type to define the clusters in terms of total(low-pass-filtered) height fields rather than height anomaly fields.

1.4 Thesis Objectives

The major goal of this study is to use climate variables, such as sea level pressure (SLP) and sea surface temperature (SST) to discover patterns relating to ocean climate indices (OCIs) in the North Atlantic.

This work presents results from study of interannual variability in the North Atlantic. Ocean climate indices have been developed that summarize the behavior of selected areas of the world's oceans[50] and they can be linked to major patterns of climate variability. The ocean climate indices are often defined through eigenvalue analysis techniques such as Principal Components Analysis (PCA) and Singular Value Decomposition (SVD). The main advantages of using PCA and SVD for analysis of the ocean data include ability to store the most important information without redundancy and noise. They also allow compression, which increases storage effectiveness through reduction of the dimensionality of the space. However, these techniques impose a condition that all discovered signals must be orthogonal to each other[50].

In this project a different approach such as data mining is used. Data mining techniques are firstly applied to the Ocean Climate Indices. In addition, we study connection of variabilities among different parts of the ocean, as data mining and clustering provide more information about connection between different parts of the ocean. This section describes the methods used in the analysis of interannual variability of the Labrador Sea and North Atlantic SST. The approach is based on the use of data mining techniques such as clustering and Empirical Orthogonal Functions (EOF) analysis. Data mining clustering techniques are used on data sets to reveal well-separated groups of data at various levels of detail. We run three clustering algorithms - K-Means, Expectation Maximization, and Farthest-First - on the same data and test how well these algorithms perform in terms of the outcome and the quality of clustering. This work shows preliminary results, details specific findings, and outlines the data mining processes and techniques of extracting classes for ocean applications. This project started with creation of the database, which currently contains data from different observational systems with wide range of temporal and spatial attributes. The database approach is shown do be an effective tool for archiving, accessing and retrieving large volume of oceanographic data in data mining applications.

Chapter 2

Methods Of Data Analysis

2.1 Clustering

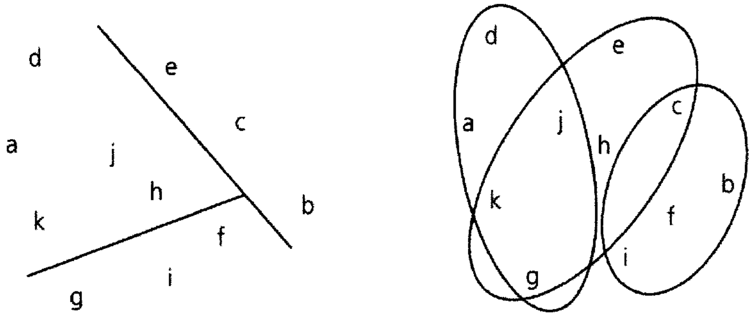
Cluster analysis is one of the descriptive data mining methods. Clustering techniques apply when there is no class to be predicted as in supervised classification (Section 1.2.2) and the instances are to be divided into natural groups. This section focuses on clustering approaches that were chosen for this study - K-Means, Expectation Maximization(EM) and Farthest-First.

The K-Means is an iterative distance-based clustering algorithm[25, 23]. Farthest-First is a sequential search algorithm to identify iteratively the cluster centers and it is a combination of hierarchical clustering and distance-based clustering[48, 58] that are described below. EM is an algorithm that assigns each object to a cluster according to a weight representing the probability of membership[25].

Clustering involves decomposition of a data set into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups. The result of clustering can be expressed in different ways and the clustering output may take a form of a diagram that shows how the instances in a data set fall into clusters. Examples of the diagrams are shown on Figure 2.1 and Figure 2.2.

In some clustering algorithms, such as *partition-based clustering*, the task is to partition a data set into k disjoint sets of points such that the points within each set as homogeneous as possible.

For instance, given the set of n data points $D = \{x(1), ..., x(n)\}$, the task is to find K clusters $C = \{C_1, ..., C_K\}$ such that each data point $x(i)$ is assigned to unique cluster C_k. Each cluster is exclusive and any single instance of a data set belongs only to one cluster. A diagram on Figure 2.1(a) shows an example of partion-based clustering.Some other clustering algorithms allow one instance be inside of more than one cluster, and, therefore, the corresponding diagram will show the *overlapping* clusters(Venn diagram)(Figure 2.1b).



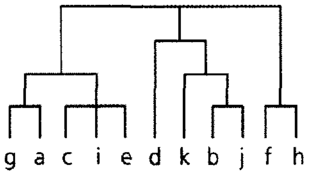
(a) (b)

Figure 2.1: (a)Associating a cluster number with with each instance; (b)Overlapping subsets representing each cluster - Venn diagram

Algorithms such as *probabilistic model-based clustering* associate instances with clusters probabilistically (Figure 2.2a), where every instance belongs to a particular cluster with certain probability. Others, that are called *hierarchical clustering algorithms*, produce a hierarchical structure of clusters, so that at the top level we have only few clusters, and each of those clusters divides itself into its own sub-clusters at one level down, and further down to individual instances. Figure 2.2b shows one possible hierarchical clustering of a eleven-instance data set. Hierarchical methods permit a convenient graphical display, in which the entire sequence of merging or splitting the clusters is shown, and such displays are also called *dendrograms*. There are two kinds of hierarchical techniques: *agglomerative* and *divisive*. They construct their hierarchy in the opposite direction, and produce different results. Agglomerative methods start when all objects are apart, that is at the first step we have n clusters, and in each step two clusters are merged until only one cluster left. On the other hand, divisive methods start when all objects are together and in each following step a cluster is split up until there are n clusters.

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

(a)



(b)

Figure 2.2: (a)Associating instances with clusters probabilistically; (b)Hierarchical structure of clusters

The choice among the algorithms should be dictated by the nature of the mechanisms that are thought to underlie the particular clustering phenomenon[56]. However, very often, for pragmatic reasons, the choice is usually dictated by the clustering tools that are available, which in our study is the ability of the WEKA data mining software to perform the clustering.

2.1.1 K-Means

K-Means is an example of partition-based clustering. K-Means starts by randomly picking K cluster centers, where the number of clusters K is known beforehand. Then, it assigns each of these points to a cluster whose mean is closest(in a Euclidean distance sense described in Appendix III), and it computes a new mean for each cluster. These means are taken to be new center values for their respective clusters. This process is repeated with the new cluster centers, and iteratively continues until the same points are assigned to each cluster and the cluster centers stop changing. The final clusters are quite sensitive to the initial cluster centers and completely different arrangements can arise from small changes in the initial random choice[25]. A brief description of the K-Means clustering algorithm is included in the Appendix IV. Nothing is known *a priori* about how to choose the number of clusters K, and the first step in cluster analysis is to define this number[56, 26, 4].

2.1.2 Expectation Maximization

The Expectation-Maximizaion (EM) is similar to the K-Means clustering method described above and it can be viewed as an extension of the K-Means algorithm. Unlike K-Means though, instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight which is a probability of membership[25]. The probability of membership is calculated. Thus, there is no strict boundaries between clusters and the new means that are computed based on weighted measures. It is an iterative algorithm for "maximizing a likelihood score function given a probabilistic model with missing or hidden data" [26] and it starts with initial guess of the parameter values of the mixture model(referred as the parameter vector). It iteratively rescores the objects against the mixture density produced by the parameter vector. Then, those rescored objects are used to update the parameter estimates[58]. Each object is assigned a probability that it would possess a certain set of attribute values given that it was a member of a given cluster. A brief description of the EM algorithm is included in the Appendix V.

2.1.3 Farthest-First

The *Farthest-First traversal* was first introduced by Hochbaum and Shmoys[28]. The Farthest-First starts with picking any data point, and assigning it its own cluster. Then it chooses the point farthest from it, then the point farthest from the first two (the distance of a point x from a set S is the usual $min\{d(x,y) : y \in S\}$) and process continues until k points, which is a some specified threshold, are obtained. These points are taken as cluster centers and each remaining point is assigned to the closest center. If the distance function is a metric, the resulting clustering is within a factor two of optimal[14], that is for any k if T is the solution returned by Farthest-First traversal, and T^* is the optimal solution, then

$$cost(T) \leq 2cost(T^*)$$

A brief description of the Farthest-First algorithm is given in Appendix VI.

2.2 Principal Component(EOF) Analysis

Principal Component Analysis(PCA) or Empirical Orthogonal Function Analysis(EOFA) is a multivariate statistical technique with many applications to oceanographic, atmospheric and other geophysical fields.

A typical example of the PCA application in oceanography can be considered the ocean climate indices. The latter are simple characteristics of interannual variability of ocean and atmosphere. This application has been successful in describing important aspects of the North Atlantic variability. Cluster analysis and PCA are suitable for this type of applications and at the same time they may result in different types of classifications.

The purpose of PCA is to reduce a data set that contains a large number of variables to a data set that contains fewer new variables. Despite the fact that the new variables are fewer than in the original data set, they describe the large fraction of the variability of the original data set. Therefore, if we have multiple observations of (K) data vector $\mathbf{x}$, we want to find (M^*) vectors $\mathbf{u}$ whose elements are functions of the elements of the $\mathbf{x}$ values, that contain most of the information in the original collection of $\mathbf{x}$ values, and whose dimensionality $M^* \ll K$.

PCA detects linear dependencies between variables and replaces groups of correlated variables by new uncorrelated variables, the principal components. The elements of this new vector $\mathbf{u}$ are called *principal components.* The choice of how many components to extract is arbitrary and it depends on the explained variance in each particular case. Information will be lost, if choosing a small number of components that fail to explain the variability in the data very well [26]. Therefore, in choosing an appropriate number k of principal components, one approach is to increase k until the squared error quantity above is smaller than some acceptable degree of squared error.

For high-dimensional data sets, in which the variables are often relatively well-correlated, it is not uncommon for a relatively small number of principal components(from 5 to 10) to capture 90% or more of the variance in the data.

Some basics of PCA are described in the Appendix IX.

Chapter 3

Data and Database

3.1 Ocean Data Collection

The oceanographic data that were collected for this project reflect different ocean features and characteristics and are related to different physical processes in the ocean, such as ocean currents, circulation, ocean-atmosphere interaction, water masses. Currently, observational methods include satellites, ocean buoys with equipped sensors, tide gauges, in-situ measurements, and model data. The major data sources are:

- ⋆ Coriolis Data Center
- ⋆ Marine Environmental Data Service, Integrated Science Data Management(GDSI)
- ⋆ Met Office Hadley Centre
- ⋆ Marine Environmental Data Service/Integrated Science Data Management(GDSI), Fisheries and Oceans Canada
- ⋆ Fisheries and Oceans Canada/Ocean and Ecosystem Science/ Bedford Institute of Oceanography
- ⋆ NOAA Satellite and Information Services/National Oceanographic Data Center

- ⋆ National Snow and Ice Data Center

The primarily physical ocean parameters acquired are summarized in the Table 3.1.

Data Source	Data Set Acquired	Disk Usage
Met Office Hadley Centre	Sea Ice and SST	694688
MEDS/ODSI Fisheries and Oceans Canada	Arctic Buoy Data,1979-1999	1065368
	ARGO- Canadian Tracked Data	24576
	Drifting Buoys	70240
Coriolis Data Center	Coriolis Global Profiles Distribution	28400560
	Lagrangian buoys	20480
	GOSUD GDAC global distribution	104948
	ARGO GDAC global distribution	44876356
	Mercator Weekly Distribution	69632
Bedford Institute of Oceanography	Hydrographic(Climate) data	6892396
	Satellite derived (NOAA/AVHRR) SST for NW Atlantic	7794597
NOAA Satellite and Information Services National Oceanographic Data Center	World Ocean Database	9664920
National Snow and Ice Data Center	Sea Ice Concentration	11840
	Ice Extent	842676
	Ice Velocity	437856
	Water Temperature	275453
	Ice Thickness	785495

Table 3.1: Major Data Sources and Data Sets

Each data set also includes data on one or more oceanographic parameters, particularly:

- ⋆ **Arctic Buoy Data(1979-1999)** include temperature, pressure, position, ice-velocity, and ice buoy/CTD. The data sets are divided by year from 1979 to 1999. It is collected from all drifting buoys reporting from north of 66 °.
- ⋆ **Argo- Canadian Tracked Data** contain argo-profile and argo-drift data sets with parameters such as temperature, pressure and salinity.
- ⋆ **Drifting Buoys** data set include buoy position, date, time and variables such as surface and subsurface water temperature, air pressure, air temperature, wind speed and wind direction.
- ⋆ **Coriolis Global Profiles Distribution** include types of profiles such as XBT, CTD, CTD from US Ocean Climate Library and Argo float.
- ⋆ **Lagrangian Buoys** from Coriolis Data Center is a trajectory data that is collected in real-time

by floating buoys.

- ⋆ **GOSUD GDAC Global Distribution** data include sea surface salinity collected by research and opportunity ships.

- ⋆ **Hydrographic (Climate)** data from the Bedford Institute of Oceanography acquired in the categories such as climate complete profiles, climate time series, and climate seasonal cycle. The data comes from a variety of sources including hydrographic bottles, CTD casts, profiling floats, spatially and temporally averaged Batfish tows, and expendable, digital or mechanical bathythermographs. Near real-time observations of temperature and salinity from the Global Telecommunications System (GTS) are also included.

- ⋆ **Satellite derived(NOAA/AVHRR) SST** data for the Northwest Atlantic from 1982 to 2005 is another data set acquired from the Bedford Institute of Oceanography.

- ⋆ **World Ocean Database** is coming from National Oceanographic Data Center and it includes drifting buoy data, expendable demographically data, high resolution CTD data, mechanical bathythermograph data, moored buoy data, profile float data, and plankton data.

Most of these data sets are in NetCDF or ASCII format. Selection, organization and retrieval of the data for a particular time period and geographical location may be a time-consuming and cumbersome process, since some data may overlap, while others may be missing, and some may not have appropriate quality control flag. This is where the database can be applied as an effective tool for storage and retrieval of the data.

3.2 Database

This section describes the oceanographic database(hereafter “the database”) that was experimentally created using three major data sources from the collection of data described in the previous section. Those data sources include:

- ⋆ Coriolis Data Center

★ Marine Environmental Data Service, Integrated Science Data Management(GDSI)

★ Met Office Hadley Centre

3.2.1 Relational Database Management System

The database is composed of multiple tables. Figure 3.1 shows a typical data table instance from the MEDS data set. The table divides data into rows, with a new entry (or record) on every row. The data in each row is further broken down into cells (or fields), each of which contains a value for a particular attribute of the data. For instance, the record for ID 477296 is divided into separate fields for 'ID', 'pressure', 'temp', 'salinity', 'sig', 'depth', 'source', 'profileID'. The rows within a table can be arranged by ID, or by any other user specified criteria. It is necessary to have some method of identifying a specific record in the table, and in our example, it is identified by 'ID', which is a number that is unique to each row or record. This unique field is called the *primary key* for that table. A primary key is an unchanging and it is a unique identifier for each record.

ID	pressure	temp	salinity	sig	depth	source	profileID
477296	5438	2.3	34.88	27.85	5437	meds	61117
477295	5435	2.39	34.86	27.83	5434	meds	61115
477291	5251	2.3	34.89	27.86	5250	meds	61116

Figure 3.1: Typical data table instance from MEDS data set

The relational database is composed of multiple tables that contain interrelated pieces of information. By adding more tables into database, we use the fundamental concept of RDBMS that is creating relationships between the tables that make up the database. In applying this concept, we relate records in different tables to one another through the use of *foreign keys.* Foreign key is a column or combination of columns used to establish or enforce a link between data in two tables. It serves as a point of commonality to link records in different tables together[13]. For example, Figure 3.2 shows the instances of two tables(relations) where the referencing table that has a foreign key column references the primary key column of the referenced table.

spaceID	long	lat	time
61227	179.47	-61.19	20814.5
61217	179.48	-61.1	20804.6
61125	179.53	-58.08	20782.4

Primary Key

Foreign Key

ID	pressure	temp	salinity	profileID
20045	649.1	2.53	34.49	61227
19039	479.3	2.6	34.36	61217
18046	294.9	3.41	34.01	61125

Figure 3.2: Instances of two tables(relations) with Primary and Foreign Keys. It is also an example of one-to-one relationship between tables.

A relationship could be *one-to-one* or *one-to-many*. In an one-to-one relationship, a record in one table is linked to one and only one record in another table. Figure 3.2 is also an example of one-to-one relationship. In one-to-many relationships a record in one table is linked to multiple records in another table. Figure 3.3 illustrates an instance of one-to-many relationship from Argo-data, January, 2007.

As soon as a foreign key is set up, the relational database only will allow entry of those values into the Data Table, 'profileID' field, which also exist in the ID Table, 'spaceID' field. This way foreign key constraints can significantly help in enforcing the data integrity of the tables in a database and reducing the occurrences of "bad" or inconsistent field values[53].

3.2.2 Building Relational Data Model

The data and the relationships among the data are represented as tables in the database, and each individual table in the database has a number of columns with a unique name. A typical relation

spaceID	long	lat	time
51208	179.897	-5.019	20430.3

Primary Key

Foreign Key

ID	pressure	temp	salinity	profileID
4672	4.4	29.691	35.565	51208
4673	10.5	29.692	35.565	51208
4674	16.8	29.692	35.563	51208
4675	24	29.679	35.562	51208
4676	31.3	29.66	35.561	51208

Figure 3.3: One-to-many relationship between tables

in the database is represented as a two-dimensional table with the rows corresponding to individual records and the columns corresponding to attributes, and it holds information about the objects that are represented in the database. The attributes can appear in any order and the relation will still be the same, therefore they will convey the same meaning. For example, location and time information is represented by 'ID' table with columns for attributes 'spaceID', 'longitude', 'latitude', and 'time'. Similarly, the information on pressure, pressure-qc, temperature, temperature-qc, salinity and salinity-qc is represented by 'DATA' table, with columns for attributes 'profileID', 'pressure', 'presQC', 'temp', 'tempQC', 'salinity', 'salQC'. Figure 4.3 shows instances of 'ID' and the 'DATA' tables. It is important to mention that there is a relationship between 'ID' and 'DATA' tables: each point in 'ID' table has data associated with it which is stored in 'DATA' table. There is no explicit link between these two tables and it is only by knowing that the attribute 'spaceID' in the 'ID' table is the same as the 'profileID' in 'Data' table, we can establish that the relationship exists.

As in the relational data model, a user sees the database as a number of tables which applies only to the logical structure of the database. It does not apply to the physical structure of the database, which can be implemented using a variety of storage structures[11].

spaceID	longitude	latitude	time
70126	179.9	-61.19	20844.6
70116	179.83	-61.19	20834.6
61204	179.89	-58.12	20792.3

ID	pressure	pressQC	temp	tempQC	salinity	salQC	profileID
14897	269.4	1	2.36	1	34.14	1	70126
19789	309.1	1	2.61	1	34.21	1	70116
12876	259.4	1	3.66	1	34.07	1	61204

Figure 3.4: A sample instance of a relational schema

3.2.3 Ocean database main features

The ARGO data that is organized and stored in the database, can be extracted and displayed in any manner and order. Through the database approach we eliminated the data redundancy by integrating the files in a way that multiple copies of the same data are not stored. However, sometimes it is necessary to duplicate key data items to model relationships. Therefore, through the database we can control the amount of redundancy inherent in the database. This feature is especially important for the oceanographic data that comes from different data sources that may overlap spatially and temporarily. By controlling data redundancy, we are greatly reducing the risk of inconsistencies that may occur during data manipulation, and if data value is stored once in the database, any update to it will be performed only once and the updated value will be available immediately.

The database approach also allowed us to put constraints on the data to filter out all the data that is not necessary or the data that carries an error. Constraints apply to data items within a single record, as well as they apply to relationships between records. For example, integrity constraint states that a value in the 'longitude' field cannot be greater than 180 and cannot be less than -180;

or for 'latitude' field the value cannot be less than -90 and cannot be greater than 90. Therefore, RDBMS allows us during the database design to define and enforce integrity constraints.

During the design stage of the database we also reinforced the quality and validity of the data, as if a measurement procedure has poor validity, any conclusions we draw from data will be misleading. As such, we assured the quality of the data by placing additional fields such as 'tempQC', 'pressQC', 'salQC' in the database tables. The value of those fields is restricted to 1(one), which denotes the quality control of each individual record. Therefore, no data were imported into the database, if the values of those fields are different from 1.

3.3 Structural Query Language(SQL)

As we stated in Section 1.1.5, SQL is the language that every relational database understands and is used to communicate and to administer the database. SQL statements that we are using currently in our work can be divided into two broad categories, each concerned with a different aspect of database management:

- ⋆ **Statements used to define the structure of a database** These statements define the relationships among different pieces of data, definitions for database, table and column types, and database indices. In the SQL specification, this component is referred to as Data Definition Language(DDL).

- ⋆ **Statements used to manipulate data** These statements control adding and removing records, querying and joining tables, and verifying data integrity. In the SQL specification, this component is referred as Data Manipulation Language(DML).

3.3.1 Data Definition Language

Data Definition Language(DDL) is used to create, modify or remove tables and other database objects. It includes statements such as 'CREATE TABLE', 'ALTER TABLE', 'DROP TABLE', 'CREATE INDEX', and so on.

- ⋆ Example of 'CREATE TABLE' statement:
  - mysql> create table medsoceanecosJM0708 select longitude, latitude, time, pressure, profileID from medsoceaDATA2, medsoceanID where medsoceanID.spaceID = medsoceanDATA2.profileID and medsoceanID.latitude >= 40 and medsoceanID.latitude <= 80 and medsoceanecID.longitude >= -70 and medsoceanID.longitude <= 0 and medsoceanecID.time >= 21001 and medsoceanID.time <= 21245
- ⋆ Example of 'ALTER TABLE' statement:
  - mysql> alter table medsoceanDATA5 rename to argoMarAug9494;
  - mysql> alter table argoMarAug9494 drop profileID;
- ⋆ Example of 'DROP TABLE' statement:
  - mysql> drop table argogdacMA0406;

3.3.2 Data Manipulation Language

Data Manipulation Language(DML) provides a set of operations that support the basic data manipulation operations on the data that is stored in the database. Data manipulation operations usually include the following:

- ⋆ insertion of new data into the database
  - The SQL statements 'INSERT' are used to insert new rows of data into database.
- ⋆ modification of data stored in the database
  - The SQL statemts 'UPDATE' are used to modify existing rows of data
- ⋆ retrieval of data contained in the database
  - The SQL statements 'SELECT' are used to read and extract data from the database. This portion of the language has its own name and it is a Data Query Language(DQL). The SQL 'SELECT' staments can also be referred as *SQL queries*

⋆ deletion of data from the database

- The SQL statements 'DELETE' are used to remove rows of data from the database

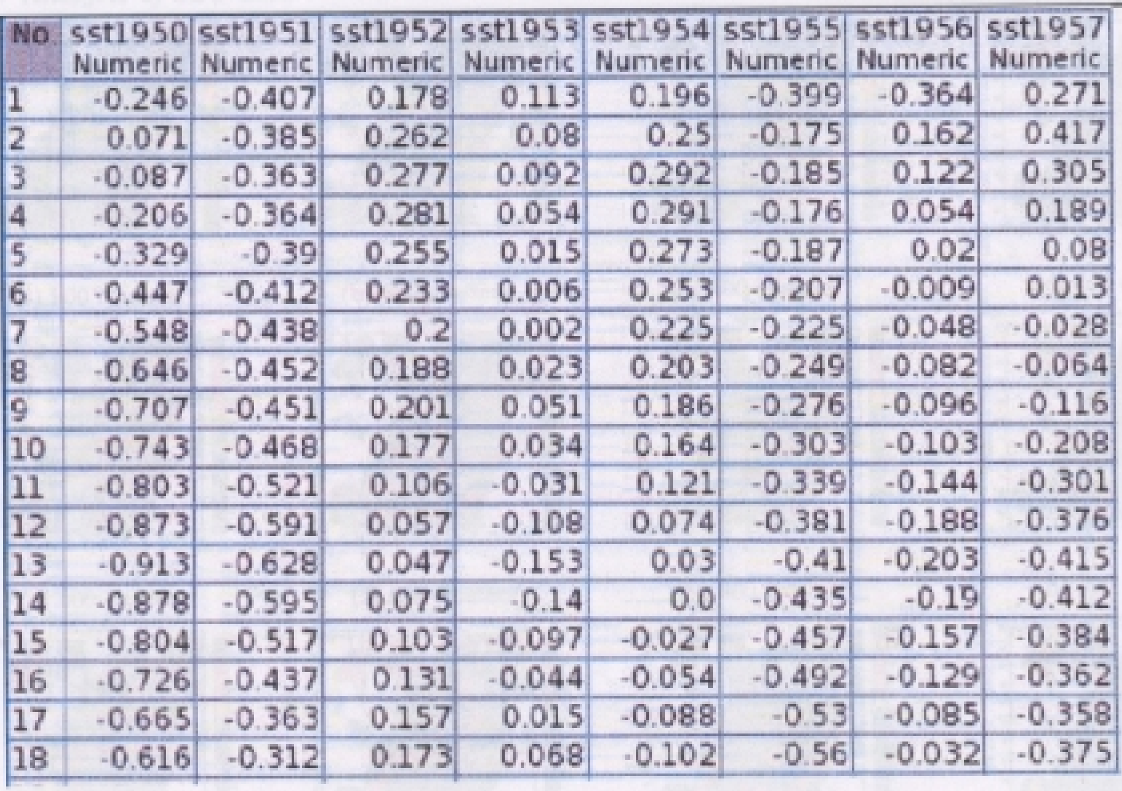
Chapter 4

Cluster Analysis of SST data

4.1 EM, Farthest-First, and K-Means algorithms

The data used in this study is the SST analyzed observation for the North Atlantic of the Met Office Hadley Center. The gridded SST includes some uncertainty which is related to the sampling spatial and temporal resolution. The observing system in the Southern part of the region was traditionally better than in the Northern sub-polar part of the ocean. Hence, the bias of the data due to the observational errors is higher in the Northern than in the Southern part of the study area[17], [57], [41].

The clustering algorithms that are used in this study are K-Means, Expectation Maximization, and Farthest-First. All three algorithms are provided within the WEKA framework(Section 1.2.4). For all three algorithms we used the same 1950 -1997 SST data set, which was too large in terms of WEKA memory usage. As such, the original SST data set was pre-processed and for each year the annual mean anomaly was calculated. As an example, Table 4.1 shows the extract of the SST data matrix that is uploaded inside of WEKA. The SST data matrix consists of 48 columns and 1468 rows, where columns represent the the annual mean value for each year of observation and the rows represent the geographical locations of the points in the ocean.



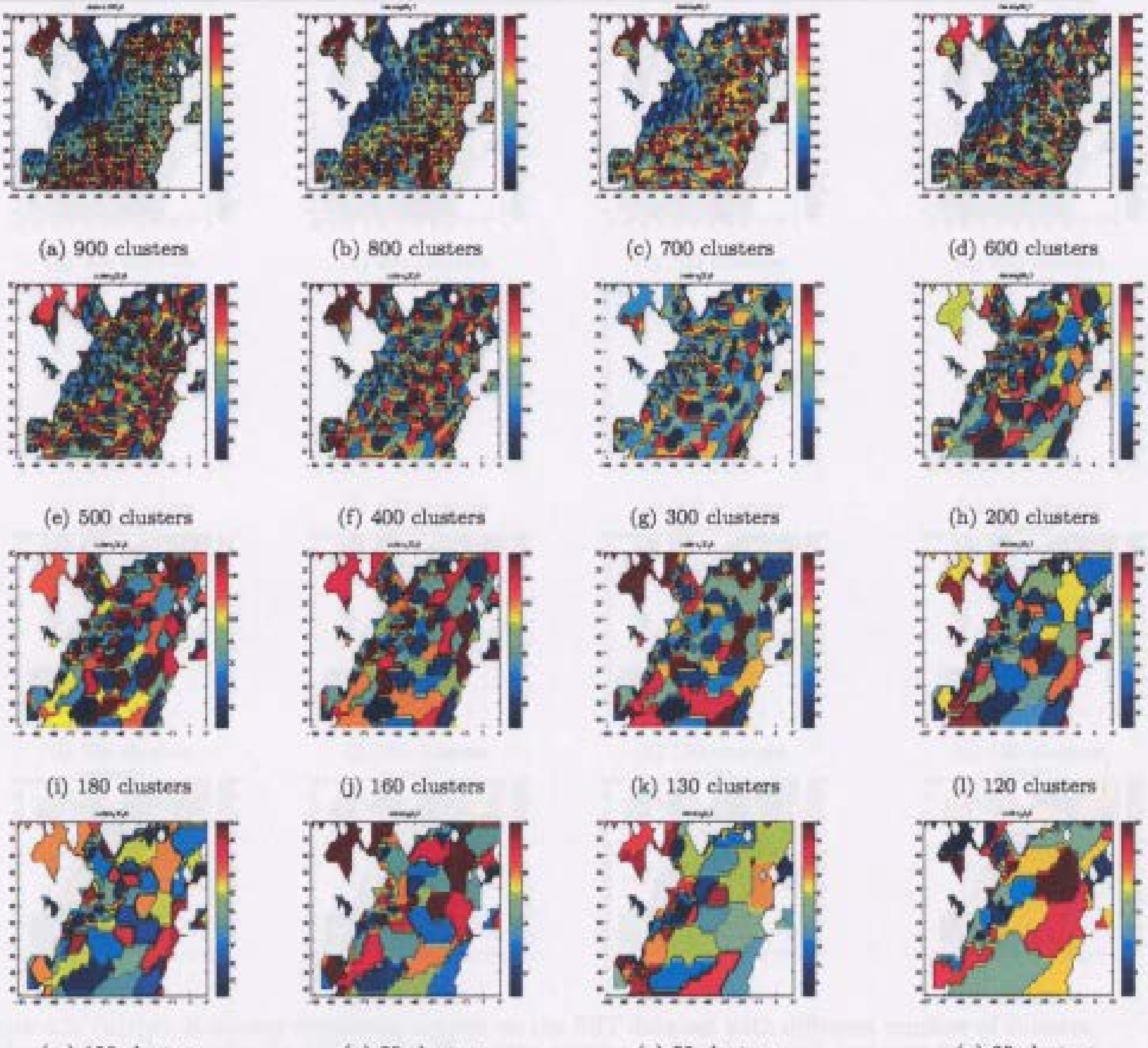
No	sst1950 Numeric	sst1951 Numeric	sst1952 Numeric	sst1953 Numeric	sst1954 Numeric	sst1955 Numeric	sst1956 Numeric	sst1957 Numeric
1	-0.246	-0.407	0.178	0.113	0.196	-0.399	-0.364	0.271
2	0.071	-0.385	0.262	0.08	0.25	-0.175	0.162	0.417
3	-0.087	-0.363	0.277	0.092	0.292	-0.185	0.122	0.305
4	-0.206	-0.364	0.281	0.054	0.291	-0.176	0.054	0.189
5	-0.329	-0.39	0.255	0.015	0.273	-0.187	0.02	0.08
6	-0.447	-0.412	0.233	0.006	0.253	-0.207	-0.009	0.013
7	-0.548	-0.438	0.2	0.002	0.225	-0.225	-0.048	-0.028
8	-0.646	-0.452	0.188	0.023	0.203	-0.249	-0.082	-0.064
9	-0.707	-0.451	0.201	0.051	0.186	-0.276	-0.096	-0.116
10	-0.743	-0.468	0.177	0.034	0.164	-0.303	-0.103	-0.208
11	-0.803	-0.521	0.106	-0.031	0.121	-0.339	-0.144	-0.301
12	-0.873	-0.591	0.057	-0.108	0.074	-0.381	-0.188	-0.376
13	-0.913	-0.628	0.047	-0.153	0.03	-0.41	-0.203	-0.415
14	-0.878	-0.595	0.075	-0.14	0.0	-0.435	-0.19	-0.412
15	-0.804	-0.517	0.103	-0.097	-0.027	-0.457	-0.157	-0.384
16	-0.726	-0.437	0.131	-0.044	-0.054	-0.492	-0.129	-0.362
17	-0.665	-0.363	0.157	0.015	-0.088	-0.53	-0.085	-0.358
18	-0.616	-0.312	0.173	0.068	-0.102	-0.56	-0.032	-0.375

Table 4.1: Extract from the SST data matrix

Having some points in a metric space, and specific number of clusters K, the goal is to partition the points into clusters $C_1, ..., C_k$ and find the cluster centers $\mu_1, ..., \mu_k$, in such a way that will allow us to minimize some cost of clustering which is the maximum radius of its clusters.Therefore, the objective of our clustering analysis is to partition the observations of the SST data into groups that are internally homogeneous and externally heterogeneous from group to group. The constitution of homogeneous groups of observations can be interpreted as a reduction of the dimension of the space where the total number of observations are grouped into several subsets.

Farthest-First

The final output of *Farthest-First* clustering is shown on Figure 4.1. This clustering approach produced clusters that are spatially contiguous, and the geographically neighboring clusters represent ocean areas with the relatively close SST values. The FF clustering technique that is used on SST data revealed relatively well-separated groups of data at various levels of detail and allowed the cluster information to be displayed with good clarity.



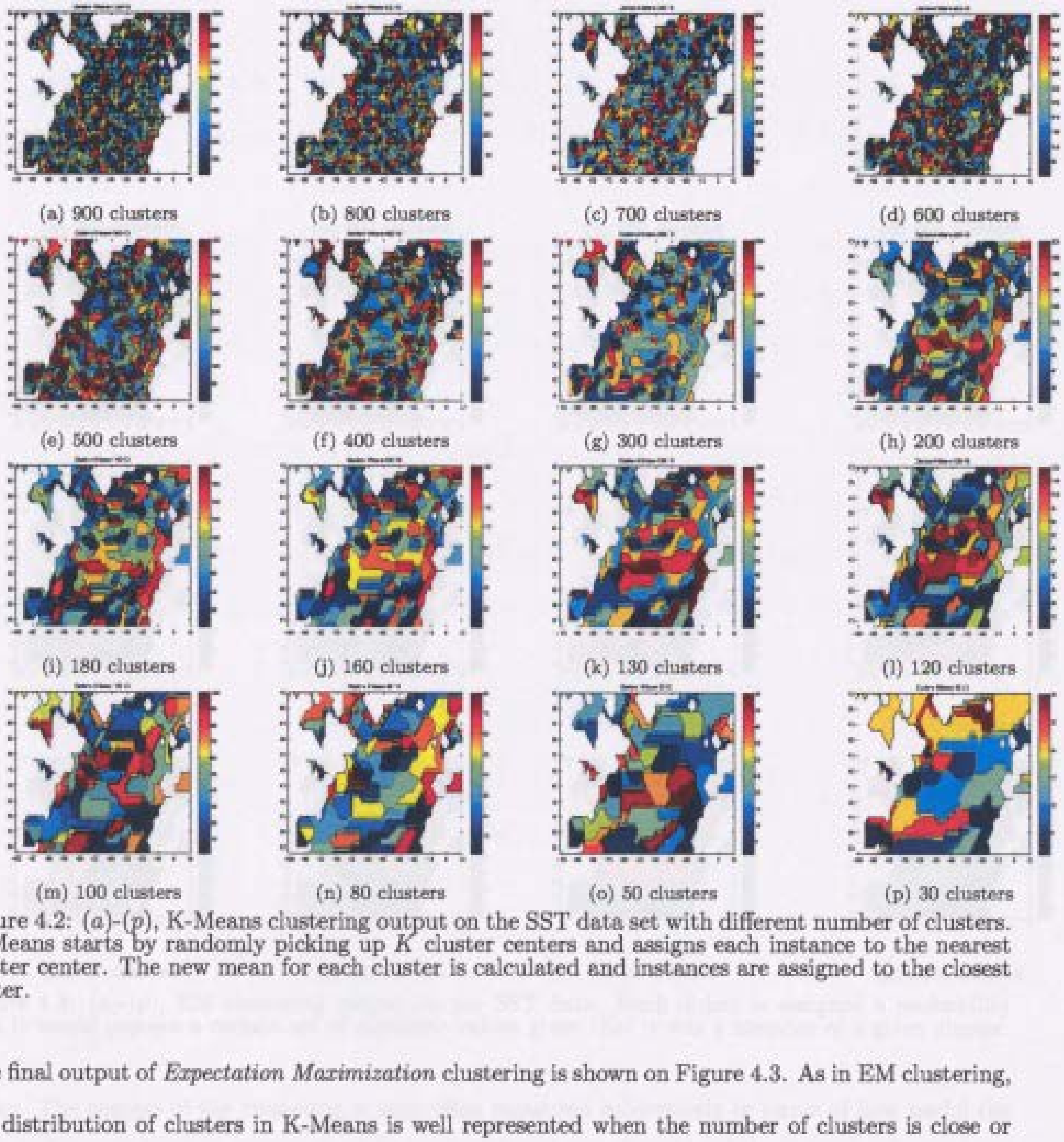
(a) 900 clusters (b) 800 clusters (c) 700 clusters (d) 600 clusters (e) 500 clusters (f) 400 clusters (g) 300 clusters (h) 200 clusters (i) 180 clusters (j) 160 clusters (k) 130 clusters (l) 120 clusters (m) 100 clusters (n) 80 clusters (o) 50 clusters (p) 30 clusters

Figure 4.1: (a)-(p), Farthest-First clustering output on the SST data set with different number of clusters. The Farthest-First traversal of n data points yields a sequence of "centers such that for any k, the first k of these centers define a k-clustering which is within a factor two of optimal[14].

K-Means Clustering

The final output of *K-Means* clustering is shown on Figure 4.2. Unlike Farthest-First algorithm, the K-Means clustering mainly failed to discover the contiguous areas in the ocean when the number of clusters is high. This clustering approach started to show some patterns and illustrate the ocean areas of homogeneous behavior for the the number of clusters close or below 200.

Expectation Maximization



(a) 900 clusters (b) 800 clusters (c) 700 clusters (d) 600 clusters

(e) 500 clusters (f) 400 clusters (g) 300 clusters (h) 200 clusters

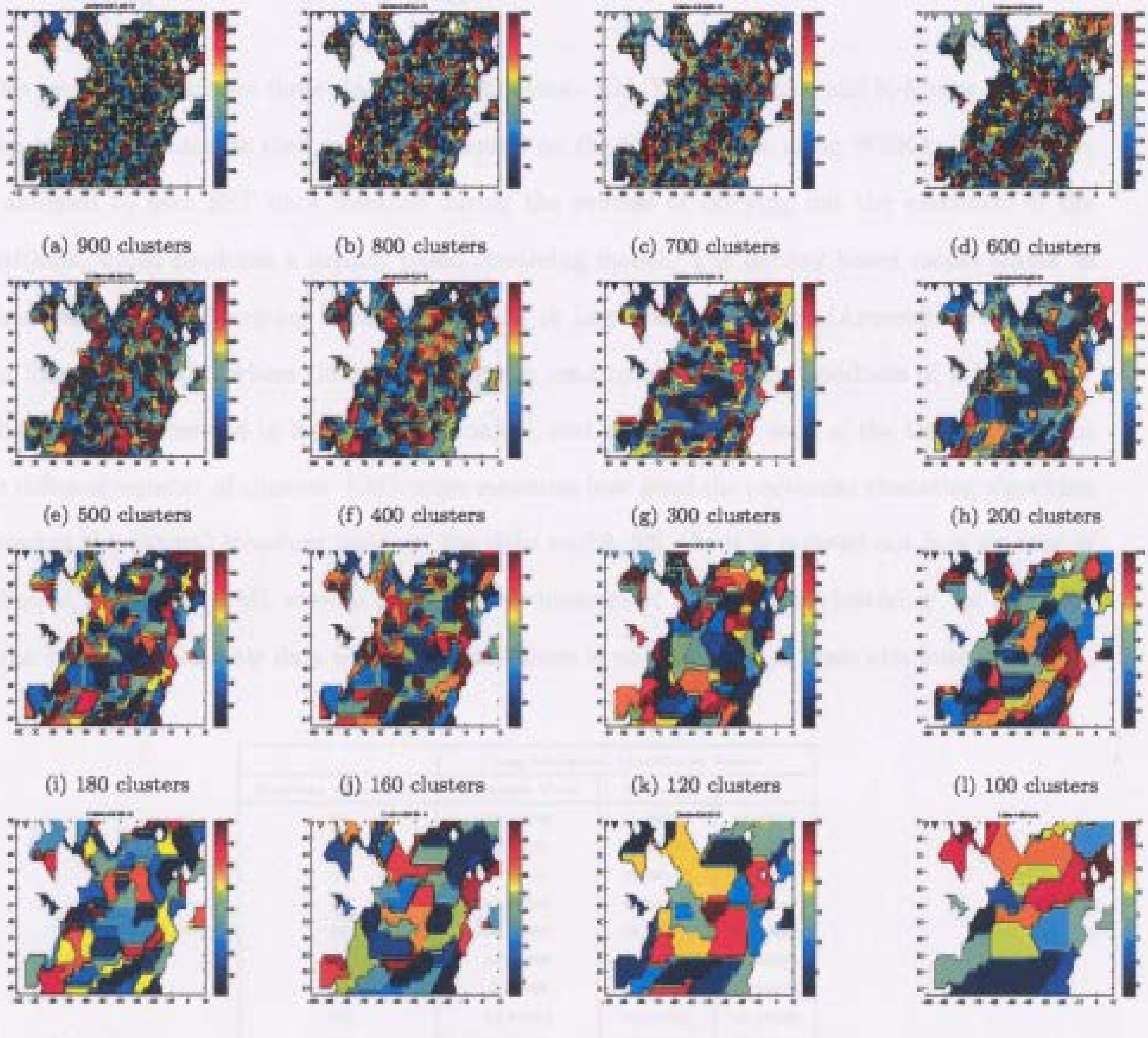
(i) 180 clusters (j) 160 clusters (k) 130 clusters (l) 120 clusters

(m) 100 clusters (n) 80 clusters (o) 50 clusters (p) 30 clusters

Figure 4.2: (a)-(p), K-Means clustering output on the SST data set with different number of clusters. K-Means starts by randomly picking up K cluster centers and assigns each instance to the nearest cluster center. The new mean for each cluster is calculated and instances are assigned to the closest center.

The final output of *Expectation Maximization* clustering is shown on Figure 4.3. As in EM clustering, the distribution of clusters in K-Means is well represented when the number of clusters is close or below the 200. The EM algorithm also allows one to select number of clusters automatically by cross validation, which did yield 17 clusters shown on Figure 4.3(p).

Implementation of all three algorithms resulted in assigning cluster labels to the SST data Instances. The outcome of each clustering algorithm that was applied to the SST data set is a numeric value. The actual value for new instances is of less interest than the structure that was learned from clustering, as each of the algorithms clusters the SST time series associated with the points in the



(a) 900 clusters (b) 800 clusters (c) 700 clusters (d) 600 clusters

(e) 500 clusters (f) 400 clusters (g) 300 clusters (h) 200 clusters

(i) 180 clusters (j) 160 clusters (k) 120 clusters (l) 100 clusters

(m) 80 clusters (n) 50 clusters (o) 30 clusters (p) auto clusters

Figure 4.3: (a)-(p), EM clustering output on the SST data. Each object is assigned a probability that it would possess a certain set of attribute values given that it was a member of a given cluster.

ocean. The success of the clustering is very often measured subjectively in terms of how useful the results appear to a human user[54],and in the literature on data mining there is no clear benchmark for the methods of clustering, which makes it difficult to compare the methods[14]. For evaluating the performance of the three algorithms and comparison of those methods, we used Log Marginal Likelihood Score and a two-tailed $t - test$, which are also implemented within a WEKA framework and described in the next section.

4.2 Comparison of the algorithms

In this section we compare three clustering algorithms - EM, Farthest-First, and K-Means - in terms of the clustering outcome that were implemented on the SST data set using WEKA. The clusters are assigned to each SST data instance during the process of carrying out the execution of the algorithms, which produces a density based clustering model. The density based model allows us to calculate the Log Marginal Likelihood (LML) or Log Likelihood score (Appendix VIII). Just as in linear regression, where the squared error is used to measure "the goodness of fit", the log-likelihood is used instead in logistic regression[56], and calculated for each of the three algorithms with different number of clusters. LML score measures how good the particular clustering algorithm represents the natural structure inside of the data set[58, 56]. As it is pointed out in a number of studies[44, 58, 34, 28], LML score is more suitable measure of "goodness of clustering" for real data sets(as oppose to synthetic data sets)[58], where there is no class label or class attribute is present.

	Log Marginal Likelihood Score		
Number of clusters	Farthest First	K-Means	EM
30	28 41795	10 33353	19 99088
40	34 2925	15 71887	25 42251
50	37 9951	22 25172	29 53311
60	41 90246	25 81179	33 50167
70	44 40617	28 28872	37 87987
80	47 49627	30 16531	42 03299
90	50 0089	33 31069	44 53679
100	52 64222	36 48804	46 86435
110	55 02508	37 81095	49 03036
120	57 25882	40 63261	51 53814
130	59 70435	43 52468	53 00262
140	61 62236	45 50791	55 40301
150	63 44433	46 97814	58 27816
160	64 51198	48 51238	60 52072
170	66 26683	50 06564	62 49992
180	65 94599	51 08784	64 54388
190	67 67236	52 77762	66 06232
200	70 23401	54 29775	67 66619
300	75 09048	62 04334	71 19217
400	81 77457	71 15912	74 78973
500	85 02074	74 56963	78 8818
600	87 37726	71 37307	80 96763
700	85 67269	62 99919	69 67469
800	79 08232	49 18326	50 49671
900	74 69723	32 35758	27 68141

Table 4.2: Results on LML Score

One of the features of LML score that is important to mention is that LML score depends on the number of attributes and instances in a data table, and therefore, different algorithms can be compared with each other in terms of LML score only if they run on the same data set. Table 4.2 shows the LML scores results from running three above mentioned algorithms on the SST data set with different number of clusters. The runs with those three algorithms were performed on the same parameters, and LML scores were obtained from 100 runs with the same random seed from 1 to 100, similar to what was done by Li et al.[55] in their work.

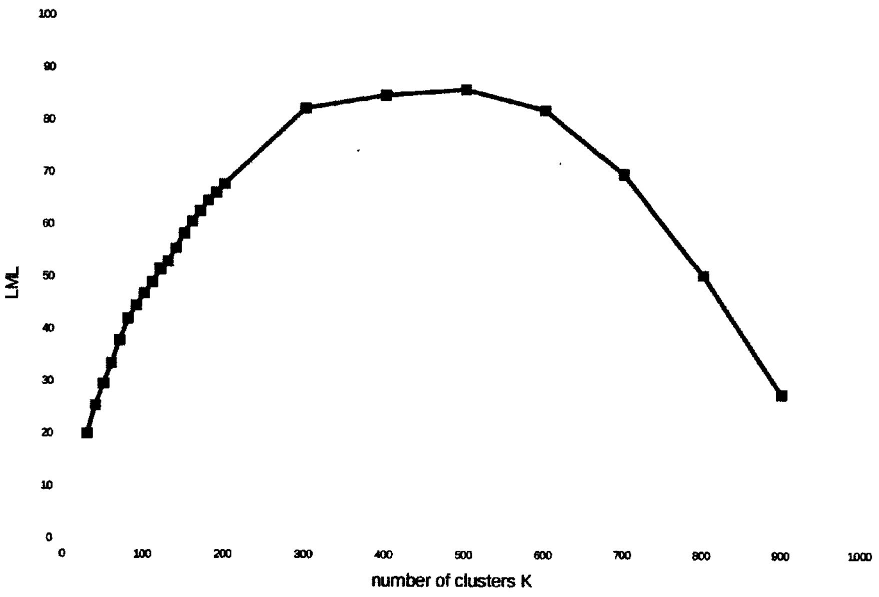


Figure 4.4: Log Marginal Likelihood as a function of cluster size K for a typical run.

Figure 4.4 shows how the log marginal likelihood varies as function of cluster size K under one particular experimental condition. Curves for other experimental conditions of all three algorithms are relatively similar. As K increases, the marginal likelihood first increases and then decreases, which is a penalty for solutions with many clusters and complexity.

We compare three algorithms with each other by running two-tailed $t - test$ with 0.05 significance level. Table 4.3 refers to the results of the two-tailed $t - \textbf{test}$ with the annotation $w/t/l$, which indicates that a specific result is statistically better or worse at the significance level specified, where w denotes that the algorithm wins in the corresponding row in w data sets, ties in t data sets, and loses in l data sets, compared to the algorithm in the corresponding column.

Farthest-First slightly predominate over both EM and K-Means algorithms, having 13 wins and 9 loses over EM, and 15 wins and 7 loses over K-Means.

	Log Marginal Likelihood Score	
	EM	*K-Means*
EM		$12 - 3 - 9$
FarthestFirst	$13 - 2 - 9$	$15 - 2 - 7$

Table 4.3: Summary of Experimental Results

EM minimally outperforms K-Means having 12 wins and 9 loses, which corresponds to the results of the test run and experimental comparison of clustering algorithms conducted by Li et al.,[58], and Meila and Heckerman[44]. As the authors suggested, EM is preferable in many applications. We also would suggest that Farthest-First and EM algorithms can be used in the SST time series data sets. Yet, we will continue to use the outcomes from all three algorithms further in this work, partially because we would like to evaluate the quality of clustering by adopting a widely used approach in the area of data mining, such as comparing the results to a "ground truth". The results of clustering will be assessed on the basis of the external knowledge of how clusters should be structured using EOF analysis of the patterns of sea level pressure associated with North Atlantic Oscillation.

Chapter 5

EOF Analysis of SLP

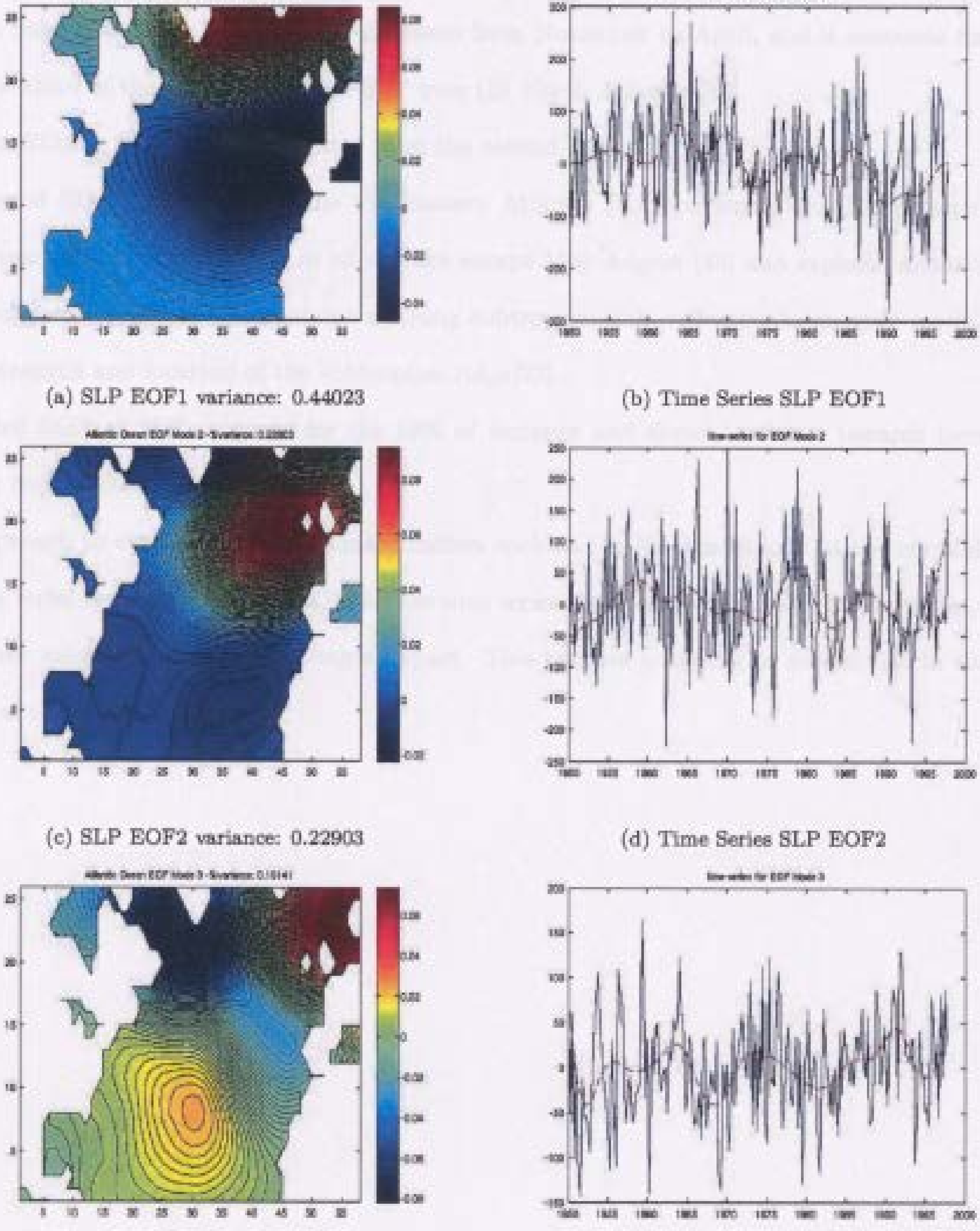
5.1 EOF Analysis of SLP

The EOF method, described in Section 2.2, reduces the dimensionality of the data by introducing k principal components (EOFs) that explain most of the variance in the original data with the assumption that the rest of the EOFs can be disregarded without losing a lot of information. The EOFs with smaller magnitudes do not contribute much of the information, therefore the reduction in the dimensionality of the original data is very significant.

The data that is used for the EOF analysis are time series of monthly mean anomalies of the Sea Level Pressure (SLP). The data covers the winter months only (DJF), for the period between 1950 and 1997. The data cover the area of the North Atlantic.

Having the SLP data in a matrix form, where the rows consist of time series from various points in the area of our interest, that is the North Atlantic, we would like to find the strongest temporal and spatial patterns in the SLP data by using EOF method. The eigenvalue for each EOF of SLP represents the variance explained by the EOF. The strongest patterns, as it is shown on Figures 5.1(a, c, e) capture the largest variation of the SLP data. The first three EOFs account for most of the variation in the data, explaining 44%, 23%, and 18% of total variance.

The first EOF of SLP for the North Atlantic and Labrador Sea is shown on the Figure 5.1(a). The



(a) SLP EOF1 variance: 0.44023

(b) Time Series SLP EOF1

(c) SLP EOF2 variance: 0.22903

(d) Time Series SLP EOF2

(e) SLP EOF3 variance: 0.18141

(f) Time Series SLP EOF3

Figure 5.1: (a)-(f). The first three strongest EOFs of the mean SLP anomalies(1950-1997) over the North Atlantic and Labrador Sea, and the percentage of the total variance they explain.

first EOF spatial pattern shows two large regions in the ocean whose mean level pressures are inversely related, which characterizes and represents the North Atlantic Oscillation(NAO), that is the fluctuations of the atmospheric sea level pressure difference between the Iceland and Azores. The NAO is most noticeable during the cold season from November to April, and it accounts for more than one-third of the total variance in SLP over the North Atlantic [33].

The first EOF in SLP is well separated from the second EOF.

The second EOF 5.1(c,d) represents the Eastern Atlantic (EA) pattern. The EA pattern has a center near Ireland. EA appears in all months except May-August [33] and explains about 33% of total variance of SLP. It also contains a strong subtropical link, reflecting large-scale modulations in the strength and location of the subtropical ridge[33].

The third EOF of SLP accounts for the 18% of variance and shows tendency towards increasing starting from 1980s.

One approach to evaluating ocean climate indices such as the NAO is to look at the correlation of the time series representing the NAO with the time series associated with the SST time series, where the higher value will indicate a stronger impact. This task we are going to accomplish in the next section of this work.

Chapter 6

The North Atlantic Ocean SST index

The SST interannual short-term fluctuations are driven primarily by the atmosphere through anomalous air-sea fluxes, while the long-term SST patterns with periods over a few decades are driven by ocean dynamics, both the wind-driven circulation and the meridional overturning circulation[5],[39]. This section is designed to verify the previous studies using data mining outcome of the SST dataset. Two approaches are used to calculate NAO SST index. The first NAO SST index is defined as the difference of the SST centroids for clusters which show highest positive and negative correlation with the first three SLP EOFs. The second NAO SST index is defined as the difference of the SST centroids in the years of anomalous positive and negative first three SLP EOFs.

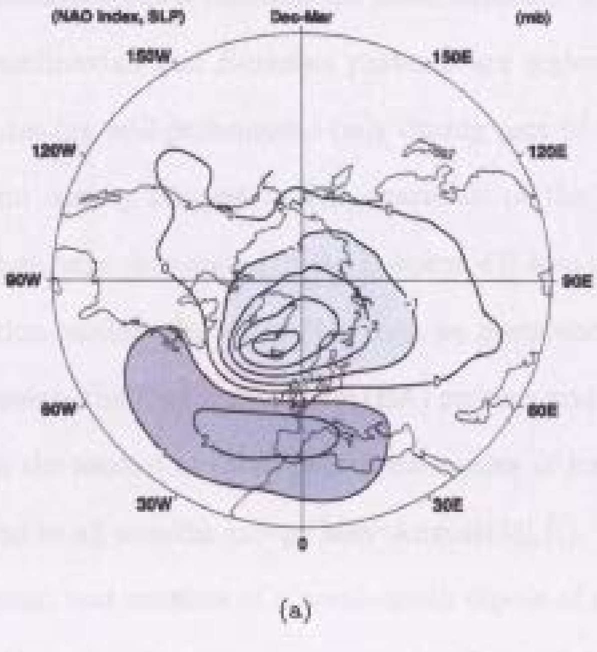
6.1 North Atlantic Oscillation(NAO) and SST

The North Atlantic Oscillation(NAO) is defined as the atmospheric sea level pressure(SLP) difference between Iceland and the Azores. When the Iceland low pressure center is deeper than usual, the Azores High is stronger than usual, and vice versa. The NAO is a measure of the strength of the westerly winds blowing across the North Atlantic Ocean in the 40 degrees N and 60 degrees

N latitude belt, and it is an important feature of atmospheric variability throughout the year, although it is less dominant during warmer seasons[24]. Previous studies have demonstrated that the NAO is correlated with large-scale changes in the Sea Surface Temperature(SST) across the North Atlantic[5], and NAO driven changes in SST are also most noticeable during the winter season. For rest of the year, the influence of the atmospheric forcing related to NAO is mild and therefore, the influence on the Atlantic Ocean is minimal. During the positive phase of the NAO, the high pressure system intensifies and the low pressure system weakens, which creates a large pressure gradient between Azores and Iceland. The seesaw in winter temperatures between western Greenland and Europe is a clear evident of high NAO index with stronger than normal westerly winds[24]. A high NAO pattern is distinguished in the northeast Atlantic by a reinforcement of the westerlies that are pushed further south, and hence by warmer winter temperatures than normal. It is also recognized that the existence of an exceptionally strong positive phase of the NAO is the source of temperature anomalies and changes in atmospheric moisture transport[31]. The impact of the two phases of the NAO can be felt across the entire Atlantic and the surrounding continents, with its greatest effect on the storms that move to Europe, creating an area of forward-moving current between clockwise and counterclockwise circulation patterns that channels the weather systems from the United States to Europe. When the pressure difference between the two systems is large, the storms propogate towards Scandinavia and northern France. When the NAO index is negative, the winter storms propagate over the southern United States and southern Europe, the Middle East, and northern Africa. The direction these storms can take causes large changes in the temperature and the weather over Europe from December through March[33]. A positive NAO on average can increase rainfall in northern Europe and warm the air at the same region. A negative NAO, on the other hand, will bring rain to southern Europe, drop the temperatures in northern Europe [33].

Numerous research studies are done to investigate what governs the NAO variability, whether it is predictable and at what extent, and whether the ocean plays a role in determining the evolution of the NAO. As pointed out by Greatbatch[24], the role of the ocean, and in particular sea surface temperature(SST), in regulating the NAO has attracted much attention, but remains controversial. The important thing is if variability of the NAO is driven by that of underlying SST, then the NAO can be predicted on the longer than three-week time scale, but the SST must be predictable by itself[24]. According to J. Hurrell et.al [33], "statistical analysis have revealed patterns in North Atlantic Sea Surface Temperatures(SST) that precede specific phases of the NAO by 6-9 months". Studies from observational data done by Czaja and Frankignoul [12] indicate a significant correlation between the wintertime NAO and the leading mode of anomalous SSTs from the previous summer, which can serve as an evidence for the oceanic forcing of NAO.

The observed departure of SLP and SST, and air temperature over the land, associated with one standard deviation(positive) of the NAO index is shown in Figure 6.1. The change in winter temperature associated with the NAO extends all the way across the Eurasian continent from the Atlantic to the Pacific[24]. This is evidence that the NAO is not a regional North Atlantic phenomenon[24]. The NAO is important not just for winter surface temperature variability in the North Atlantic sector, but for winter surface temperature variability over the northern hemisphere as a whole. The NAO is also closely related to a hemispheric mode of variability that is called the Arctic Oscillation(AO) described by Thompson and Wallace[52], where the AO corresponds to the first EOF of SLP variability over the northern hemisphere. The spatial structure of the AO and the NAO are closely related and both, the AO and the NAO correlate to the same physical phenomenon[24]. Interestingly, some studies investigate the NAO in the relation to the recruitment of the certain fish stock in the North Sea through the influence of the NAO on the SST[19],[18],[43].



(a)

(b)

Figure 6.1: (a)-(b), SLP and SST change associated with one standard deviation(positive) of NAO index.(Adopted from J. Hurrell[32])

6.2 Eastern Atlantic Pattern

Modes other than NAO also play important role in determining the changes of SST in the North Atlantic. Yet, unlike well established NAO pattern, the other modes of atmospheric circulation, such as the Eastern Atlantic, Scandinavian and Eurasian patterns are regional. According to Barnston and Livezey[1], their signatures are well pronounced only during part of the year and the occurrence of this modes is not uniform during the year. The character of the modes, the shape and the intensity of their action centers vary seasonally[1]. As Pokorna[47] also points out, very few authors pay attention to the circulation modes other than NAO. As we mentioned in Section 5.1, the second EOF in Figure 5.1(c,d) represents the Eastern Atlantic (EA) pattern and explains about 23% of total variance. The EA pattern is the second of three prominent modes of low-frequency variability over the North Atlantic, appearing in all months except May-August[33],[1]. The EA pattern has similar structure to the NAO in winter, and consists of a north-south dipole of anomaly centers which span the entire North Atlantic. However, its anomaly centers are located farther southward [47]. The EA pattern has lower-latitude center and contains a strong subtropical link, reflecting large-scale modulations in the strength and location of the subtropical ridge[33]. This subtropical link also makes the EA pattern distinct from the NAO pattern. The EA pattern which represents the second EOF in our study is similar to that shown in the Barnston and Livezey study[1]. The main cell of the EA is located westward of the British Isles and the body of high opposite values extends to lower latitudes, over central Atlantic, northern Africa and southern Europe. During the positive phase of the EA, the daily maximum, minimum and mean temperature are higher than average and south winds dominate[47]. The positive phase of the EA pattern is also associated with above-average precipitation over northern Europe and Scandinavia, and with below-average precipitation across southern Europe[1]. The EA pattern exhibits very strong multi-decadal variability in the 1950-2000 record, with the negative phase prevailing during much of 1950-1955 and 1967-1977, and the positive phase occurring during much of 1957-1967,1977-1982,1985-1990, and 1977-2000. The positive phase of the EA pattern was particularly strong and persistent during 1997-2000. The NAO and the EA are the zonal modes and influence temperature more than other variables. They also show high correlation with wind directions and these correlations are generally higher in the winter when the circulation modes are better pronounced[47].

Interestingly, the EA pattern also depends on the procedure used to derive it[42]. As Lionello P.,et al. pointed out, "still, the kind of variability associated with the EA pattern seems important and physically real, as it is also detected in studies using alternative techniques, like cluster analysis done by Kimoto and Ghil"[37].

To expand previous interrelations between the NAO, the EA pattern and SST anomalies, the next section will present the results from the study of the connection between SST clusters and major patterns in SLP. The clustering outcomes that are used to calculate correlation coefficients are obtained from the FarthestFirst, the Expectation Maximization(EM), and the KMeans algorithms, described in the sections(2.1.3),(2.1.2), and (2.1.1), and displayed in the Figures 4.1, 4.2, and 4.3. Here we will explore if the SST clusters can be potentially applied in deriving climate indices for the North Atlantic related to NAO, EA and the third EOF of SLP. The outcome of clustering is evaluated on the basis of the external knowledge about how clusters should be structured. Thus, we adopted the approach of comparing the results to a "ground truth" and the correctness of clustering is estimated against the existing knowledge.

6.3 Correlation analysis of SST clusters centroids and SLP EOFs

This section introduces the correlation analysis of clustered SST with the modes of variability of atmospheric circulation. The evidence of relationships between SLP and SST patterns is analyzed by the previous studies in local or regional settings by many scientists. The SST variability in the North Atlantic also has been evaluated in a number of research studies. In particular, it was found that large-scale temperature anomalies occurred in the North Atlantic Ocean on interannual to decadal time scales [16],[39],[27],[51]. It has also been shown that the SST variability can be correlated as part of the mixed layer response to variability in surface fluxes [7], [2], [49]. Although, "forcing by surface flux variations cannot account for all the observed features of Atlantic Ocean variability"[20].

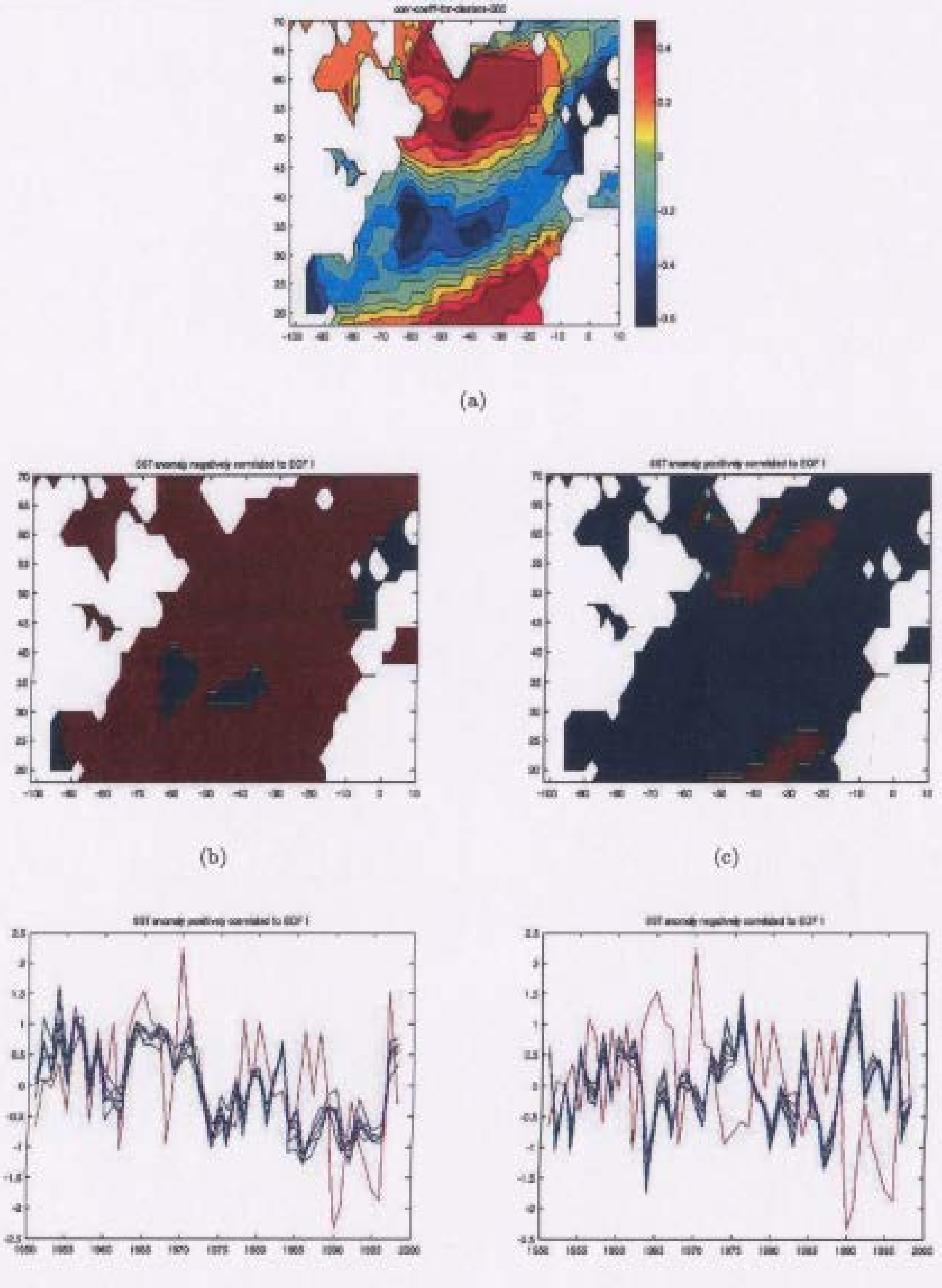
The research by Cayan[7] is done for the wider spatial settings, where the author examines thermodynamic forcing of the upper ocean by relating latent and sensible heat flux to changes in SST over Northern Oceans. The main conclusion for his study is the strong similarity in the configuration anomalous heat flux and SST tendency patterns in their association with major SLP modes[7]. In his work, Cayan[7] uses two separate analysis, Canonical Correlation Analysis and composites according to atmospheric circulation anomaly modes. He demonstrates that heat flux and SST anomalies covary with the patterns that extend over the ocean basins. The CCA performed by Cayan illustrates the spatial configuration of the connection between heat flux and SST. The structure of the patterns of these fields, their interrelationships, and the magnitude of their anomaly centers are similar in the North Atlantic and North Pacific. It was established that the heat flux is consistent with atmospheric circulation and significantly correlated with tendencies in SST anomalies.

In our analysis, the SST shows large-scale patterns of variability that are related to the patterns of major modes of SLP variability. These relationships are steady and regular, and the consistency of these correlations are confirmed by the use of three different data clustering algorithms, such as FarthestFirst, EM, and KMeans. The amount of variance ranges from 18% to 44% over most of the North Atlantic, which also authenticated by similar results in the studies done by Cayan[7].

The clustering outcomes that are presented in the Section4.1 for the SST field for all these three algorithms are not similar, that is, each of the algorithms interpreted the same SST data uniquely distinct from each other and the maps yielded from each classification show noticeably different outcomes for each algorithm. Despite those explicit differences, portrayal of correlation coefficients between major SLP modes and the SST anomalies depicts consistency of the results with small dissimilarities from one algorithm to another.

The correlation coefficient between time series of dominant SLP EOFs(Fig.5.1(a, c, e) and SST clusters(Fig.4.1, 4.2, 4.3) are calculated. The optimal number of SST clusters is selected based on the analysis done in the previous section(4.2), and the final number of clusters for the further analysis is determined to be equal to 500. Figures 6.2(a), 6.3(a), 6.4(a), 6.5(a), 6.6(a), 6.7(a), 6.8(a), 6.9(a), 6.10(a) show the correlation coefficients for the three dominant modes of SLP and for the three clustering algorithms. The complete results for the calculations of the correlation coefficient between various number of SST clusters and dominant SLP EOFs can be viewed in the Appendix X.

For EOF1, correlation maps for all three algorithms show positive correlation near the southeast of Greenland and in the subtropics. Between these two positive centers lies an area of negative correlations. Figures 6.2(a,b), 6.5(a,b), 6.8(a,b), also show a negative correlation west of Norway, which may reflect the impact of the NAO on Norwegian and North Seas. Time series of SST anomalies that are positively and negatively correlated with first EOFs of SLP for all three algorithms are shown on Figures 6.2(d,e), 6.5(d,e), 6.8(d,e). The time series show that the correlation between SST and EOFs has the opposite sign and it is consistent through all three algorithms. Figures 6.3, 6.6, and 6.9 show the results of correlation between EOF2 and SST clusters for all three algorithms. Figures 6.3(a,b,c), 6.6(a,b,c), 6.9(a,b,c) identify two large and spatially consistent regions of positive and negative correlation centers. The center of the positive correlation corresponds to the region located near the Azores and Canary Islands, and the negative center is in the subtropics. Figures 6.4, 6.7, and 6.10 show the results of correlation between EOF3 and SST clusters for all three algorithms. Figures 6.3(a,b,c), 6.6(a,b,c), 6.9(a,b,c) identify three regions of positive and negative correlation centers, where positive centers are located in the subtropics, and Faroe Island, and the negative center is located south of Newfoundland. The time series for the SST anomaly and third EOF for SLP are consistent with the results for the first two EOFs, and show negative sign of correlation over all time period. The time series that are presented on the Figures 6.2(d,e), 6.3(d,e), 6.4(d,e), 6.5(d,e), 6.6(d,e), 6.7(d,e), 6.8(d,e), 6.9(d,e), 6.10(d,e) for all three algorithms are the averaged SST-s in the areas of highest correlation of the SST with the NAO and they are used to define the NAO climate index. All the results are consistent with the previous studies of the North Atlantic Oscillation and Eastern Atlantic Pattern [12],[2],[24].



(a)

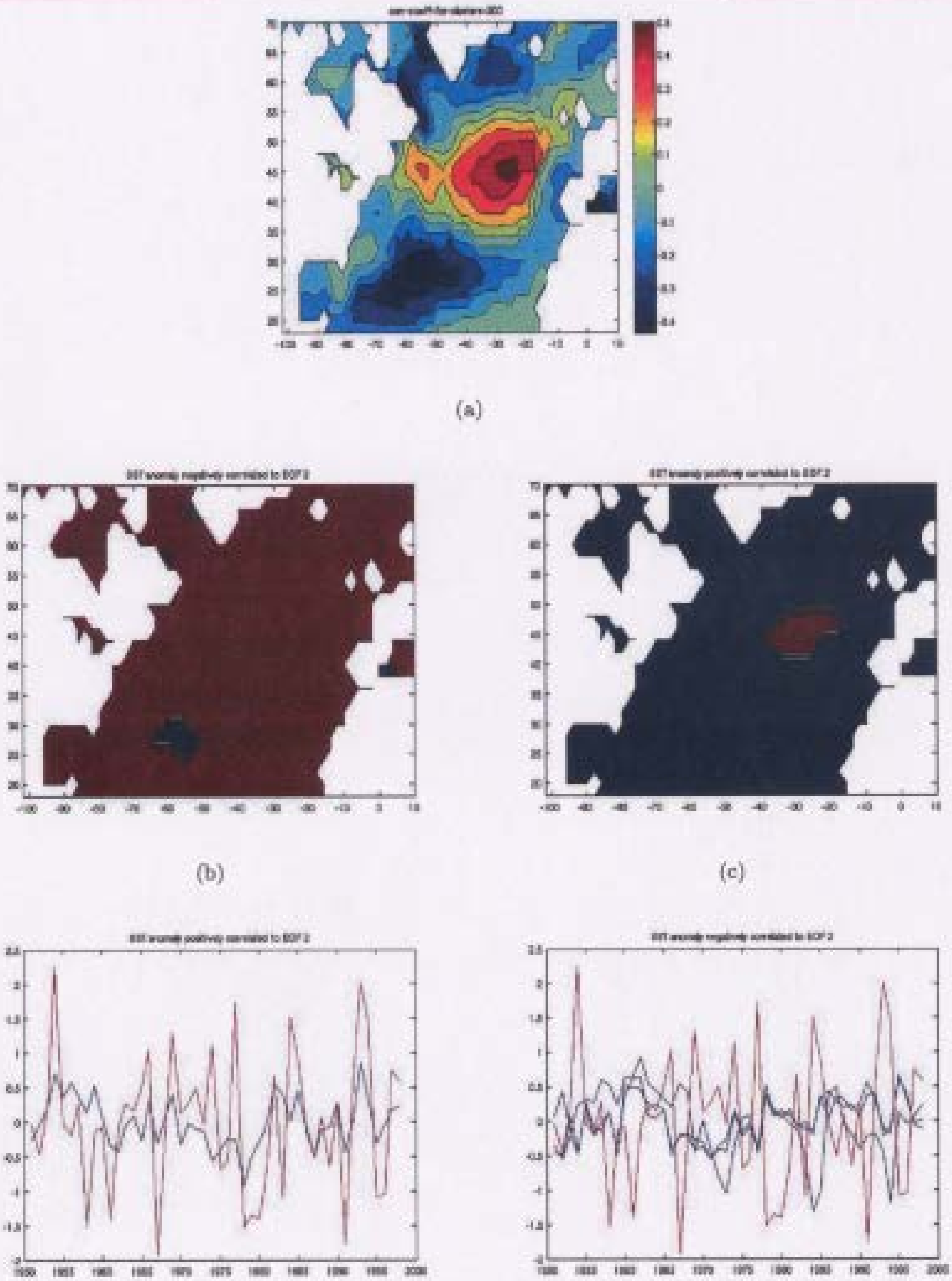
(b)

(c)

(d)

(e)

Figure 6.2: (a)Correlation coefficients for the FarthersFirst clustering output for 500 clusters and EOF1 of SST data set;(*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF1 for FartherstFirst clustering algorithm using 500 clusters.



(a)

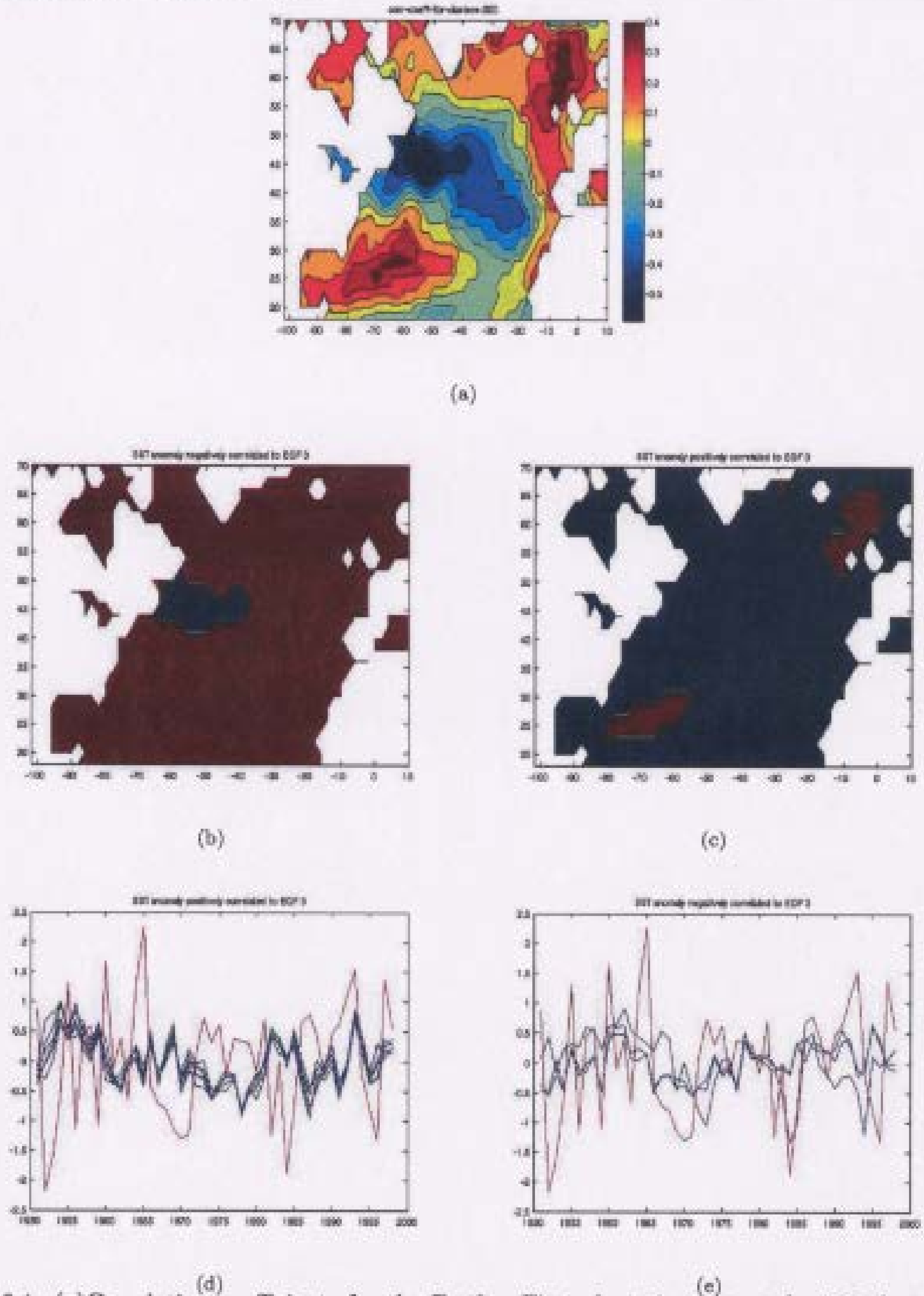
(b)

(c)

(d)

(e)

Figure 6.3: (a)Correlation coefficients for the FarthersFirst clustering output for 500 clusters and EOF2 of SST data set;(*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF2 for FartherstFirst clustering algorithm using 500 clusters.

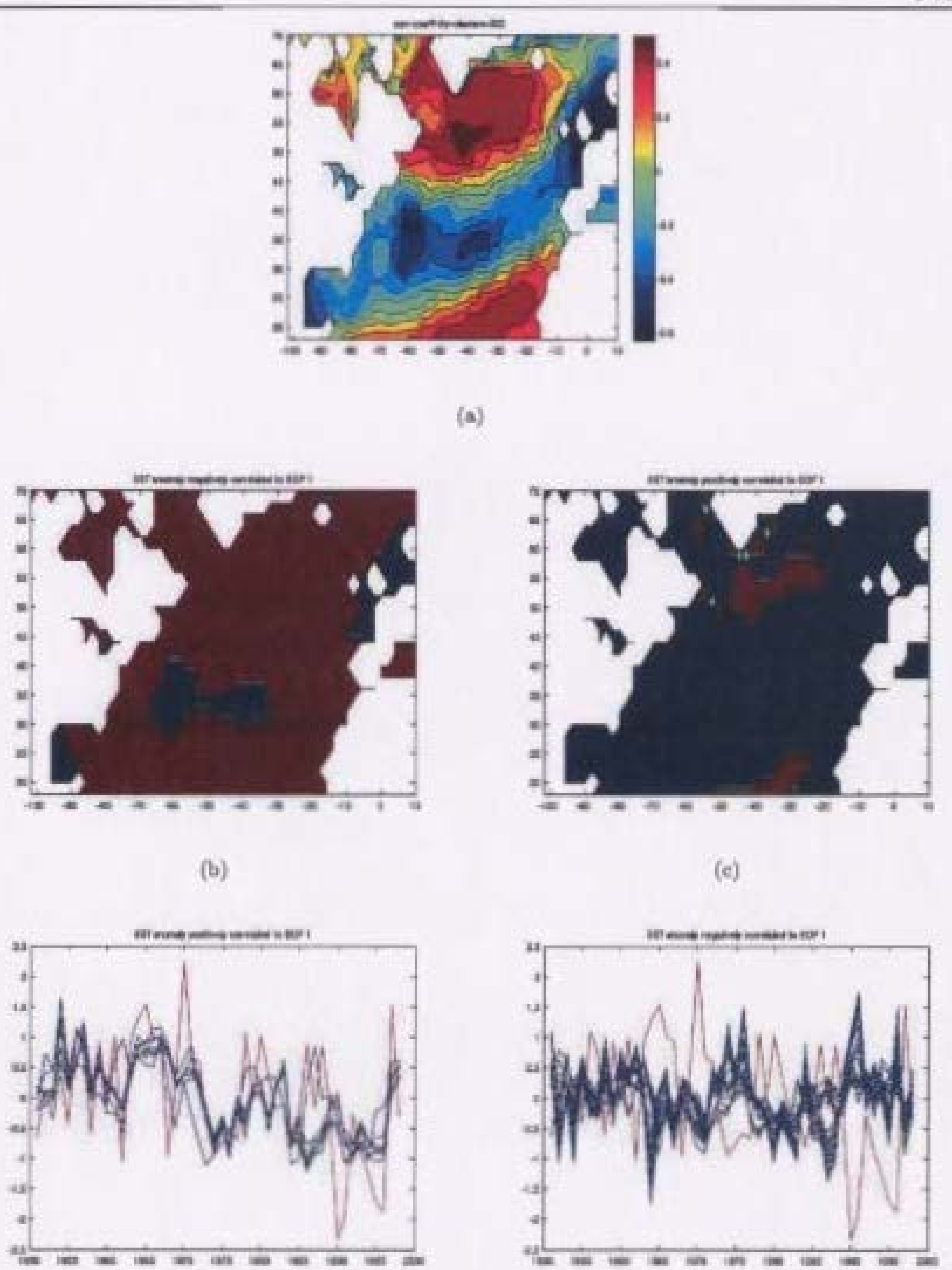


(a)

(b) (c)

(d) (e)

Figure 6.4: (a)Correlation coefficients for the FarthersFirst clustering output for 500 clusters and EOF3 of SST data set;(*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF3 for FartherstFirst clustering algorithm using 500 clusters.



(a)

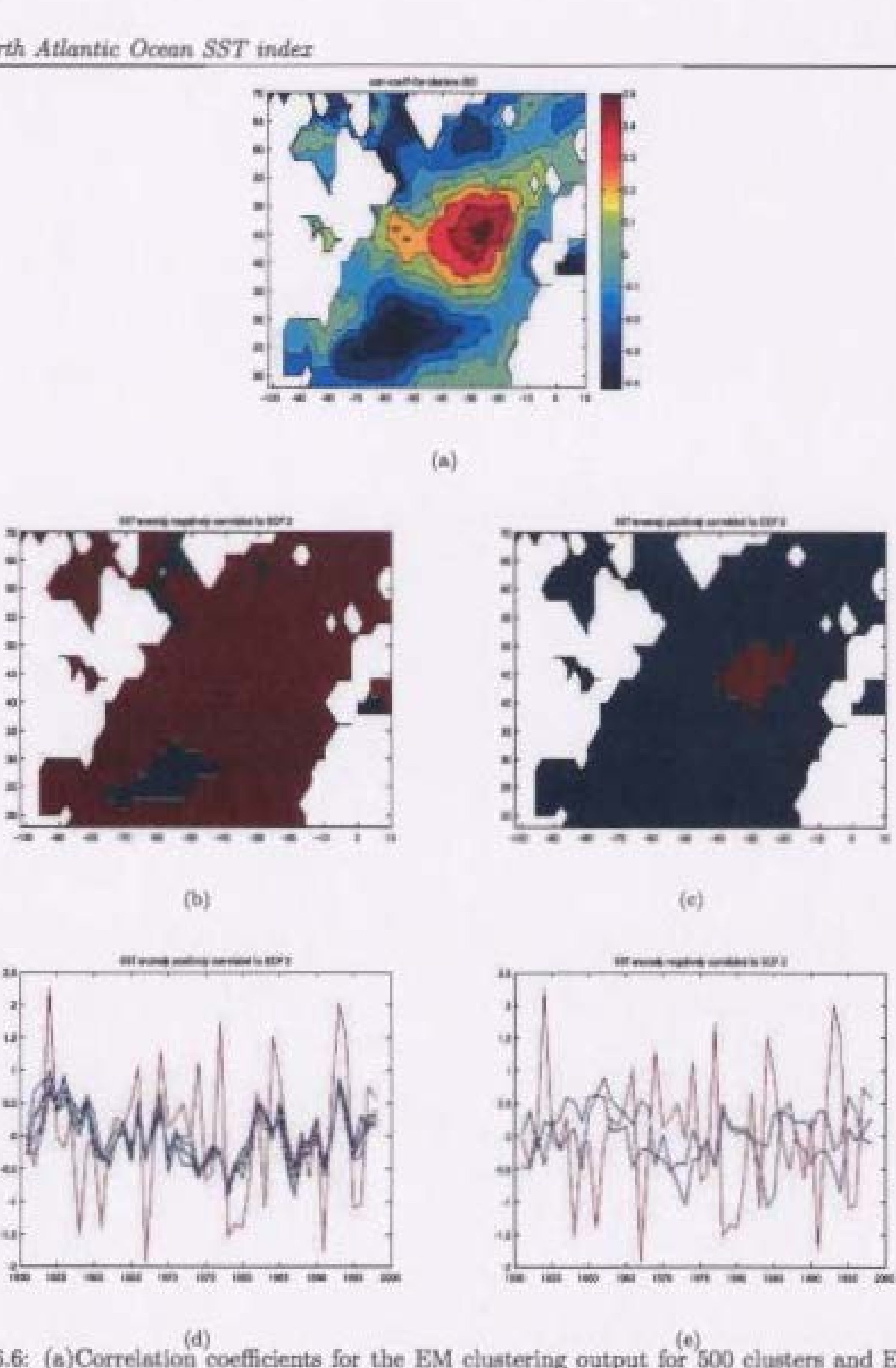
(b)

(c)

(d)

(e)

Figure 6.5: (a)Correlation coefficients for the EM clustering output for 500 clusters and EOF1 of SST data set; (*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF1 for EM clustering algorithm using 500 clusters.



(a)

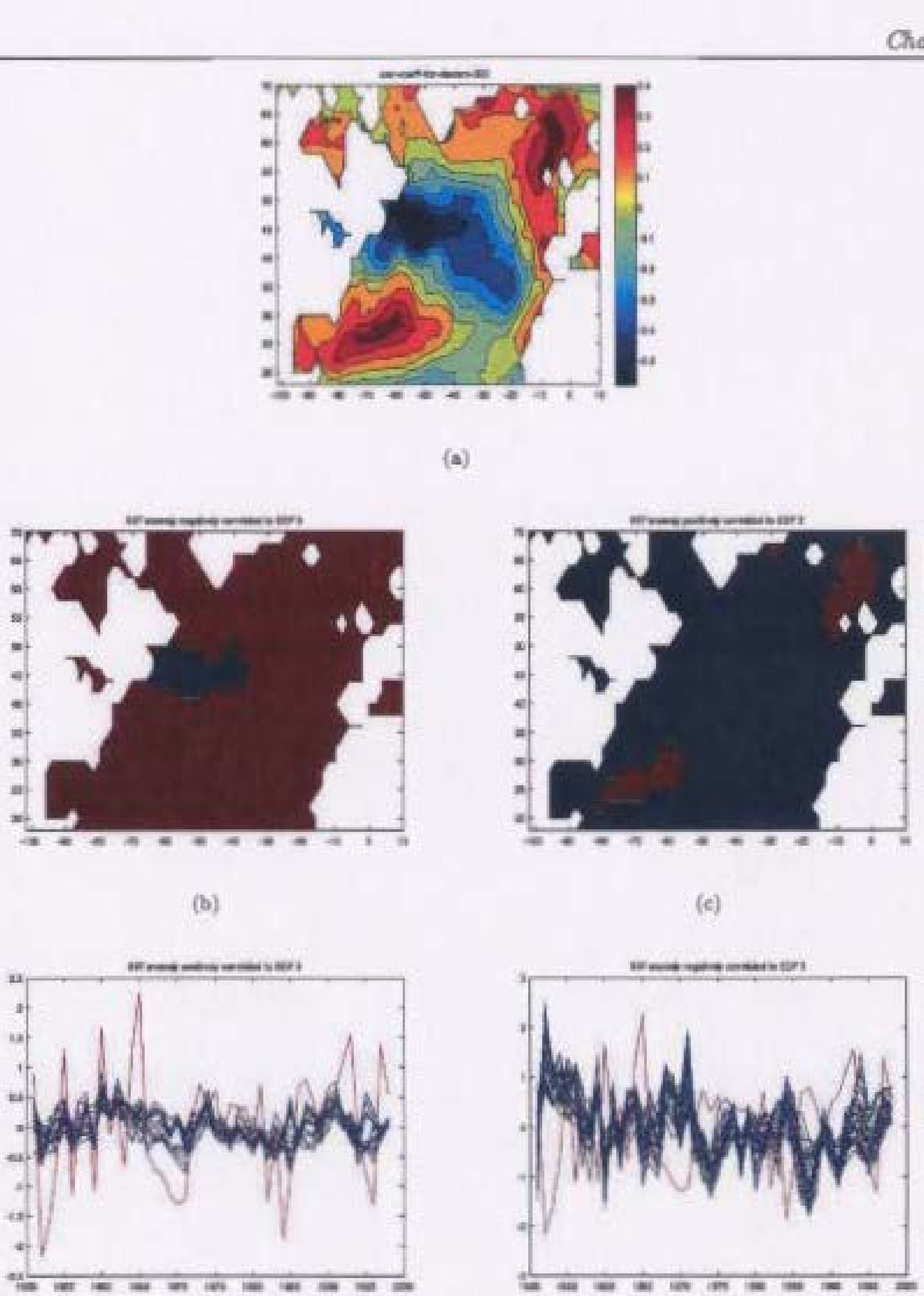
(b)

(c)

(d)

(e)

Figure 6.6: (a)Correlation coefficients for the EM clustering output for 500 clusters and EOF2 of SST data set;(b)-(e), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF2 for EM clustering algorithm using 500 clusters.



(a)

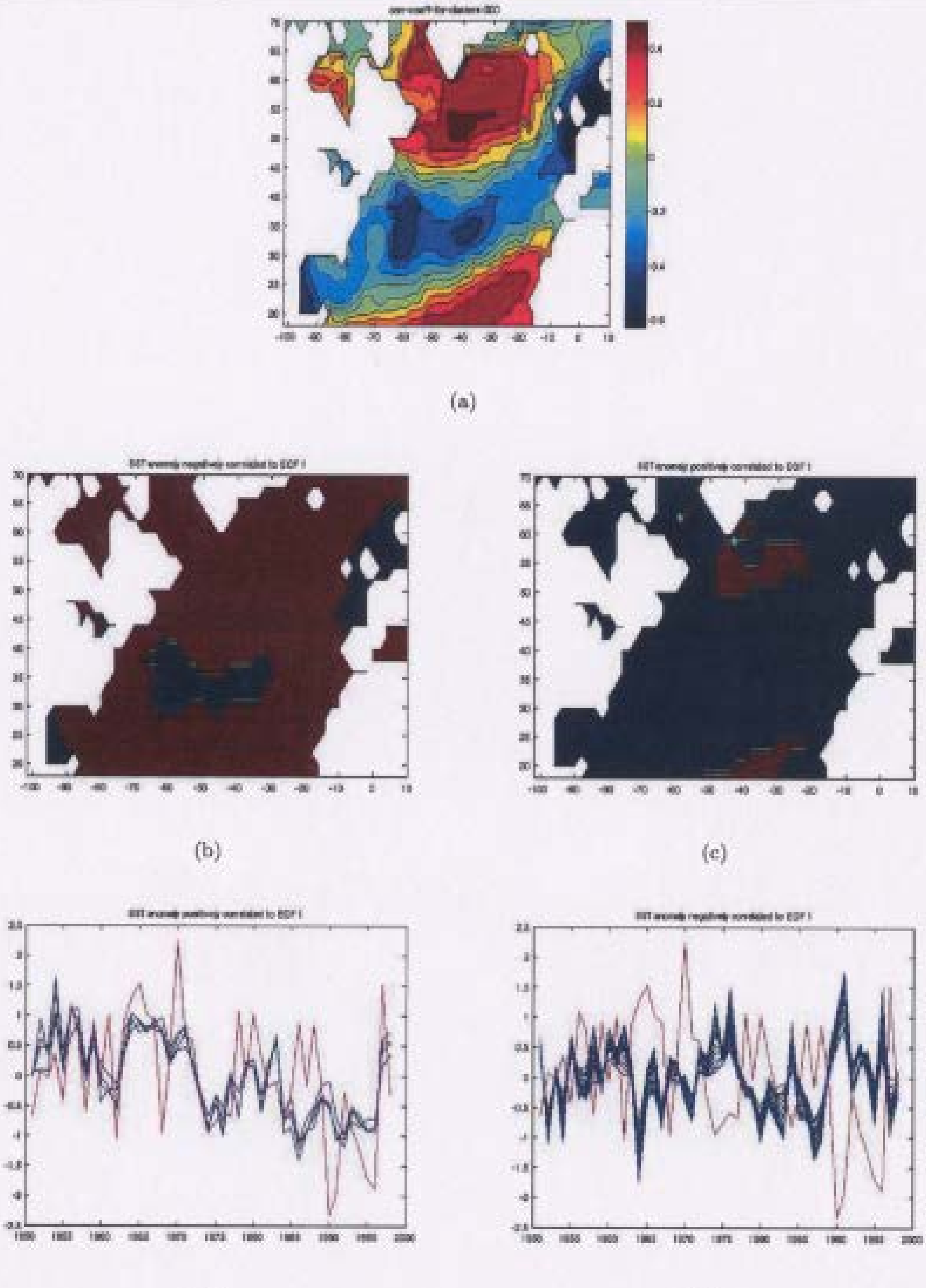
(b)

(c)

(d)

(e)

Figure 6.7: (a)Correlation coefficients for the EM clustering output for 500 clusters and EOF3 of SST data set; (*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF3 for EM clustering algorithm using 500 clusters.


SST anomaly negatively correlated to EOF1
SST anomaly positively correlated to EOF1
SST anomaly positively correlated to EOF1
SST anomaly negatively correlated to EOF1

(a)

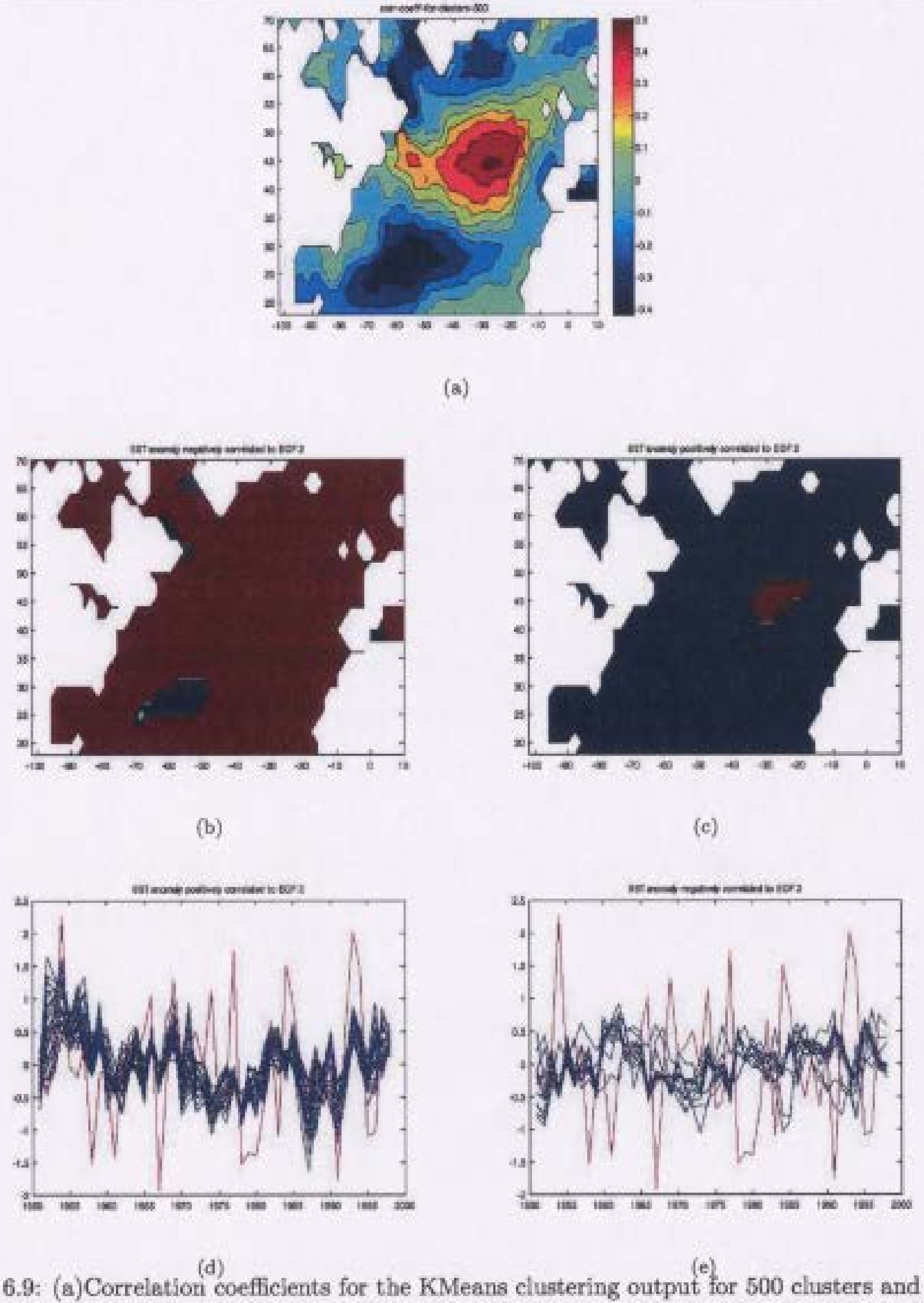
(b)

(c)

(d)

(e)

Figure 6.8: (a)Correlation coefficients for the KMeans clustering output for 500 clusters and EOF1 of SST data set; (*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF1 for KMeans clustering algorithm using 500 clusters.



(a)

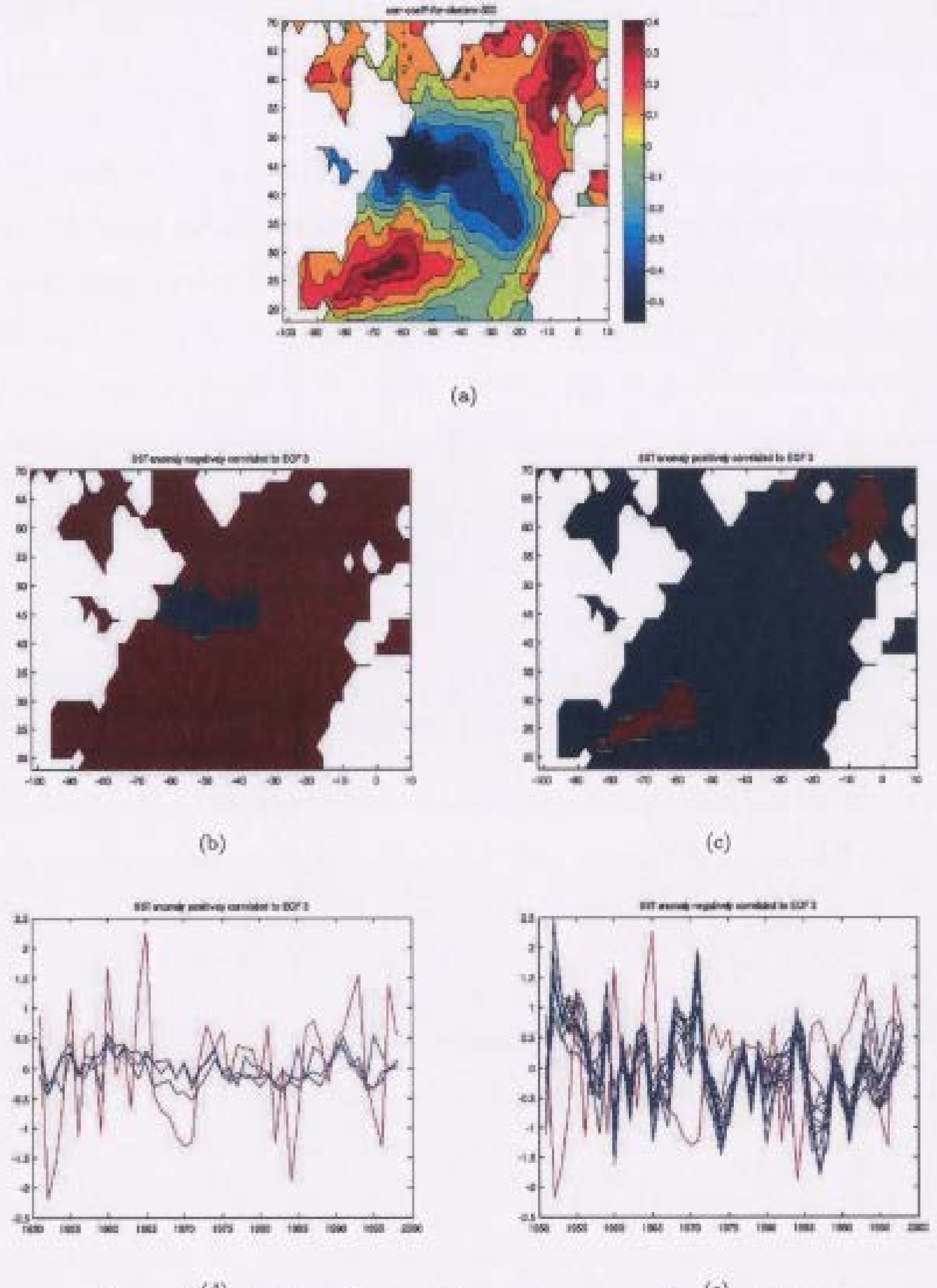
(b)

(c)

(d)

(e)

Figure 6.9: (a)Correlation coefficients for the KMeans clustering output for 500 clusters and EOF2 of SST data set; (*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF2 for KMeans clustering algorithm using 500 clusters.



(a)

(b)

(c)

(d)

(e)

Figure 6.10: (a)Correlation coefficients for the KMeans clustering output for 500 clusters and EOF3 of SST data set; (*b*)-(*e*), Areas of maximum positive and negative correlation and the time series of SST anomaly positively and negatively correlated to EOF3 for KMeans clustering algorithm using 500 clusters.

6.3.1 SST patterns related to anomalous high and low indices of NAO and EA

To test the consistency of patterns between the heat flux and SST tendency anomalies, Cayan [7] examines this linkage for the case of strong anomalous atmospheric circulation. The composite differences associated with positive and negative extremes of the first two SLP EOFs from Cayan are shown on Fig 6.11. The differences that are significantly different from zero are indicated by shaded areas. Cayan confirmed by his findings that the major regions of significant SST anomaly tendencies are closely matched to those of heat flux, supporting the view that they are linked.

In the attempt to determine the connection between the SST and the major three modes of SLP during strong anomalous atmospheric circulation, we identified extreme positive and negative EOF amplitudes for each of the EOF of SLP. Then, the composites of SST were formed by averaging this field during the respective extreme EOFs. As in Cayan[7], the composites were expressed as a difference between averages of positive(strong) and negative(weak) phase of the EOF of SLP.

The composite differences associated with positive and negative extremes of the first three SLP EOFs for each of the three clustering algorithms and their time series are shown respectively on Figures 6.12(a,b,c,d), 6.13(a,b,c,d), 6.14(a,b,c,d), 6.15(a,b,c,d), 6.16(a,b,c,d), 6.17(a,b,c,d), 6.18(a,b,c,d), 6.19(a,b,c,d), 6.20(a,b,c,d).

For the state of the extreme NAO atmospheric circulation pattern, for all three clustering algorithms, that are shown on the Figures 6.12(a,b,c,d), 6.15(a,b,c,d), 6.18(a,b,c,d), the North Atlantic ocean is partitioned into several regions that are the following:

1. Strong positive region localized in the are of South, West and East of Greenland, between 45 degrees N and 60 degrees N. This area matches with Cayan's findings, with the exception of the region contained in the North-West of Greenland, and the sign of it is negative.

2. Second strong positive region is a relatively wide region that extends along the Gulf Stream to the central North Atlantic. This area is also in sync with the Cayan map.

3. The third major area that also corresponds to Cayan findings, is the area between 60 degrees W and North Africa.

4. Another strong, but spatially smaller area is located above Great Britain, which also can be located Cayan [7], again with the opposite sign.

5. Our findings indicate another small area, north of Iceland and North-East of Greenland, between 65 degrees N and 70 degrees N. This area has strong negative sign, but is not located on Cayan's chart.

All three algorithms show notable consistency for the regions described above, with the small deviations in the magnitude of the correlation. The time series for all three algorithms also show good consistency throughout the years. The SST patterns that are related to the positive NAO are more representative than those that relate to the negative NAO. The time series of the negative, positive and the difference of average SST for the extreme EOF1, Figures 6.12(b,c,d), 6.15(b,c,d), 6.18(b,c,d), show opposite signs during the period between years of 1990 and 1995. The reason most probably can be referred to some degree of error in the calculations or in the data itself.

We assume that the regions of significant SST differences for NAO are also closely matched to those of heat flux, supporting the view that they are linked [7].

For the EA(EOF2) atmospheric circulation pattern, for all three clustering algorithms shown on the Figures 6.13(a,b,c,d), 6.16(a,b,c,d), 6.19(a,b,c,d), the North Atlantic ocean is partitioned into the following regions:

1. The area of strong positive differences in the central North Atlantic, between 45 degrees N and 55 degrees N. The area is consistent with the findings of Cayan, and again with the opposite sign for the feature.

2. Small area of the negative difference at the 60 degrees W, whereas the Cayan's map shows that the area is slightly shifted to the East.

The major feature for the EA pattern that differs our findings from the Cayan's ones is the area that occupies the pathway that runs from West of Greenland to the Labrador through the Labrador Sea.

3. There are also small areas of the SST differences with negative signs that are located east of Greenland, above the Iceland, that are not shown at Cayan's chart.

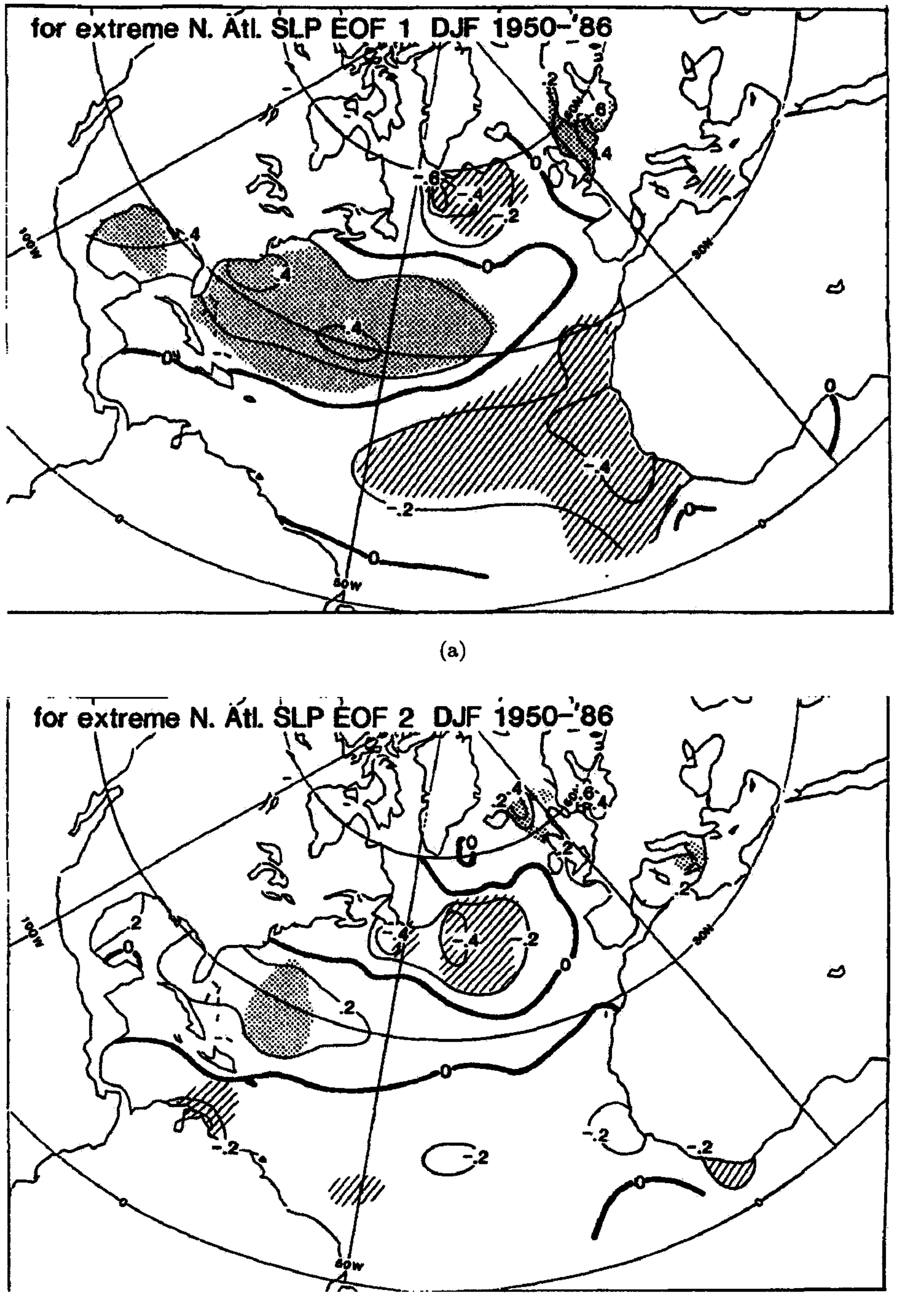
4. The small area with the negative sign fills the area in the North Atlantic at the 40 degrees N,

which is slightly shifted from the location on the Cayan's map.

As in the case with the extreme NAO circulation pattern, the time series of the negative, positive and the difference of average SST for the extreme EA(EOF2), Figures 6.13(b,c,d), 6.16(b,c,d), 6.19(b,c,d), show opposite signs during the period between the years of 1990 and 1995.

We also consider the third EOF, and run the above described procedure, as the third mode for the SLP accounts for 18 percent of variance in the data. The outcome for the EOF3 atmospheric circulation pattern, for all three clustering algorithms shown on the Figures 6.14(a,b,c,d), 6.17(a,b,c,d), 6.20(a,b,c,d).

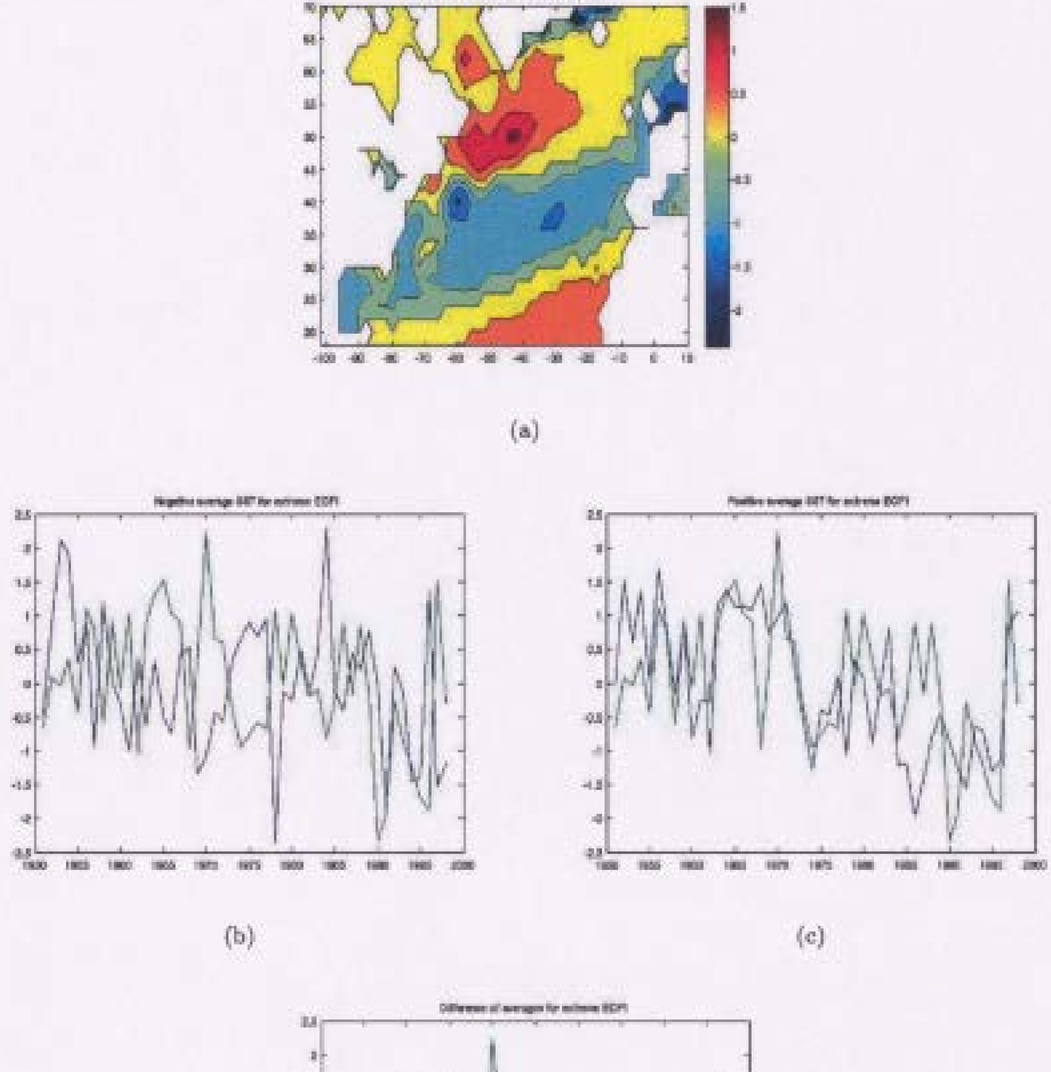
For the state of the extreme EOF3, we see one main region with negative sign that is located southeast of Newfoundland and occupied the area from 40 degrees W and 60 degrees W. The time series for the negative, positive and the difference of average SST for the extreme EOF3 are shown on Figures 6.14(b,c,d), 6.17(b,c,d), 6.20(b,c,d). The time series are less pronounced, yet they do not show opposite signs during the period between the years of 1990 and 1995, as in the case with EOF1 and EOF2.


for extreme N. Atl. SLP EOF 1 DJF 1950-'86
for extreme N. Atl. SLP EOF 2 DJF 1950-'86
100W
80W
60W
30W
0
.2
.4
.6
-.2
-.4
-.6

(a)

(b)

Figure 6.11: (*a*)-(*b*), Difference of average SST anomaly tendency associated with positive vs. negative extremes of EOF1 and EOF2 of North Atlantic SSP.(Adopted from D. Cayan[32])



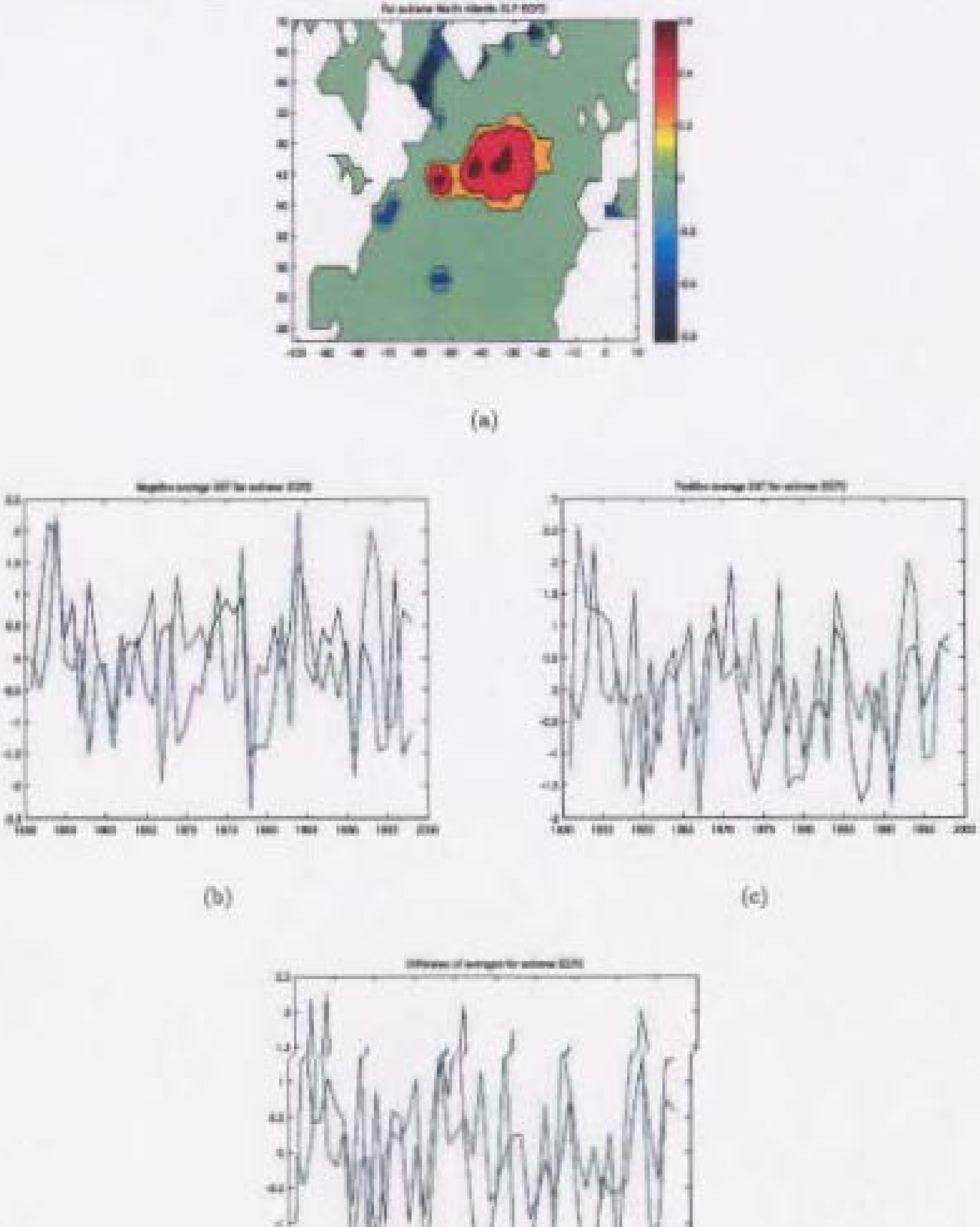
(a)

(b)

(c)

(d)

Figure 6.12: (*a*)-(*d*), Difference of average SST associated with positive and negative extremes of EOF1 of North Atlantic SLP and their related normalized normalized by one STD time-series(Axis Y for (b), (c), (d)) for the FartherstFirst algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



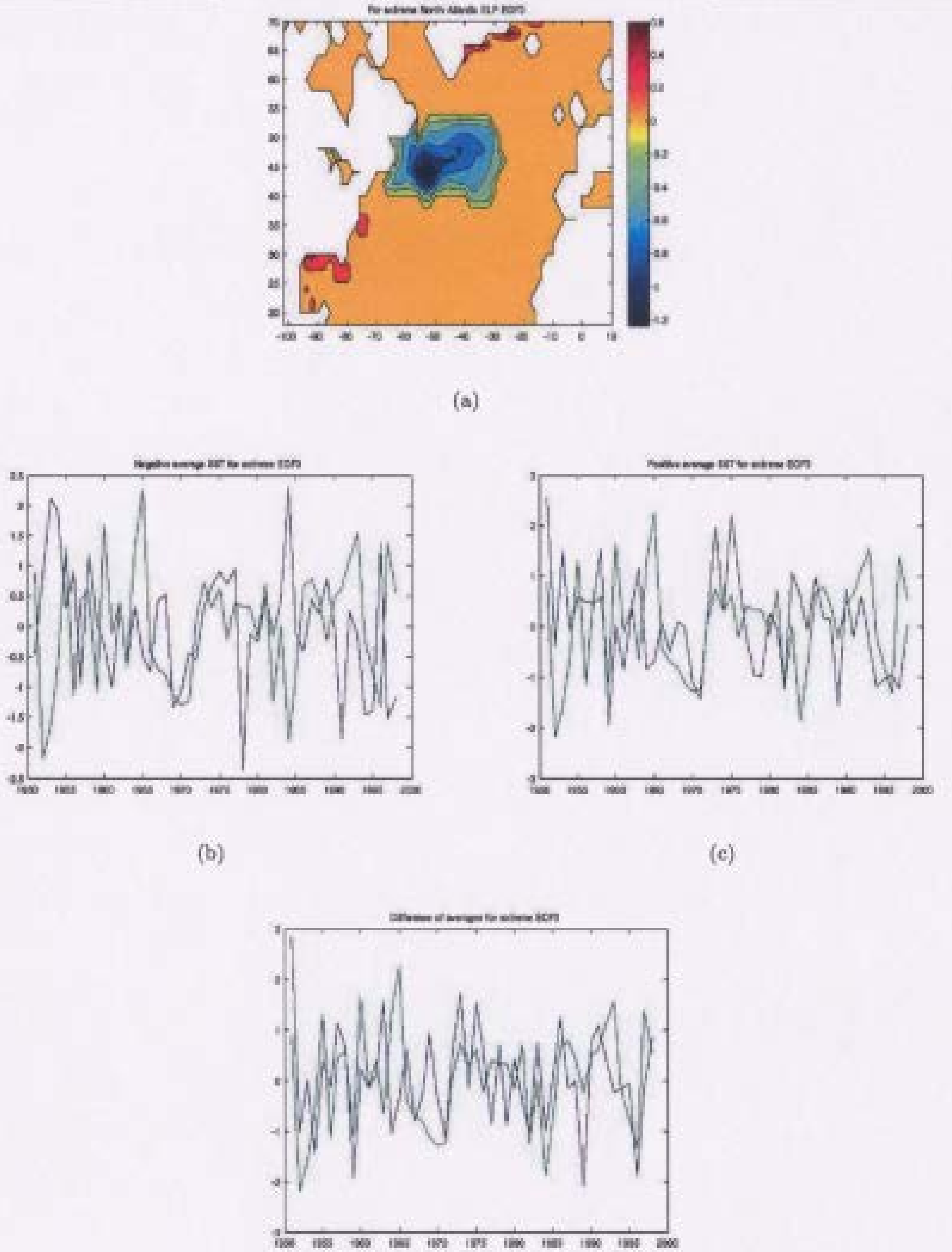
(a)

(b)

(c)

(d)

Figure 6.13: (*a*)-(*d*), Difference of average SST associated with positive and negative extremes of EOF2 of North Atlantic SLP and their related normalized normalized by one STD time-series(Axis Y for (b), (c), (d)) for the FartherstFirst algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



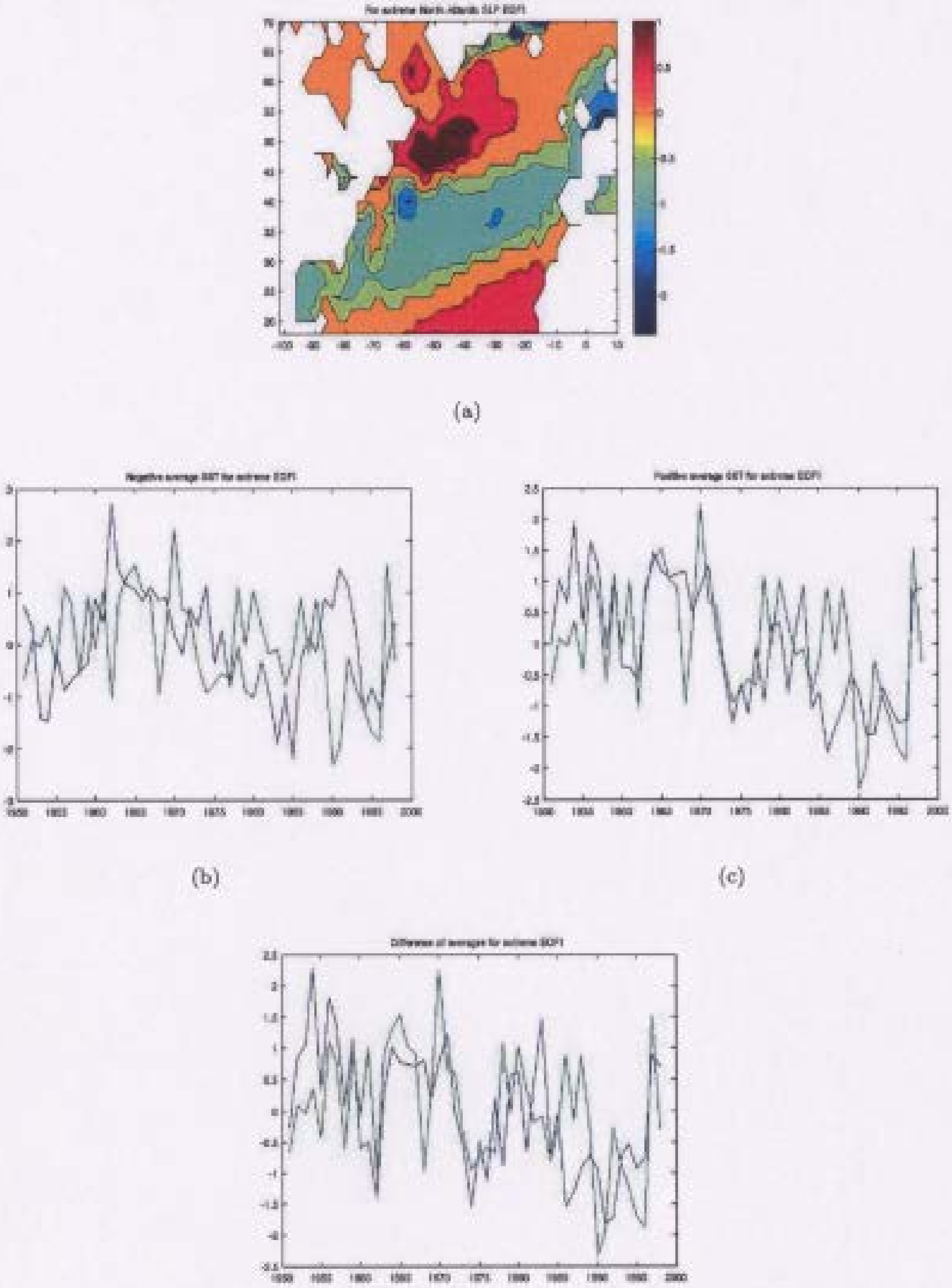
(a)

(b)

(c)

(d)

Figure 6.14: (*a*)-(*d*), Difference of average SST associated with positive and negative extremes of EOF3 of North Atlantic SLP and their related normalized by one STD time-series(Axis Y for (b), (c), (d)) for the FartherstFirst algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



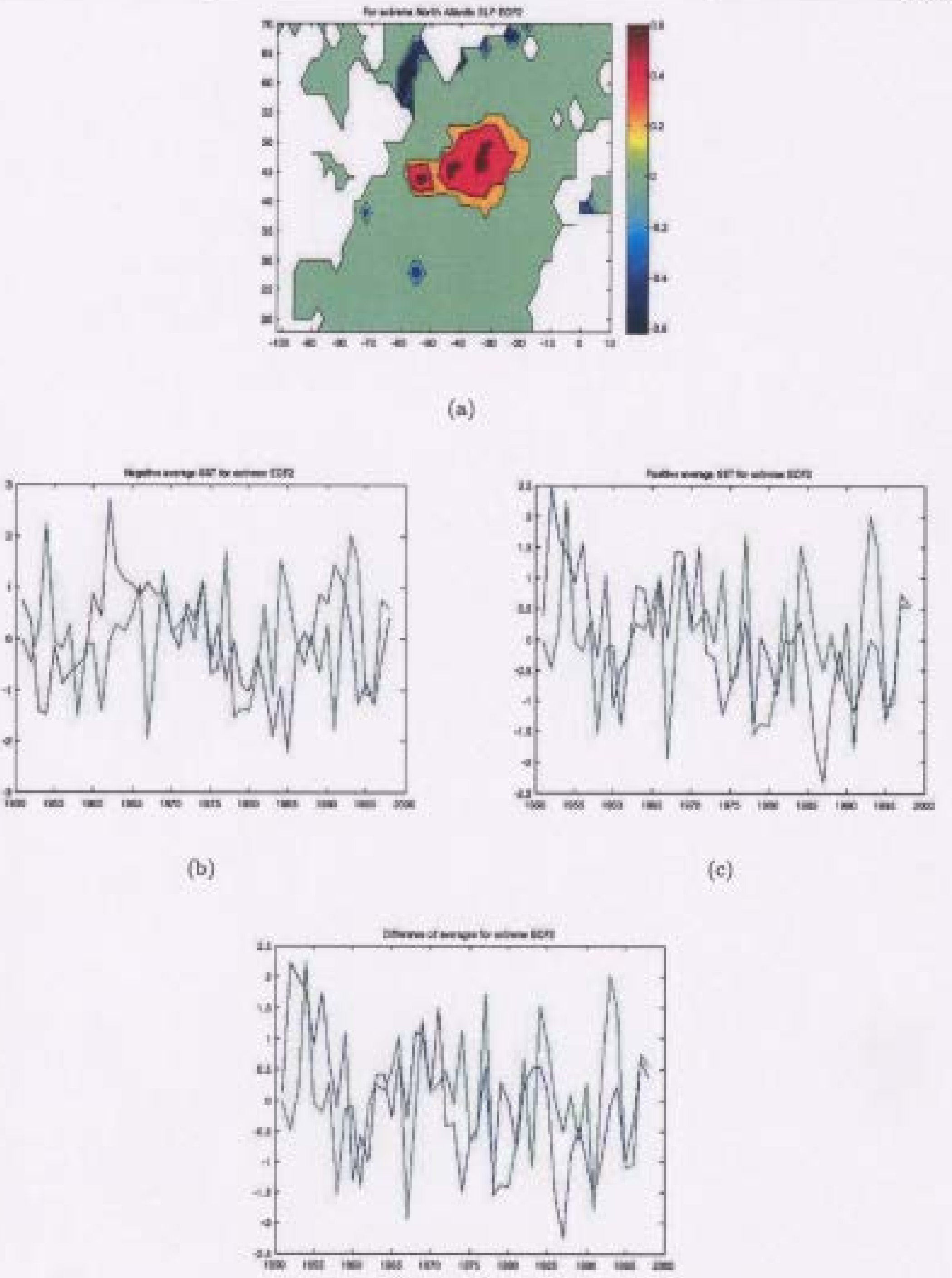
(a)

(b)

(c)

(d)

Figure 6.15: (*a*)-(*d*), Difference of average SST associated with positive and negative extremes of EOF1 of North Atlantic SLP and their related normalized by one STD time-series(Axis Y for (b), (c), (d)) for the EM algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



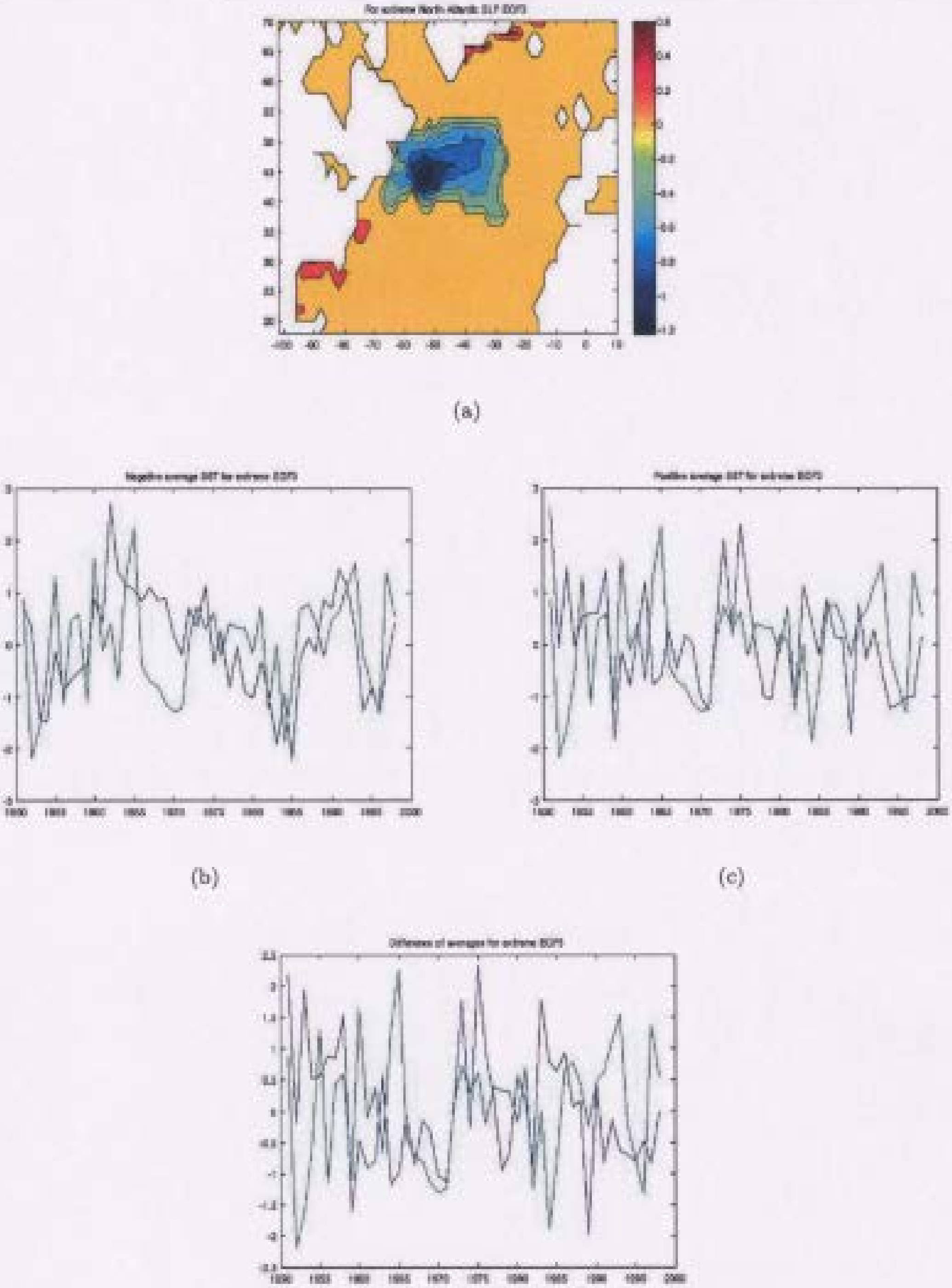
(a)

(b)

(c)

(d)

Figure 6.16: (*a*)-(*d*), Difference of average SST associated with positive and negative extremes of EOF2 of North Atlantic SLP and their related normalized by one STD time-series(Axis Y for (b), (c), (d)) for the EM algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



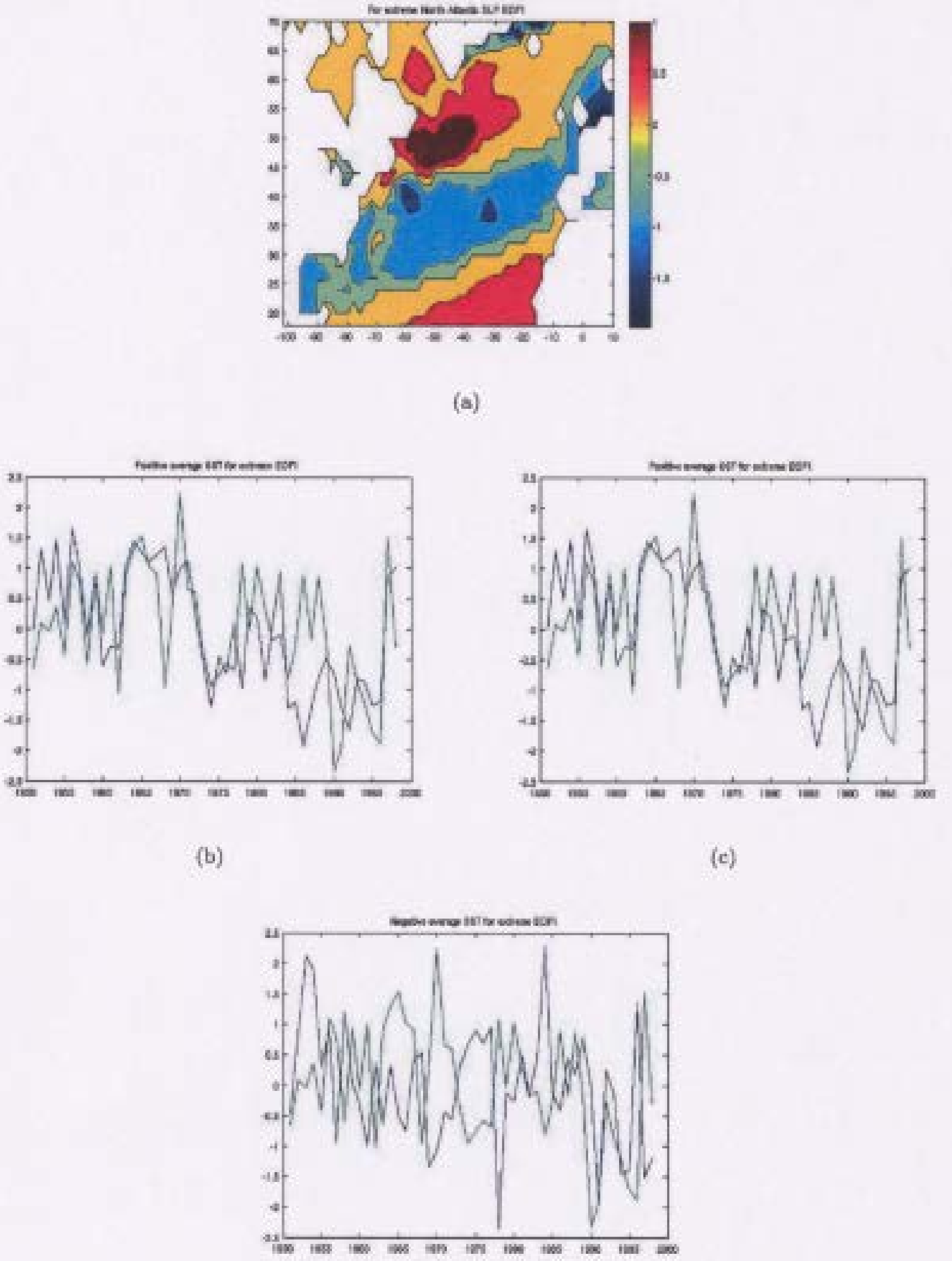
(a)

(b)

(c)

(d)

Figure 6.17: (*a*)-(*d*), Difference of average SST associated with positive and negative extremes of EOF3 of North Atlantic SLP and their related normalized by one STD time-series(Axis Y for (b), (c), (d)) for the EM algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



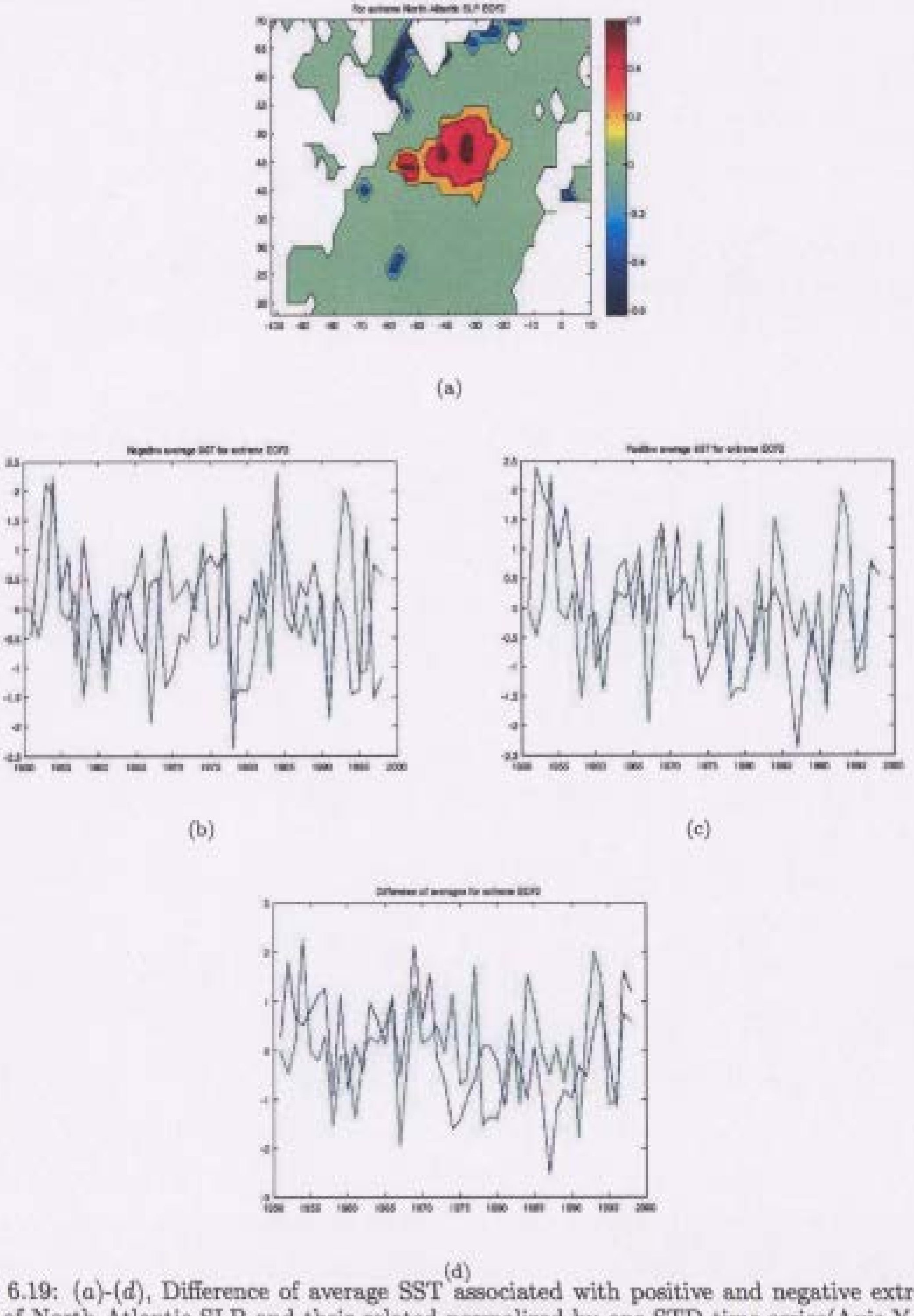
(a)

(b)

(c)

(d)

Figure 6.18: (*a*)-(*d*), Difference of average SST associated with positive and negative extremes of EOF1 of North Atlantic SLP and their related normalized by one STD time-series(Axis Y for (b), (c), (d)) for the KMeans algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



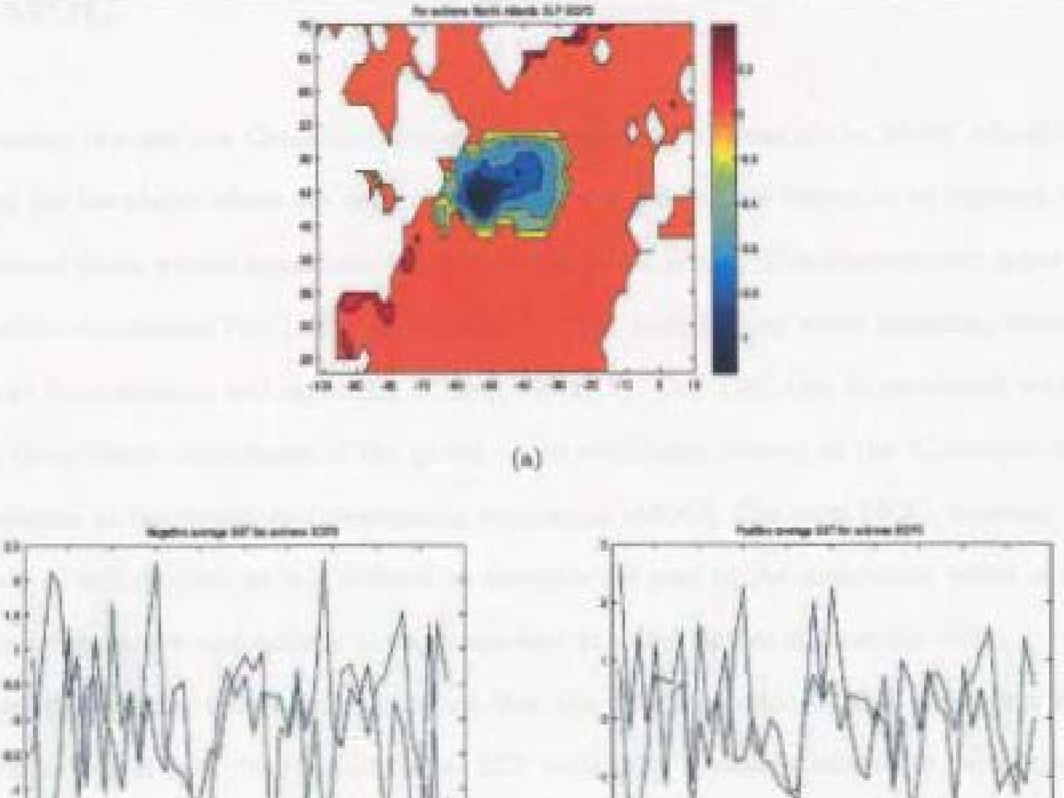
(a)

(b)

(c)

(d)

Figure 6.19: (a)-(d), Difference of average SST associated with positive and negative extremes of EOF2 of North Atlantic SLP and their related normalized by one STD time-series(Axis Y for (b), (c), (d)) for the KMeans algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.



(a)

(b)

(c)

(d)

Figure 6.20: (a)-(d), Difference of average SST associated with positive and negative extremes of EOF3 of North Atlantic SLP and their related normalized by one STD time-series(Axis Y for (b), (c), (d)) for the KMeans algorithm. Axis X for (b), (c), (d) represents time in years. Axis X for (a) represents longitude, Axis Y for (a) represents latitude.

6.4 MOC

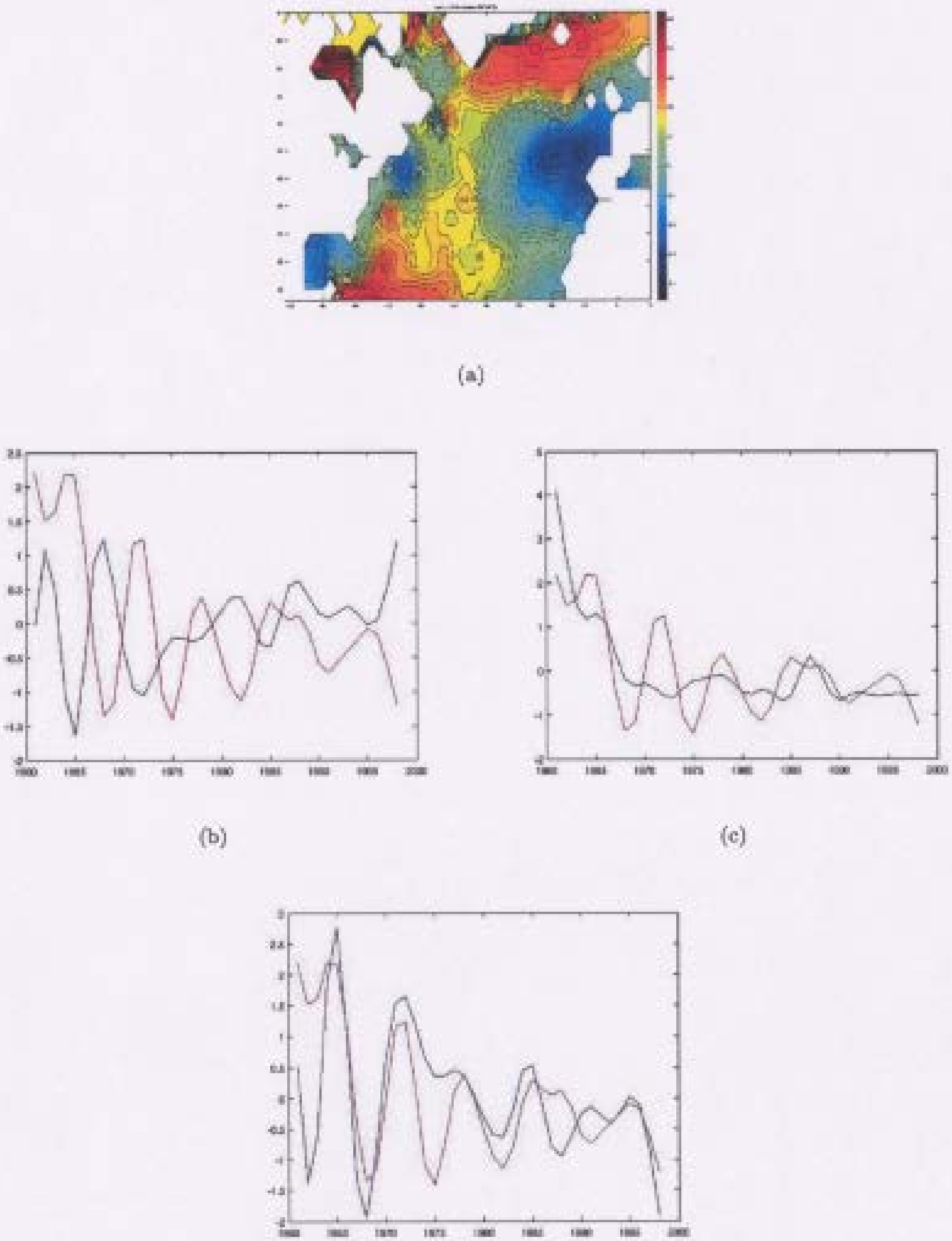
The Labrador Sea and the Greenland/Iceland/Norwegian (GIN) Seas of the North Atlantic Ocean are two of the few places where the deep waters of the world ocean are known to be renewed, and the newly formed dense waters spread into the rest of the global ocean. This phenomenon is part of the thermohaline circulation(THC). The key features of THC include deep water fomation, spreading of deep water from sources, and upwelling of deep water[17]. The THC that is associated with North Atlantic Deep Water contributes to the global ocean circulation termed as the "Conveyor Belt". It is also referred as the meridional overturning circulation (MOC). The term MOC, however, is more accurate and well defined, as it is difficult to separate the part of the circulation which is actually driven by temperature and salinity alone as opposed to other factors such as the wind.

There are observations that support the fact that the North Atlantic THC is consistent with the SST variability and that "the multidecadal SST variability is closely related to variations in the North Atlantic thermohaline circulation[41], and that the variations in the North Atlantic THC are reflected in large-scale SST anomalies[41],[55],[15]. Changes of THC are important for climate, but there are no good methods to observe THC, therefore we try to identify SST index that would provide information about MOC.

The MOC index is computed from the ECMWF coupled ocean-atmosphere-sea-ice reanalysis(data source KNMI- Royal Netherlands Meteorological Institute). Period of Study 1960 - 1998. The MOC index is defined as the maximum of the overturning streamfunction calculated from ECMWF ocean reanalysis. The correlation map and the time series of the averaged SST for all three clusters and MOC are shown on Figures 6.21(a,b,c,d), 6.22(a,b,c,d), 6.23(a,b,c,d).

The correlation maps for all three algorithms Figures 6.21(a), 6.22(a), 6.23(a) show well pronounced dipole structure with the negative correlation around Azores and positive correlation around Iceland. There are also relatively small areas with positive and negative correlation, south of Nova Scotia, in Hudson Bay, and in the Gulf of Mexico.

The time series for all three algorithms, Figures 6.21(b,c,d), 6.22(b,c,d), 6.23(b,c,d), also show good relationships between SST and the MOC. The opposite sign of correlation between the years approximately between 1987 and 1997, confirms the discrepancy in the SST data or possibly in the calculations of the clusters for those years. Close connection between SST clusters and the THC exists during the whole time period from 1957 to 1980. For the period from 1980 to 1987 we observe that both curves show the same tendency, although the magnitude varies, especially for the period between 1973 and 1976. A good correlation is shown starting from 1993 and up. The beginning of the study time period, from about 1961 to 1963 shows a similar tendency, but with a greater difference than in other years. We assume that there was not enough data or not enough high quality data for that period. More data SST should be investigated using the similar techniques to see if the results will be more accurate. Yet, we can assume that our results are comparable with the work of Latif et al. [41], where the simulations in coupled ocean-atmosphere models show that variations in the North Atlantic THC are reflected in large-scale SST anomalies. More data should be investigated to study the option of using data mining principles and cluster analysis in monitoring future changes in THC strength and its predictability using SST clusters.



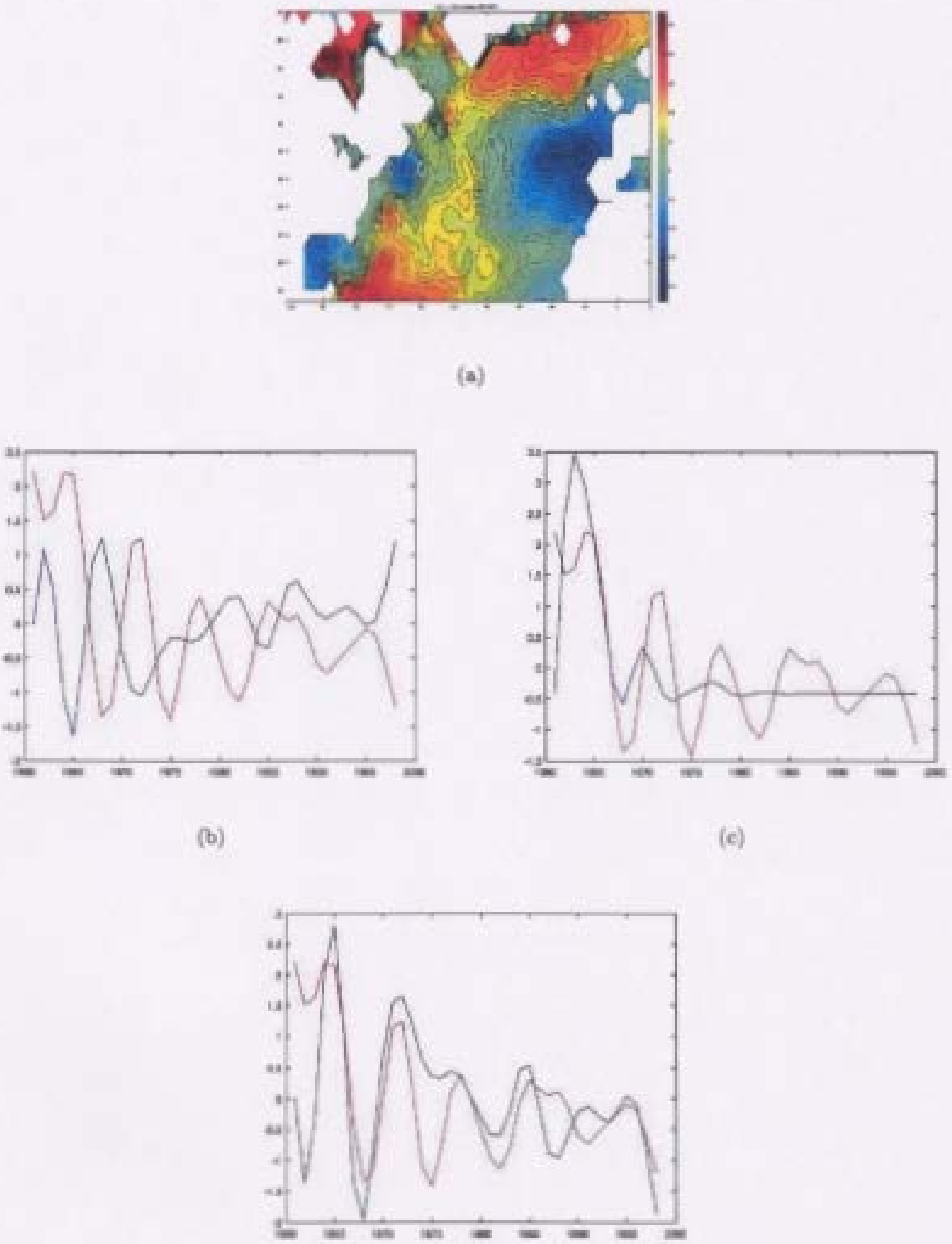
(a)

(b)

(c)

(d)

Figure 6.21: (*a*)-(*d*), Difference of average SST associated with positive and negative MOC and their related normalized time-series for the FF algorithm. Time series of MOC index(red line) and maximum of positive SST(blue line) are normalized by one STD



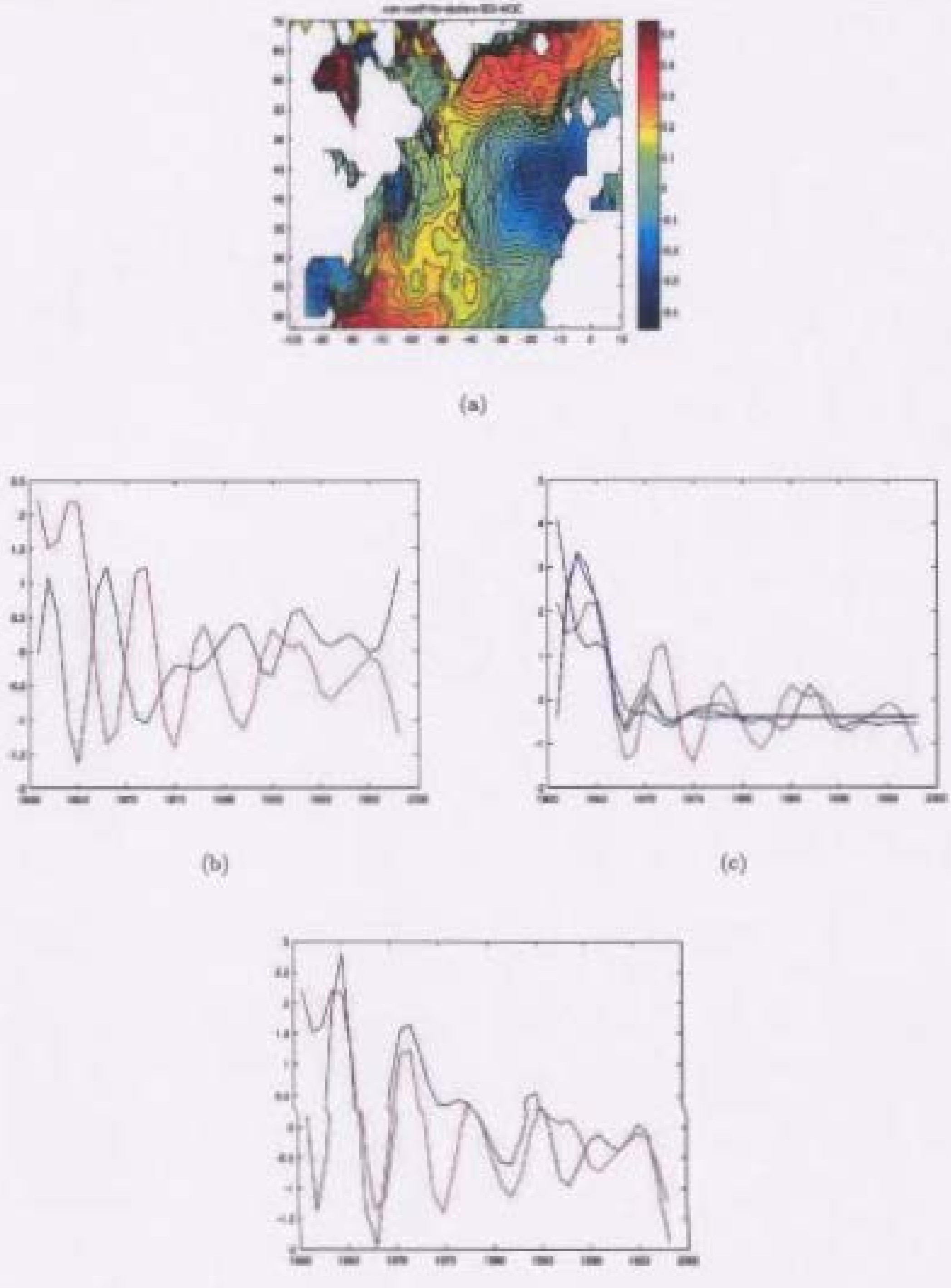
(a)

(b)

(c)

(d)

Figure 6.22: (*a*)-(*d*), Difference of average SST associated with positive and negative MOC and their related normalized time-series for the EM algorithm. Time series of MOC index(red line) and maximum of positive SST(blue line) are normalized by one STD



(a)

(b)

(c)

(d)

Figure 6.23: (*a*)-(*d*), Difference of average SST associated with positive and negative MOC and their related normalized time-series for the FF algorithm. Time series of MOC index(red line) and maximum of positive SST(blue line) are normalized by one STD

Chapter 7

Conclusion

Through application of data mining and clustering techniques to the oceanographic Sea Surface Temperature data we obtained valuable information about how these clusters map on the natural structure of the problem. We investigate the relationships between the NAO and SSTs over the North Atlantic basin and detect a significant causal relation of SST with the dominant atmospheric circulation pattern. This influence is mainly centered over the Gulf Stream, the Greenland and the subtropics areas, which are the centers of the typical tripole pattern, which may represent the effect of positive feedbacks between the atmosphere and ocean in this region. Testing three different data mining algorithms in this context allows us to better outline possible sources of uncertainty for the NAO and EA indices. As was mentioned in the previous chapter, between two main data mining classification techniques such as supervised and unsupervised, we have chosen unsupervised classification or clustering, since we can not predetermine the set of classes in advance. Thus, since the clustering is done in a completely unsupervised manner, finding that the cluster structure is reasonably mapped onto the true classes supports the hypothesis that algorithms described in this study, such as FarthestFirst, Expectation-Maximization (EM), and KMeans are capable of discovering the "true structure", the one that is inherent in the data. However, it is observed that perfect classification is not achieved and it is also observed that different methods of cluster analysis are effective at detecting different kinds of clusters, so, different clustering algorithms are biased toward finding different types of cluster structures in the data. Therefore, the approach in this study was to try to match the method to the objectives, and "apply a cluster analytical tool that is effective in detecting clusters for the problem that we want to solve" [26]. Quantifying the clustering results is difficult and the validity of clustering is often subjective, as it depends to some degree on the eye of a researcher, and if the clusters produce interesting scientific insight, it can be judged as being useful[57]. Following this strategy, we did not obtain a direct assessment of the goodness of clusters per se; in exchange, we obtained valuable information about how these clusters map on the natural structure of the problem, something that may be more interesting than evaluating a single or few indirect performance parameters. To evaluate the quality of clustering and overall results, we adopted the approach of comparing the results to a "ground truth". The results show that the clustering method applied to SST compares favorably with the approach described in previous studies of dominant patterns of interannual variability in the North Atlantic and the results achieved are comparable with those obtained by the authors [7], [20], [53], [50], [18]. This comparison should be used as a possible conformation of the validity of the method that besides Sea Level Pressure, Sea Surface Temperature can be used as another parameter linked to the North Atlantic Oscillation.

7.1 Future Work and Recommendations

The classification step is the most computationally intensive step in the process, requiring up to 24 hours to classify large amount of instances. Since in the long run we want to apply the process of unsupervised classification or clustering to the "ARGO-DATA" database described in Section 3.2 that contains millions of instances, it is vital to examine new methods for reducing the computational time and to scale the entire classification process to accommodate massive amount of data in a timely manner.

There are two approaches that can be investigated and compared with each other by the outcomes.

- ⋆ The first approach of scaling the process is to parallelize the learning algorithm by splitting it up into parallel portions and executing the splitted portions of the code on multiple processors to get the results. Parallel algorithm will perform the job faster than the serial(sequential) one, but it may be possible that the entire sequential algorithm or part of it will be inherently serial, i. e. the algorithm can not be split up into parallel portions. Therefore, the second approach can be used.

- ⋆ Second approach partitions the data by itself into subsets, allocates those subsets to different processors and applies a sequential algorithm to each data subset. Second approach does not alter the algorithm, and allows the classification results be combined from each processor to one single classifier. This approach is received an attention mainly because of two reasons: it reduces execution time and improves classification accuracy[37]. The execution time is becoming shorter because the expensive classification step is allocated among several processors. The accuracy of classification is improved because each of the distributed classifiers makes different types of errors and the resulting classification is often more accurate than that of a single classifier[47].

Most of the previous work on combining classifiers is done using supervised classification algorithms[46, 47, 37]. Since we cannot predetermine the set of classes in advance, we are using unsupervised classification or clustering. Therefore, the schema of the future work should include, first, determining the likeliness or similarity between classes that are outcomes from different classifiers, and then determining how to combine the results when classifying a particular instance. The process will involve writing the programs that will distribute the data and then gather the results.

Both, first and second approaches can be implemented and then evaluated using parameters such as execution time, speedup, and efficiency.

Furthermore, we would suggest that:

- ⋆ More research work should be done that would test existing algorithms and build new algorithms that are applicable to time series data, as the "standard" algorithms that comes with data mining software such as WEKA may not always be applicable.

- ⋆ The results of the clustering algorithms presented in this study should also be compared to results of other methods both supervised and unsupervised. The comparison will yield the methods that are much fitted to the data with similar temporal and spatial characteristics.

⋆ More research needs to be done to apply clustering methods to Ice Concentration data. Most of the work is completed successfully, and the clustering was implemented on the Ice Concentration data using Farthest-First algorithm.

Appendices

I Mathematical Relations

To understand the true meaning of the term *relation*, it is useful to review some concepts from mathematics. Suppose that we have two sets, D_1 and D_2, where $D_1 = \{2, 4\}$ and $D_2 = \{1, 3, 5\}$. The **Cartesian product** of these two sets, written $D_1 \times D_2$, is a set of all ordered pairs such that the first element is a member of D_1 and the second element is a member of D_2. An alternative way of expressing this is to find all combinations of elements with the first from D_1 and the second from D_2. Thus, we will have:

$D_1 \times D_2 = \{(2,1),(2,3),(2,5),(4,1),(4,3),(4,5)\}$

Any subset of this Cartesian product is a relation. For example, we could produce a relation R such that:

$R = \{(2,1),(4,1)\}$

We may specify which ordered pairs will be in the relation by giving some condition for their selection. For example, if we observe that R includes all those ordered pairs in which the second element is 1, then we could write R as: $R = \{(x, y)|x \in D_1, y \in D_2, \text{ and } y = 1\}$

Using the same sets, we could form another relation S as:

$S = \{(x, y)|x \in D_1, y \in D_2, \text{ and } x = 2y\}$

or, in this instance,

$S = \{(2,1)\}$

since there is only one ordered pair in the Cartesian product that satisfies this condition. We can extend the relation to three sets. Let D_1, D_2, and D_3 be three sets. The Cartesian product $D_1 \times D_2 \times D_3$ of these three sets is the set of all ordered triples such that the first element is from D_1, the second element is from D_2, and the third element is from D_3. For example, suppose we have: $D_1 = \{1, 3\}$, $D_2 = \{2, 4\}$, $D_3 = \{5, 6\}$

$D_1 \times D_2 \times D_3 = \{(1,2,5),(1,2,6),(1,4,5),(1,4,6),(3,2,5),(3,2,6),(3,4,5),(3,4,6)\}$

Any subset of these ordered triples is a relation. We can extend the three sets and define a general relation on n domains. Let D_1, D_2, D_3,…,D_n be n sets. Their Cartesian product is defined as:

$D_1 \times D_2 \times D_3 \times \ldots \times D_n = \{(d_1 \in D_1, d_2 \in D_2, \ldots, d_n \in D_n\}$

Any set of n-tuples from this Cartesian product is a relation on the n sets.

II Metric spaces

A metric space (χ, ρ) consists of a set χ and a distance function ρ: $\chi \times \chi \rightarrow \mathbb{R}$ that satisfies the three properties of a metric:

(1) Reflexivity: $\rho(x, y) \geq 0$ with equality iff $x = y$

(2) Symmetry: $\rho(x, y) = \rho(y, x)$

(3) Triangle inequality: $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$

III Distance Measures

Many data mining techniques such as *nearest neighbor* classification methods, *cluster analysis*, and *multidimensional scaling methods*, are based on similarity measures. There are two ways to obtain measures of similarity. One way is they can be obtained from the objects themselves, and second way is when it is necessary to give a precise definition of "similar". The formal definition of "similar" allows one to define "dissimilar" by applying a "suitable monotonically decreasing transformation" [27]. Therefore, if *s(i, j)* denotes the similarity and *d(i, j)* denotes the dissimilarity between objects i and j, then possible transformations include $d(i, j) = 1 - s(i, j)$ and $d(i, j) = \sqrt{2(1 - s(i, j))}$.

The definitions such as *distance* and *metric* are used to denote a measure of dissimilarity. *Distance* refers to a dissimilarity measure taken from the characteristics describing the objects – as in Euclidean distance. The Euclidean distance between the ith and jth objects is defined as

$$d_E(i, j) = \sqrt{\left(\sum_{k=1}^{p}(x_k(i) - x_k(j))^2\right)} \tag{1}$$

where n is the number of data objects with p real-valued measurements on each object; and $\boldsymbol{x}(i)$ is the vector of observations for the ith object: $\boldsymbol{x}(i) = (x_1(i), x_2(i), ..., x_p(i))$, $1 \geq i \leq n$, where the value of the kth variable for the ith object is $x_k(i)$.

Metric is a dissimilarity measure that satisfies three conditions:

1. $d(i,j) \geq 0$ for all i and j, and $d(i,j) = 0$ if and only if $i = j$;

2. $d(i,j) = d(j,i)$ for all i and j; and

3. $d(i,j) \leq d(i,k) + d(k,j)$ for all i, j and k.

A common strategy to standardize the data if the variables are not compatible is to divide each of the variables by its sample standard deviation, so that they are all regarded as equally important. The standard deviation of the kth variable X_k can be estimated as

$$\hat{\sigma} = \sqrt{\left(\frac{1}{n}\sum_{i=1}(x_k(i) - \mu_k)^2\right)} \tag{2}$$

where μ_k is the mean for variable X_k, which can be estimated using the *sample mean* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_k(i)$. In addition, knowing the relative importance of each variable, we can weight them to have the weighted Euclidean distance measure, such as

$$d_{WE}(i,j) = \sqrt{\left(\sum_{k=1}^{p} w_k(x_k(i) - x_k(j))^2\right)} \tag{3}$$

This property may not be always appropriate in the case when variables are highly correlated and one of the approaches of standardizing the data in this case is not just in the direction of each variable, as with weighted Euclidean distance, but also taking into account the *covariances* between the variables. If we assume that we have two variables X and Y, and also assume that we have n objects, with X taking values $x(1), ..., x(n)$ and Y taking values $y(1), ..., y(n)$, then the sample covariance between X and Y will be defined as

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x(i) - \bar{x})(y(i) - \bar{y}) \tag{4}$$

where $\bar{x}$ is the sample mean of the X values and $\bar{y}$ is the sample mean for the Y values. The covariance $Cov(X,Y)$ is a measure of how X and Y vary together: $Cov(X,Y)$ will have a large positive value if large values of X tend to be associated with large values of Y and small values of X with small values of Y. If large values of X are associated with small values of Y, $Cov(X,Y)$ will take a negative value. In general, with p variables we can construct $p \times p$ matrix of covariances, in which the element (k,l) is the covariance between the kth and lth variables, and from the definition of covariance above, such covariance matrix must be symmetric. The value of the covariance depends of the ranges of X and Y. This dependence can be removed by standardizing, dividing the values of X by their standard deviation and the values of Y by their standard deviation. The result is the *sample correlation coefficient* $\rho(X,Y)$ between X and Y:

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\sqrt{(\sum_{i=1}^{n}(x(i)-\bar{x})^2 \sum_{i=1}^{n}(y(i)-\bar{y})^2)}} \tag{5}$$

The same way as a covariance matrix can be formed if there are p variables, in the same manner $p \times p$ correlation matrix can be formed. Covariance and correlation capture *linear dependencies* between variables, therefore the more accurate terms are *linear covariance* and *linear correlation.* Thus independence implies a lack of correlation, but the reverse is not generally true, as it may show non-linear correlation.

There is a number of other metrics for quantitative measurements, so the problem is not so much defining one but rather deciding which is the most appropriate for a particular situation[26].

IV K-Means

As an algorithm, the K-Means method is as follows: assuming we have n data points $D = \{x_1, ..., x_n\}$, our task is to find K clusters $\{C_1, ..., C_K\}$:

```
for $k = 1, ..., K$ let $r(k)$ be a randomly chosen point from $D$;
while changes in clusters $C_k$ happen do
    form clusters:
    for $k = 1, ..., K$ do
        $C_k = \{x \in D | d(r_k, x) \leq d(r_j, x)$ for all $j = 1, ..., K, j \neq k\}$;
    end;
    compute new cluster centers:
    for $k = 1, ..., K$ do
        $r_k$ = the vector mean of the points in $C_k$
    end;
  end;
```

V Expectation Maximization(EM)

EM clustering method is designed to solve the the missing or hidden data problems in a likelihood context. Particularly, let $D = x(1), ..., x(n)$ be a set of n observed data vectors. Let $H = z(1), ..., z(n)$ represent a set of n values of a hidden variable Z, which is in one-to-one correspondence with the observed data points D; that is, $z(i)$ is associated with data point $x(i)$. We can assume Z to be discrete, in which case we can think of the unknown $z(i)$ values as cluster labels that are hidden. We can write the log-likelihood of the observed data as:

$$l(\theta) = \log p(D|\theta) = \log \sum_{H} (p(D, H|\theta) \tag{6}$$

where the term on the right indicates that the observed likelihood can be expressed as the likelihood of both the observed and hidden data, summed over the hidden data values, assuming a probabilistic model in the form $p(D, H|\theta)$ that is parametrized by a set of unknown parameters θ. Let $Q(H)$ be any probability distribution on the missing data H. Then the log-likelihood can be written as:

$$l(\theta) = \log p(D|\theta) = \log \sum_{H} (p(D, H|\theta) \tag{7}$$

Let $Q(n)$ be any probability distribution on the missing data H, We can then write the log-likelihood in the following fashion:

$$l(\theta) = \log p(D|\theta)$$

$$= \log \sum_{H} Q(H) \frac{p(D, H|\theta)}{Q(H)}$$

$$\geq \sum_H Q(H) \log \frac{p(D,H|\theta)}{Q(H)}$$

$$= \sum_H Q(H) \log p(D,H|\theta) + \sum_H Q(H) \log \frac{1}{Q(H)}$$

$$= F(Q,\theta)$$

The function $F(Q,\theta)$ is a lower bound on the function we wish to maximize(the likelihood $l(\theta)$). The EM algorithm alternates between maximizing F with respect to the distribution Q with the parameters θ fixed, and then maximizing F with respect to the parameters θ with the distribution $Q = p(H)$ fixed. Specifically:

E-step: $Q^k + 1 = \arg\max_Q F(Q^k, \theta^k)$

M-step: $\theta^k + 1 = \arg\max_\theta F(Q^k + 1, \theta^k)$

E and M steps have a simple interpretation. In the E-step we estimate the distribution on the hidden variables Q, conditioned on a particular setting of the parameter vector θ^k. Then, keeping the Q function fixed, in the M-step we choose a new set of parameters $\theta^k + 1$ so as to maximize the expected log-likelihood of observed data (with expectation defined with respect to $Q = p(H)$. In turn, we can now find a new Q distribution given the new parameters $\theta^k + 1$, then another application of the M-step to get θ^{k+2}, and so forth in an iterative manner. As sketched above, each such application of the E and M steps is guaranteed not to decrease the log-likelihood of the observed data, and under fairly general conditions this in turn implies that the parameters θ will converge to at least a local maximum of the log-likelihood function.

To specify an actual algorithm we need to pick an initial starting point(for example, start with either an initial randomly chosen Q or θ) and a convergence detection method(for example, detect when any of Q, θ, or $l(\theta)$ do not change from one iteration to the next). The EM algorithm is similar to a form of local hill-climbing in multivariate parameter space(as discussed in earlier sections of this chapter) where the direction and distance of each step is implicitly and automatically specified by the E and M steps. The method is sensitive to initial conditions, so that different choices of initial conditions can lead to different local maxima. In practice it is usually wise to run EM from different initial conditions to decrease the probability of finally settling on a relatively poor local maximum.

The EM algorithm is widely used given the broad generality of the framework and the relative ease with which an EM algorithm can be specified for many different problems.

The computational complexity of the EM algorithm is dictated by the number of iterations required for convergence and the complexity of each of the E and M steps. The algorithm can often converge to the general vicinity of the solution after only a few (10-15) iterations. The complexity of the E and M steps depends on the data and for many simple models the E and M steps need only take time linear in n, i.e., each data point need only be visited once during each iteration.

VI FarthestFirst

Hochbaum and Shmoys (1985) introduced the Farthest-First traversal of a data set as an approximation algorithm for what is sometimes called

the k-center problem, that of finding an optimal *k-clustering* under the cost function. The cost of clustering is taken to be the largest radius of its clusters.

The Farthest-First starts by assigning each data instance to its own cluster. It finds the Euclidean Distance between all instances in each pair of clusters. The maximum of these distances is chosen. Any two clusters that have minimum of this chosen distance are merge. The process is continued until the total number of clusters is above some specified threshold[59]. The algorithm of *Farthest-First traversal* is described as follows:

```
pick any $z \in S$ and set $T = \{z\}$
while $|T| < k$
$z = argmax_{x \in S} \rho(x, T)$
$T = T \cup \{z\}$
```

This builds a solution T one point at a time. It starts with any point, and then iteratively adds the point farthest from the ones chosen so far. According to Hochbaum and Shmoys[29], the solution of the farthest-first traversal may not be perfect, but it is close to optimal, that is for any k if T is the solution returned by farthest-first traversal, and T^* is the optimal solution, then

$$cost(T) \leq 2cost(T^*)$$

VII Nearest-Neighbor Clustering

In the *nearest-neighbor* or *single link* clustering method each new instance is compared with existing ones using a distance metric, and the closest existing instance is used to assign the class to the new one [27, 4]. Usually the starting point for the process is the initial clustering in which each cluster consists of a single data point, so that the procedure begins with the n points to be clustered. Nearest-neighbor defines the distance between two clusters as the distance between the two closest points, one from each cluster:

$$D_{sl}(C_i, C_j) = \min_{x,y}\{d(x,y)|x \in C_i, y \in C_j\}, \tag{8}$$

where $d(x, y)$ is the distance between objects x and y. As an algorithm, the method is as follows: assuming we are given n data points $D = \{x(1), ..., x(n)\}$, and a function $D(C_i, C_j)$ for measuring the distance between two clusters C_i and C_j. Then the algorithm for clustering can be described as follows:

```
for $i = 1, ..., n$ let $C_i = \{x(i)\}$;
while there is more than one cluster left do
  let $C_i$ and $C_j$ be the clusters
    minimizing the distance $D(C_k, C_h)$ between any two clusters;
  $C_i = C_i \cup C_j$;
  remove cluster $C_j$:
end;
```

One of the major disadvantages of this method is that it is often slow[57] and processing whole data set takes time that is proportional to the number of rows in the data table, that is, it can be computationally expensive, especially when the data set is large.

VIII Log-Likelihood

The most common score function for estimating the parameters of probability functions is the log-likelihood. If the probability function of random variables X is $f(x; \theta)$, where θ are the parameters that need to be estimated, then the log-likelihood is $logf(D|\theta)$, where $D = \{x(1), ..., x(n)\}$ is the observed data. Making the common assumption that the separate rows of the data matrix have arisen independently, this becomes

$$S_L(\theta) = -\sum_{i=1}^{n} \log f(x(i); \theta) \tag{9}$$

If f has simple functional form then this score function can usually be minimized explicitly, producing a closed form estimator for the parameters θ. However, if f is more complex, iterative optimization methods may be required. Despite the importance, the likelihood may not always be an adequate or appropriated measure for comparing models, in particular when the models are of different complexity.

IX Principal Component Analysis(PCA)

Suppose that X is an $n \times p$ data matrix in which the rows represent the cases(each row is a data vector $x(i)$) and the columns represent the variables. The ith row of this matrix is actually the transpose x^T of the ith data vector $x(i)$, since the convection is to consider data vectors as being $p \times 1$ column vectors rather than $1 \times p$ row vectors. Also, we assume that X is mean-centered so that the value of each variable is relative to the sample mean for that variable(that is estimated mean has been subtracted from each column). Let a be the $p \times 1$ column vector of projection weights(unknown at this point) that result in the largest variance when the data X are projected along a. The projection of any particular data vector x is the linear combination $a^T \times x = \sum_{j=1}^{p} a_j x_j$. We can express the projected values onto a of all data vectors in X as $Xa(n \times p$ by $p \times 1$, yielding an $n \times 1$ column vector of projected values). Furthermore, we can define the variance along a as

$$\sigma_a^2 = (Xa)^T(Xa) = a^T X^T X a = a^T V a, \tag{10}$$

where $V = X^T X$ is the $p \times p$ covariance matrix of the data. Thus, we can express σ_a^2(the variance of the projected data (a scalar) that we wish to maximize) as a function of both a and the covariance matrix of the data V. Maximizing σ_a^2 directly is not well-defined, since we can increase σ_a^2 without limit simply be increasing the size of the components of a. We impose normalization constraint on the a vectors such that $a^T a = 1$. With this normalization constraint we can rewrite our optimization problem as that of maximizing the quantity

$$[u = a^T V a - \lambda(a^T a - 1), \tag{11}$$

where λ is a Lagrange multiplier. Differentiating with respect to a yields

$$\frac{\partial u}{\partial a} = 2Va - 2\lambda a = 0, \tag{12}$$

which reduces to the eigenvalue form of

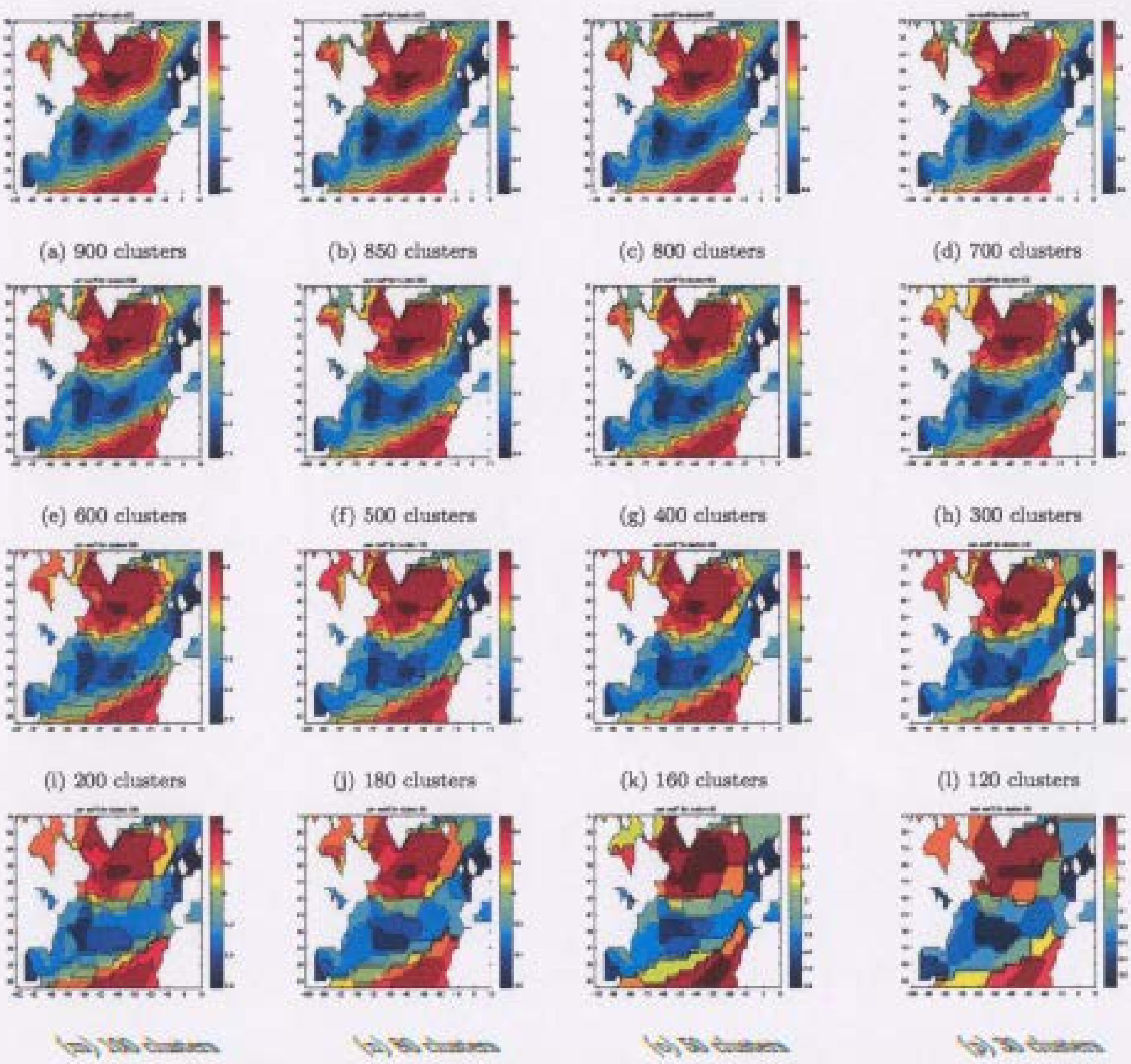
$$(V - \lambda I)a = 0 \tag{13}$$

Therefore, the first principal component a is the eigenvector associated with the largest eigenvalue of the covariance matrix V. Furthermore, the second principal component (the direction orthogonal to the first component that has the largest projected variance) is the eigenvector of the kth largest eigenvalue corresponds to the kth principal component direction). In practice we are interested in projecting to more than two-dimensions. The variance of the projected data can be expressed as $\sum_{j=1}^{k} \lambda_j$, where λ_j is the jth eigenvalue. Similarly, the squared error in terms of approximating the true data matrix X using only the first eigenvectors can be expressed as

$$\frac{\sum_{j=k+1}^{p} \lambda_j}{\sum_{l=1}^{p} \lambda_l} \tag{14}$$

Therefore, in choosing an appropriate number k of principal components, one approach is to increase k until the squared error quantity above is smaller than some acceptable degree of squared error. For high-dimensional data sets, in which the variables are often relatively well-correlated, it is not uncommon for a relatively small number of principal components(from 5 to 10) to capture 90% or more of the variance in the data.

X Correlation between SST clusters and SLP EOFs



(a) 900 clusters

(b) 850 clusters

(c) 800 clusters

(d) 700 clusters

(e) 600 clusters

(f) 500 clusters

(g) 400 clusters

(h) 300 clusters

(i) 200 clusters

(j) 180 clusters

(k) 160 clusters

(l) 120 clusters

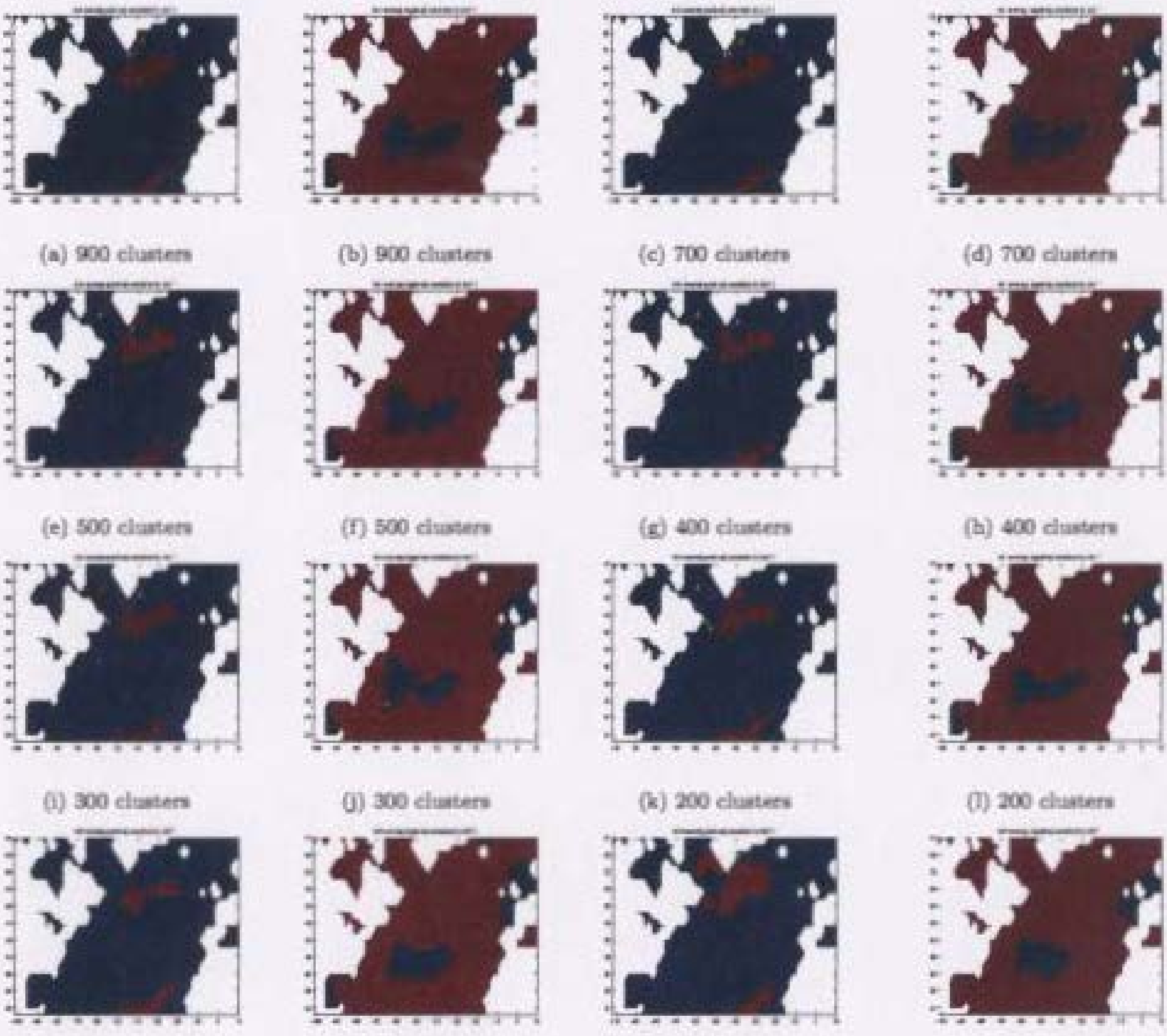
(m) [illegible] clusters

(n) [illegible] clusters

(o) [illegible] clusters

(p) [illegible] clusters

Figure 1: (a)-(p), Correlation coefficients for the KMeans clustering output and EOF1 of SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

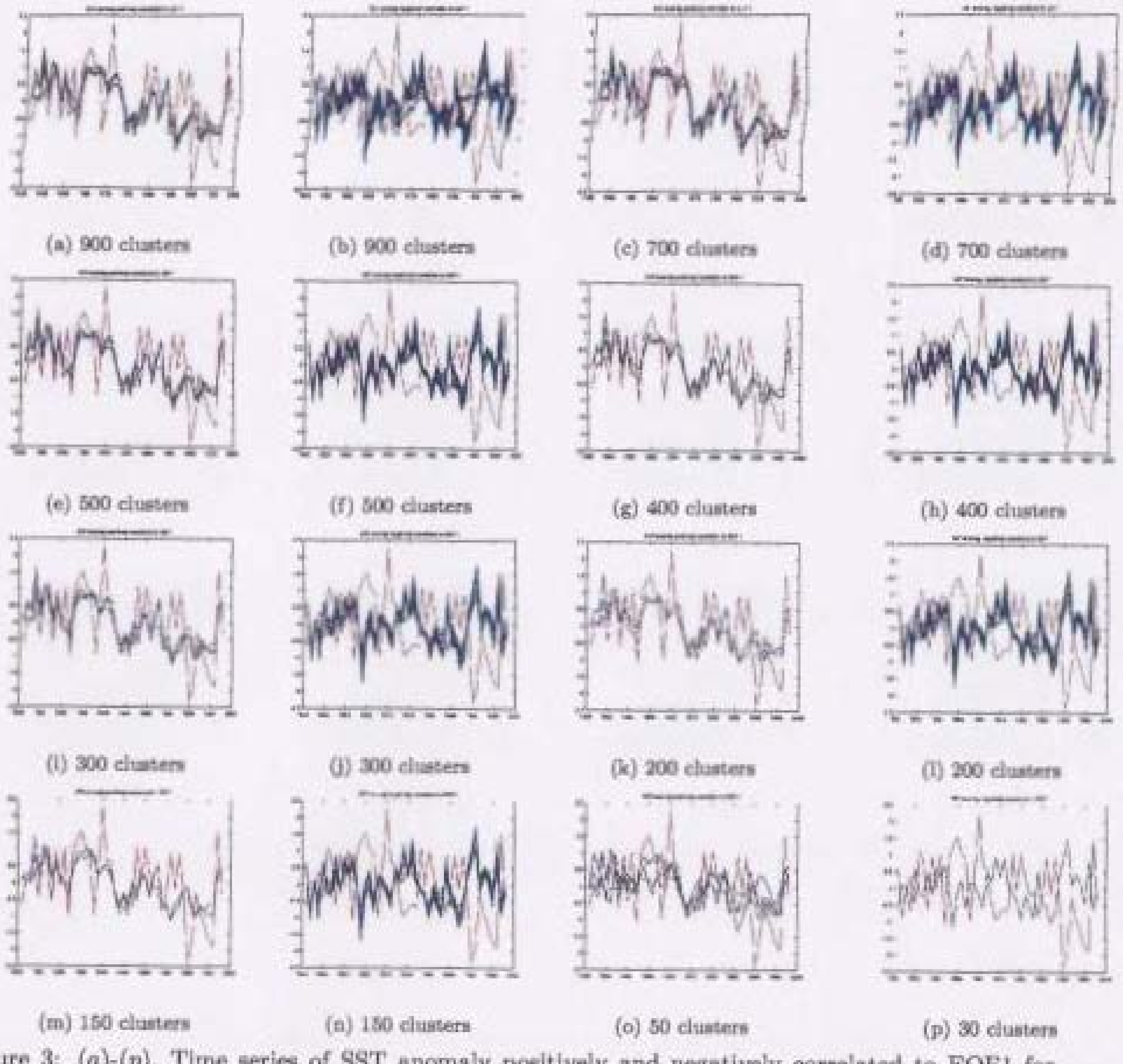
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 2: (a)-(p). Areas of maximum positive and negative correlation for the KMeans clustering output and EOF1 of the SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

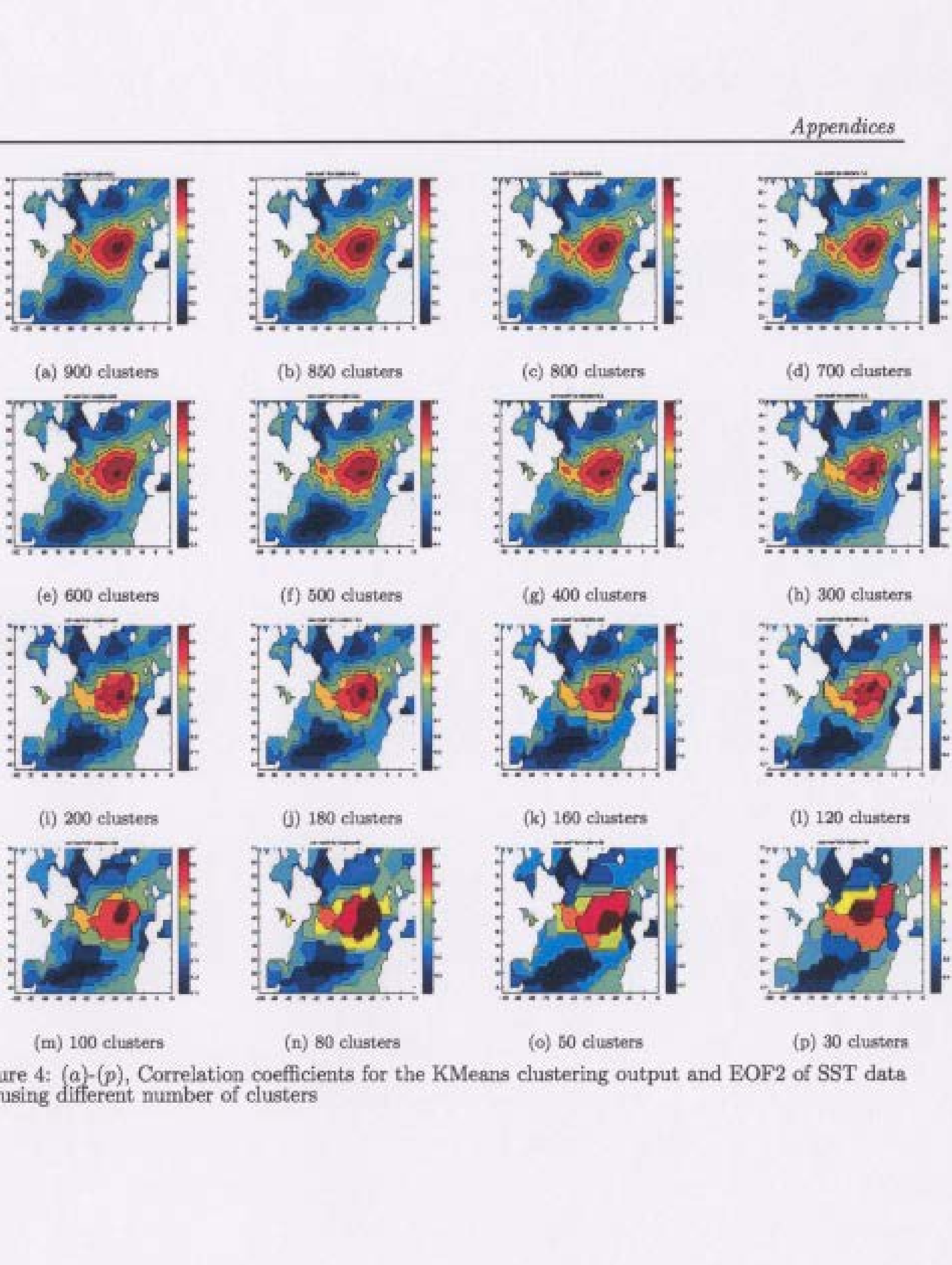
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 3: *(a)*-*(p)*, Time series of SST anomaly positively and negatively correlated to EOF1 for different number of clusters using KMeans clustering algorithm.



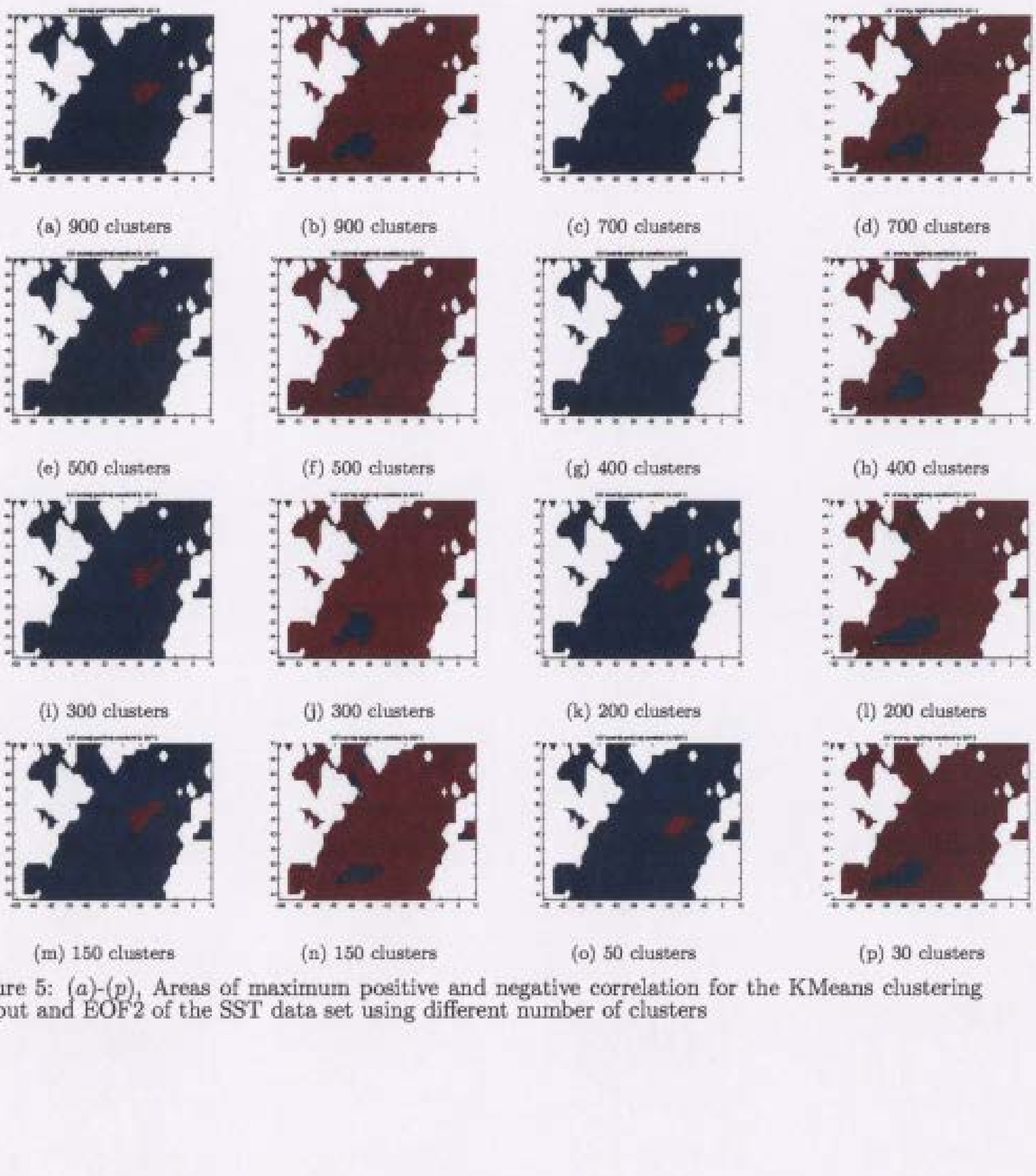
(a) 900 clusters (b) 850 clusters (c) 800 clusters (d) 700 clusters

(e) 600 clusters (f) 500 clusters (g) 400 clusters (h) 300 clusters

(i) 200 clusters (j) 180 clusters (k) 160 clusters (l) 120 clusters

(m) 100 clusters (n) 80 clusters (o) 50 clusters (p) 30 clusters

Figure 4: (a)-(p), Correlation coefficients for the KMeans clustering output and EOF2 of SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

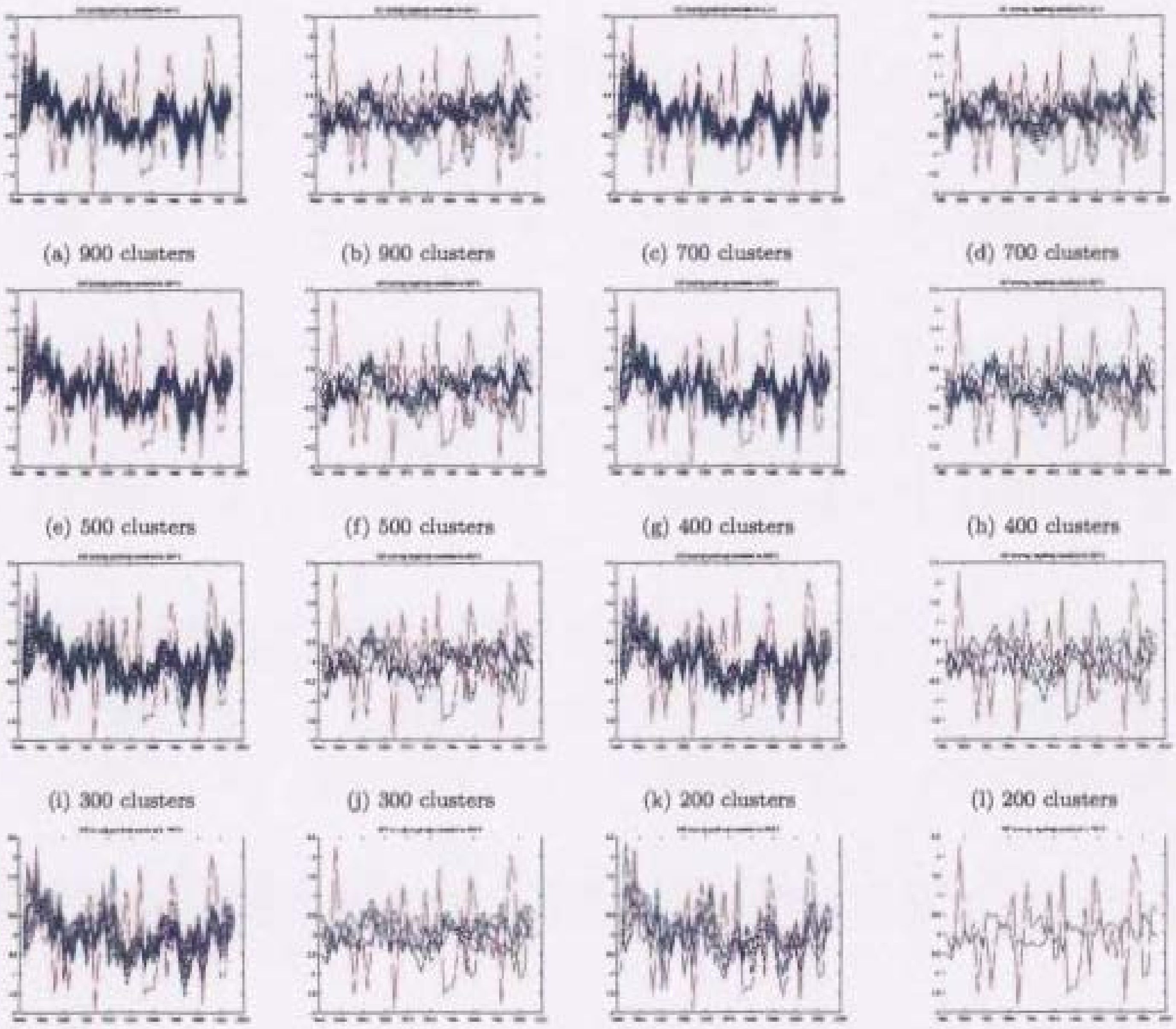
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 5: (a)-(p), Areas of maximum positive and negative correlation for the KMeans clustering output and EOF2 of the SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

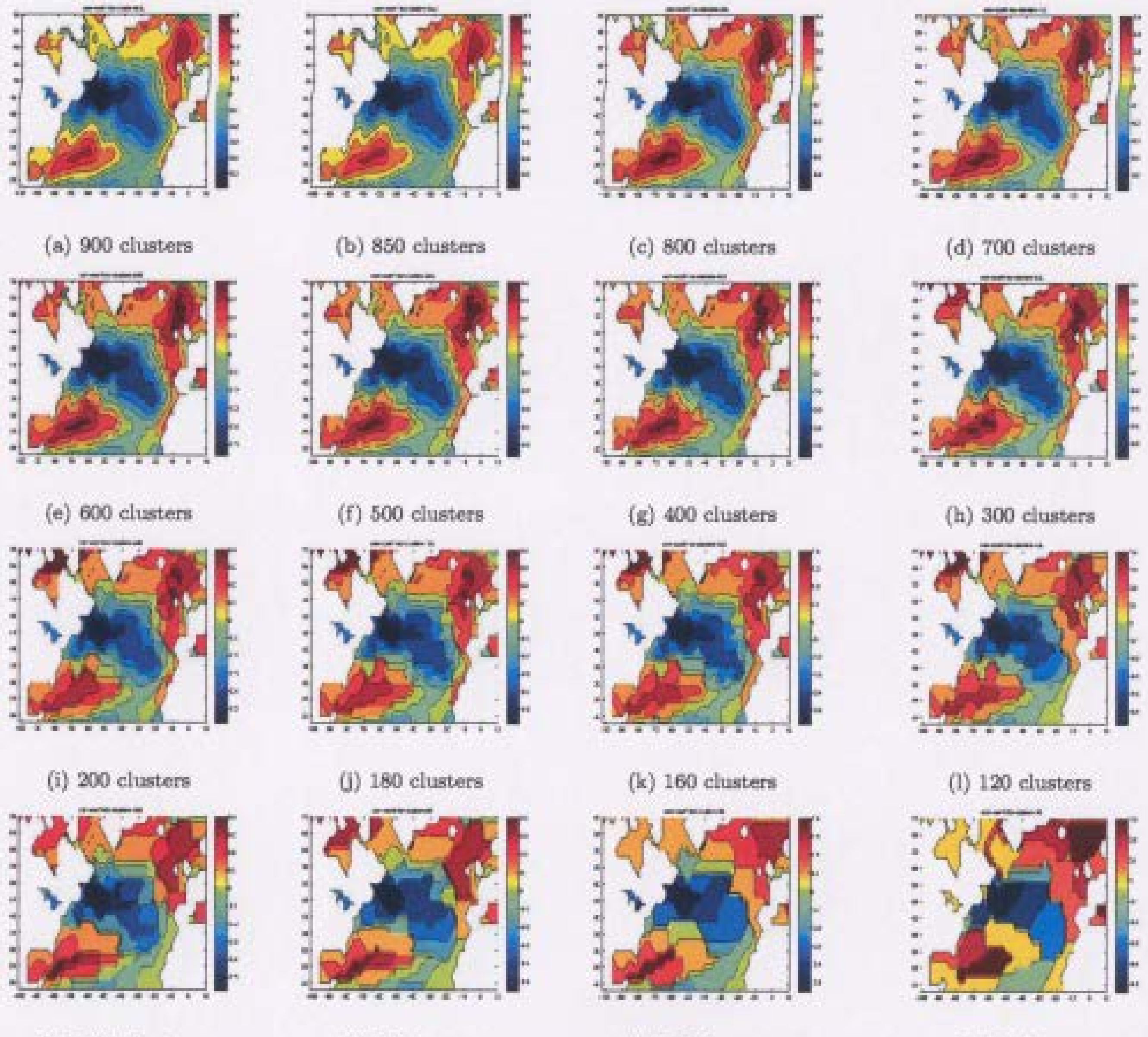
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 6: (*a*)-(*p*), Time series of SST anomaly positively and negatively correlated to EOF2 for different number of clusters using KMeans clustering algorithm.



(a) 900 clusters

(b) 850 clusters

(c) 800 clusters

(d) 700 clusters

(e) 600 clusters

(f) 500 clusters

(g) 400 clusters

(h) 300 clusters

(i) 200 clusters

(j) 180 clusters

(k) 160 clusters

(l) 120 clusters

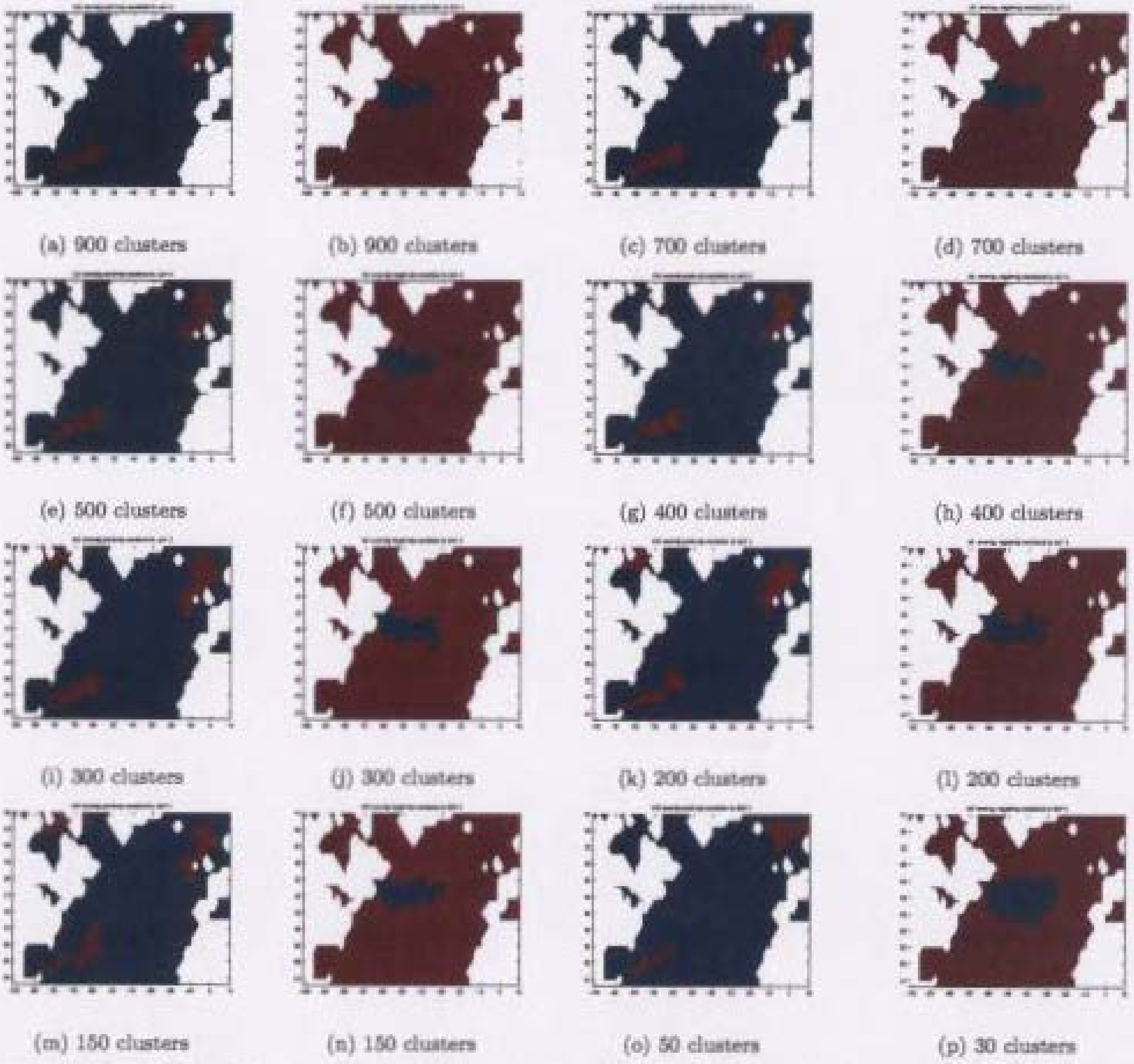
(m) 100 clusters

(n) 80 clusters

(o) 50 clusters

(p) 30 clusters

Figure 7: (a)-(p), Correlation coefficients for the KMeans clustering output and EOF3 of SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

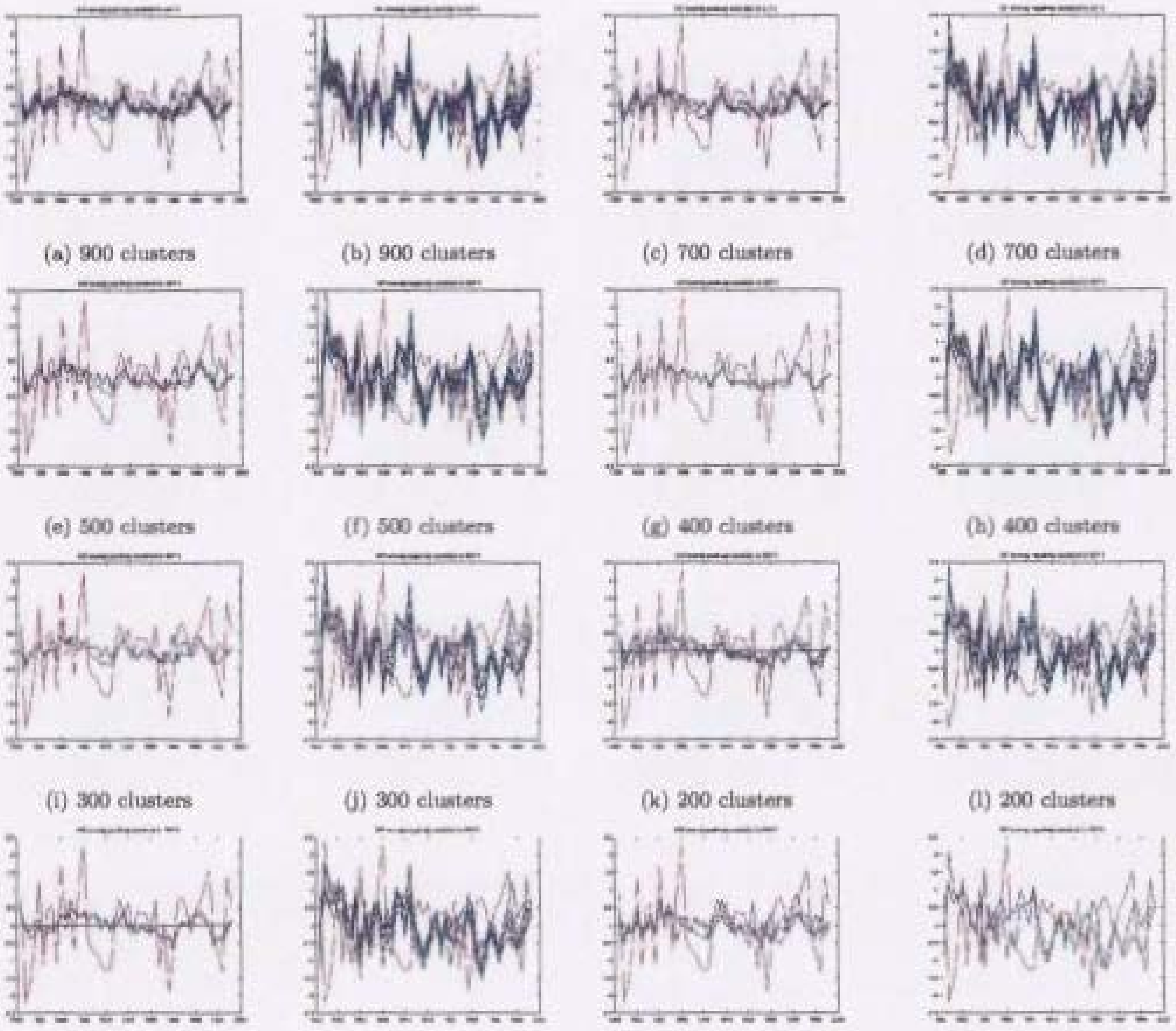
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 8: (a)-(p). Areas of maximum positive and negative correlation for the KMeans clustering output and EOF3 of the SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

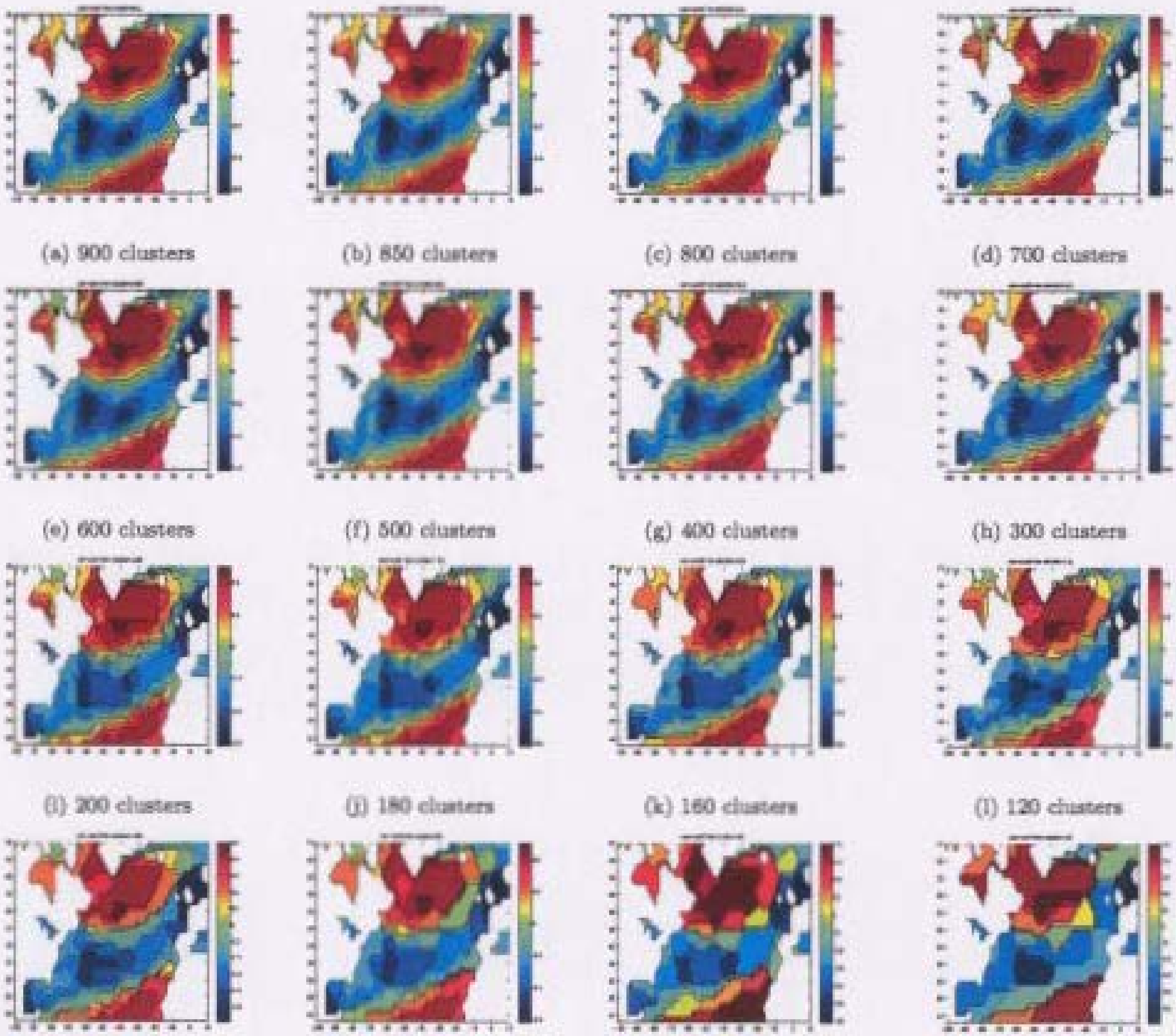
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 9: (a)-(p), Time series of SST anomaly positively and negatively correlated to EOF3 for different number of clusters using KMeans clustering algorithm.



(a) 900 clusters

(b) 850 clusters

(c) 800 clusters

(d) 700 clusters

(e) 600 clusters

(f) 500 clusters

(g) 400 clusters

(h) 300 clusters

(i) 200 clusters

(j) 180 clusters

(k) 160 clusters

(l) 120 clusters

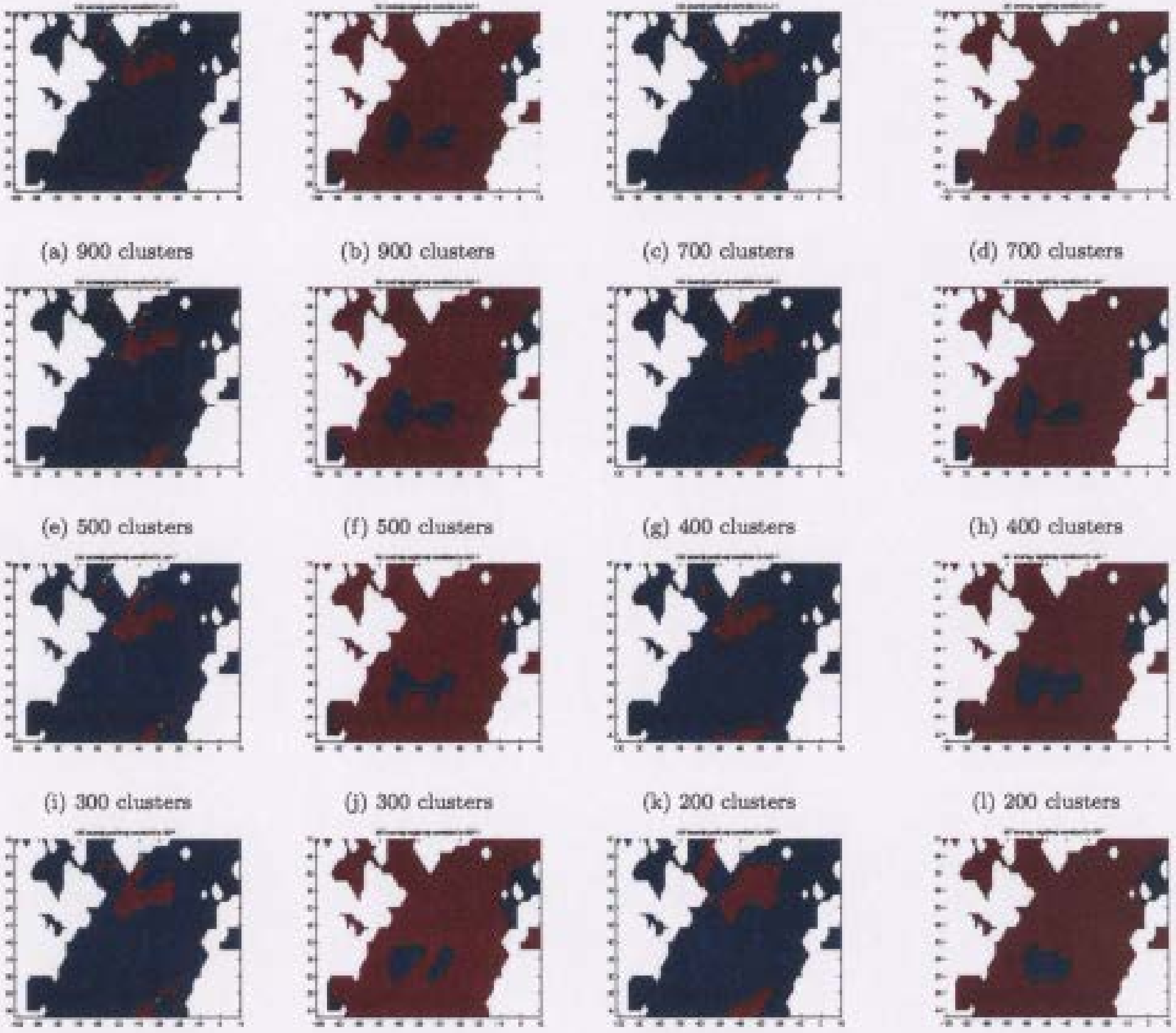
(m) 100 clusters

(n) 80 clusters

(o) 50 clusters

(p) 30 clusters

Figure 10: (*a*)-(*p*), Correlation coefficients for the EM clustering output and EOF1 of SST data set using different number of clusters



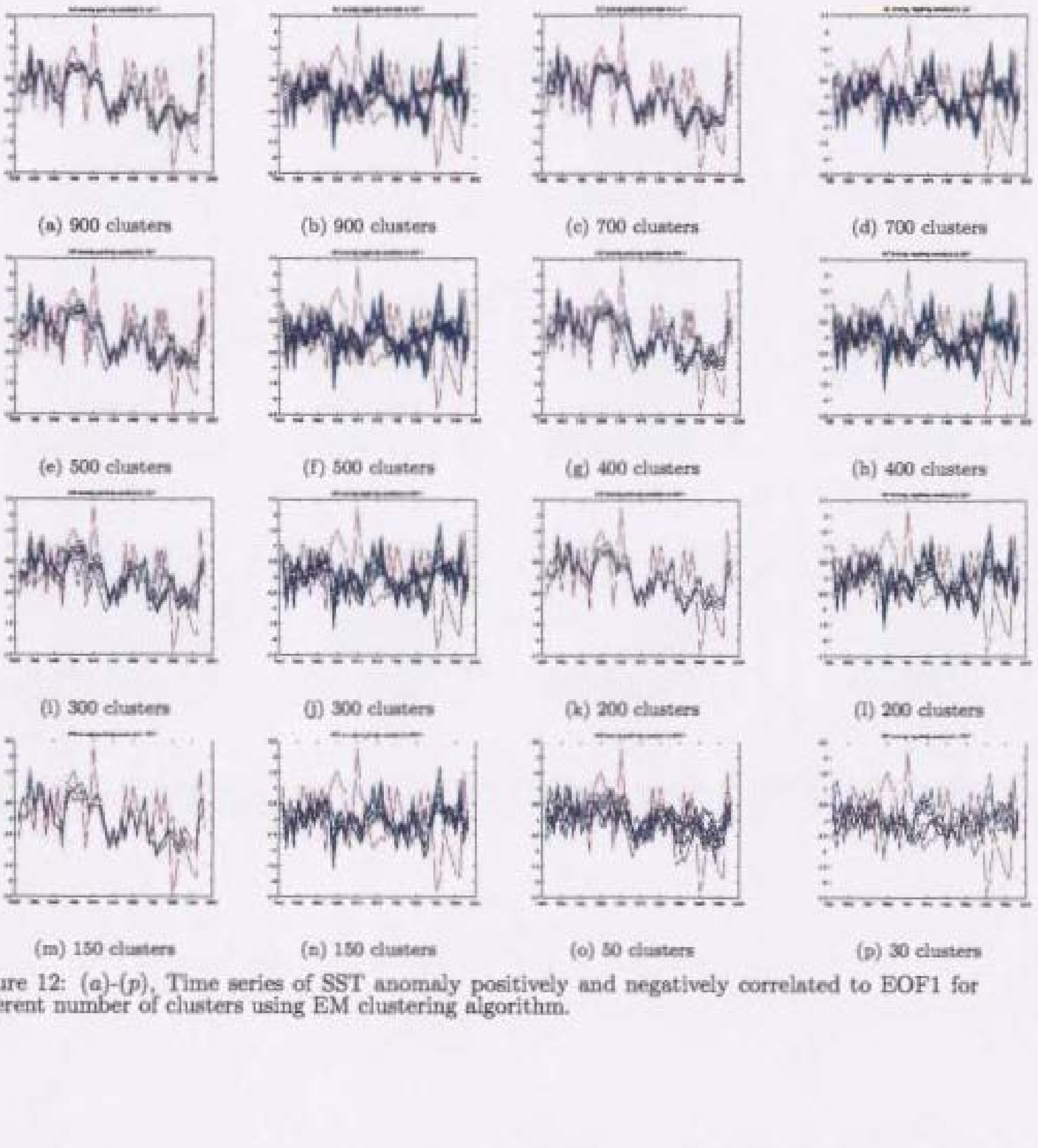
(a) 900 clusters (b) 900 clusters (c) 700 clusters (d) 700 clusters

(e) 500 clusters (f) 500 clusters (g) 400 clusters (h) 400 clusters

(i) 300 clusters (j) 300 clusters (k) 200 clusters (l) 200 clusters

(m) 150 clusters (n) 150 clusters (o) 50 clusters (p) 30 clusters

Figure 11: (a)-(p), Areas of maximum positive and negative correlation for the EM clustering output and EOF1 of the SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

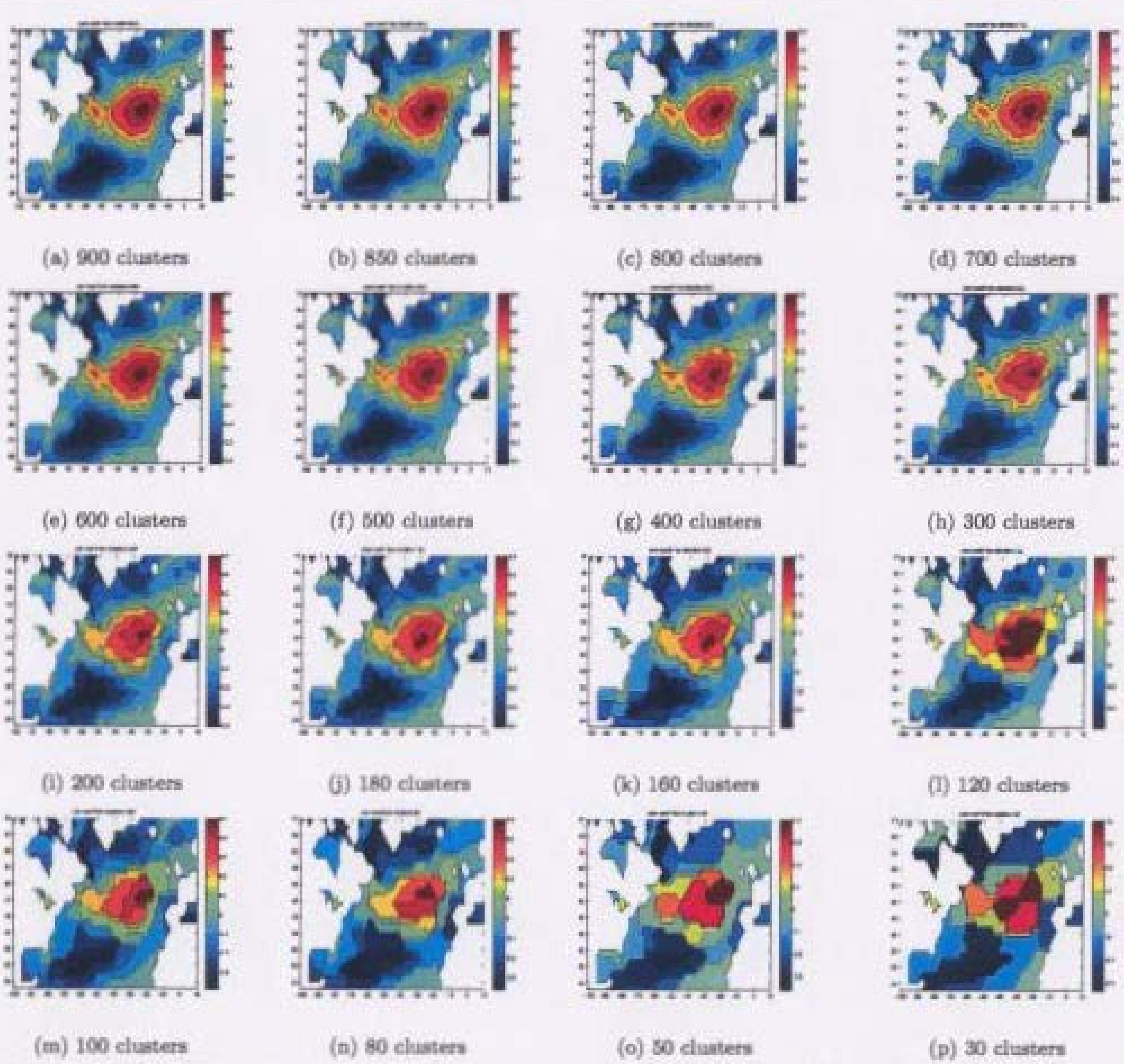
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 12: (a)-(p), Time series of SST anomaly positively and negatively correlated to EOF1 for different number of clusters using EM clustering algorithm.



(a) 900 clusters

(b) 850 clusters

(c) 800 clusters

(d) 700 clusters

(e) 600 clusters

(f) 500 clusters

(g) 400 clusters

(h) 300 clusters

(i) 200 clusters

(j) 180 clusters

(k) 160 clusters

(l) 120 clusters

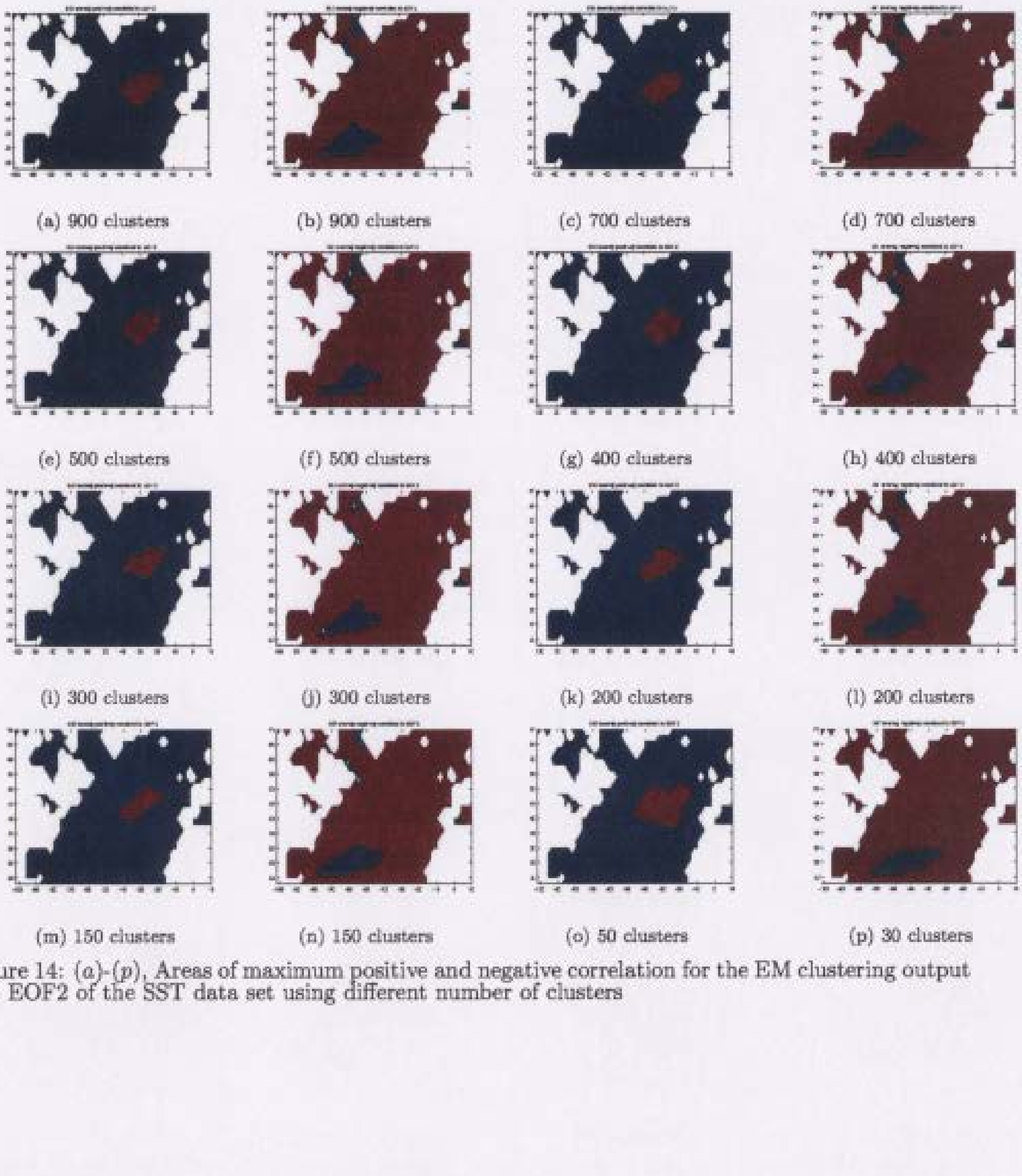
(m) 100 clusters

(n) 80 clusters

(o) 50 clusters

(p) 30 clusters

Figure 13: (a)-(p), Correlation coefficients for the EM clustering output and EOF2 of SST data set using different number of clusters



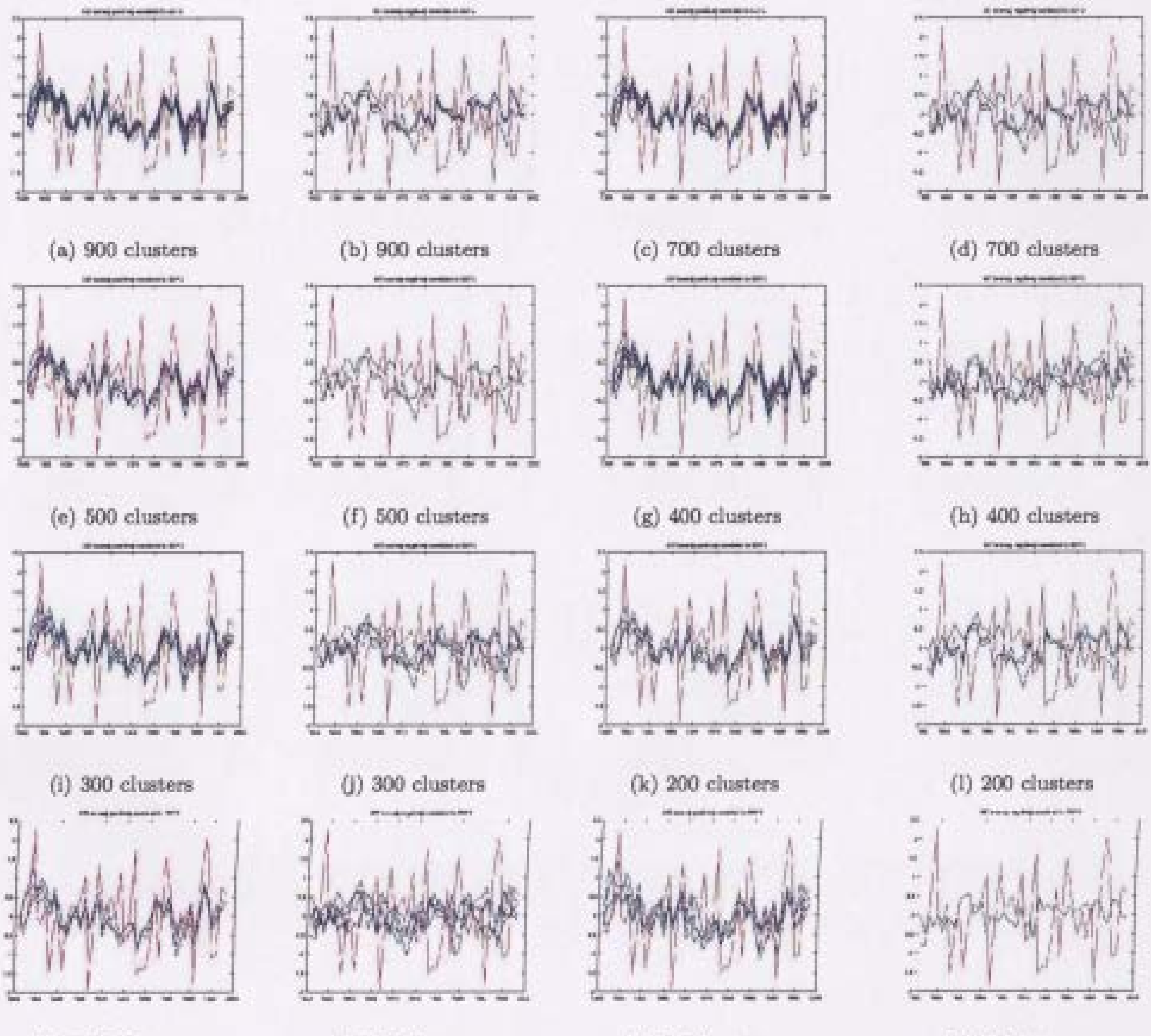
(a) 900 clusters (b) 900 clusters (c) 700 clusters (d) 700 clusters

(e) 500 clusters (f) 500 clusters (g) 400 clusters (h) 400 clusters

(i) 300 clusters (j) 300 clusters (k) 200 clusters (l) 200 clusters

(m) 150 clusters (n) 150 clusters (o) 50 clusters (p) 30 clusters

Figure 14: (*a*)-(*p*), Areas of maximum positive and negative correlation for the EM clustering output and EOF2 of the SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

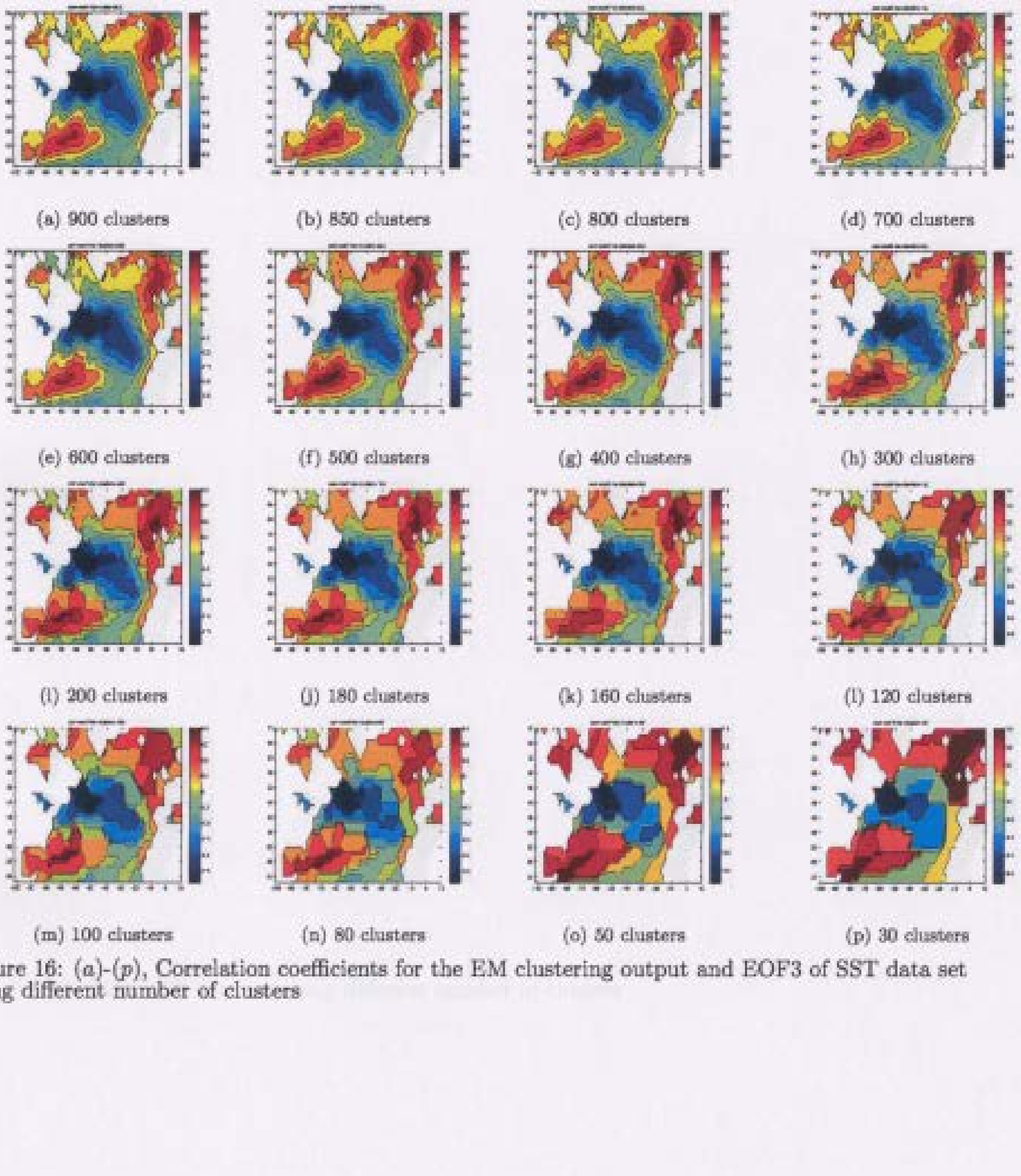
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 15: (a)-(p), Time series of SST anomaly positively and negatively correlated to EOF2 for different number of clusters using EM clustering algorithm.



(a) 900 clusters

(b) 850 clusters

(c) 800 clusters

(d) 700 clusters

(e) 600 clusters

(f) 500 clusters

(g) 400 clusters

(h) 300 clusters

(i) 200 clusters

(j) 180 clusters

(k) 160 clusters

(l) 120 clusters

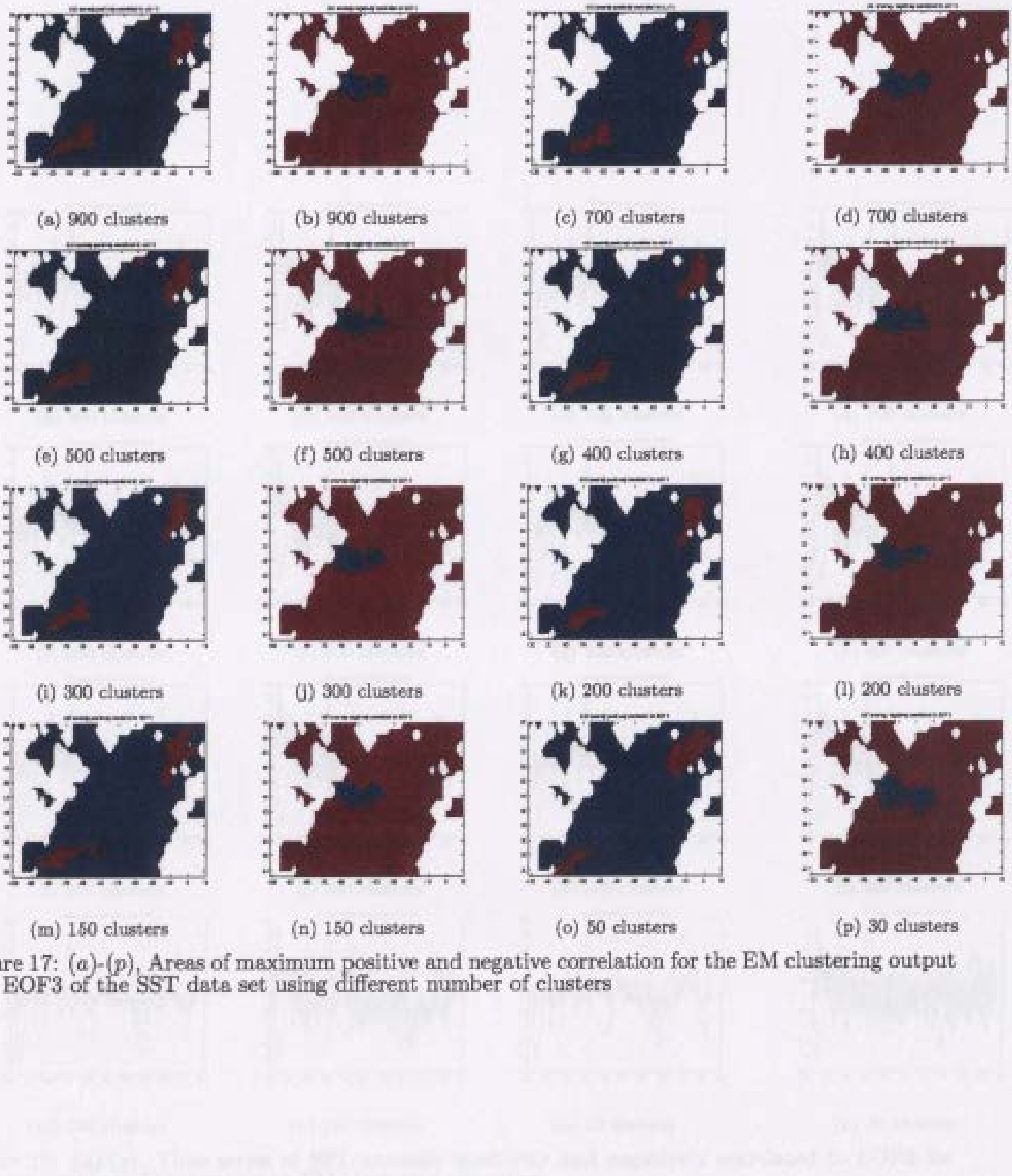
(m) 100 clusters

(n) 80 clusters

(o) 50 clusters

(p) 30 clusters

Figure 16: (a)-(p), Correlation coefficients for the EM clustering output and EOF3 of SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

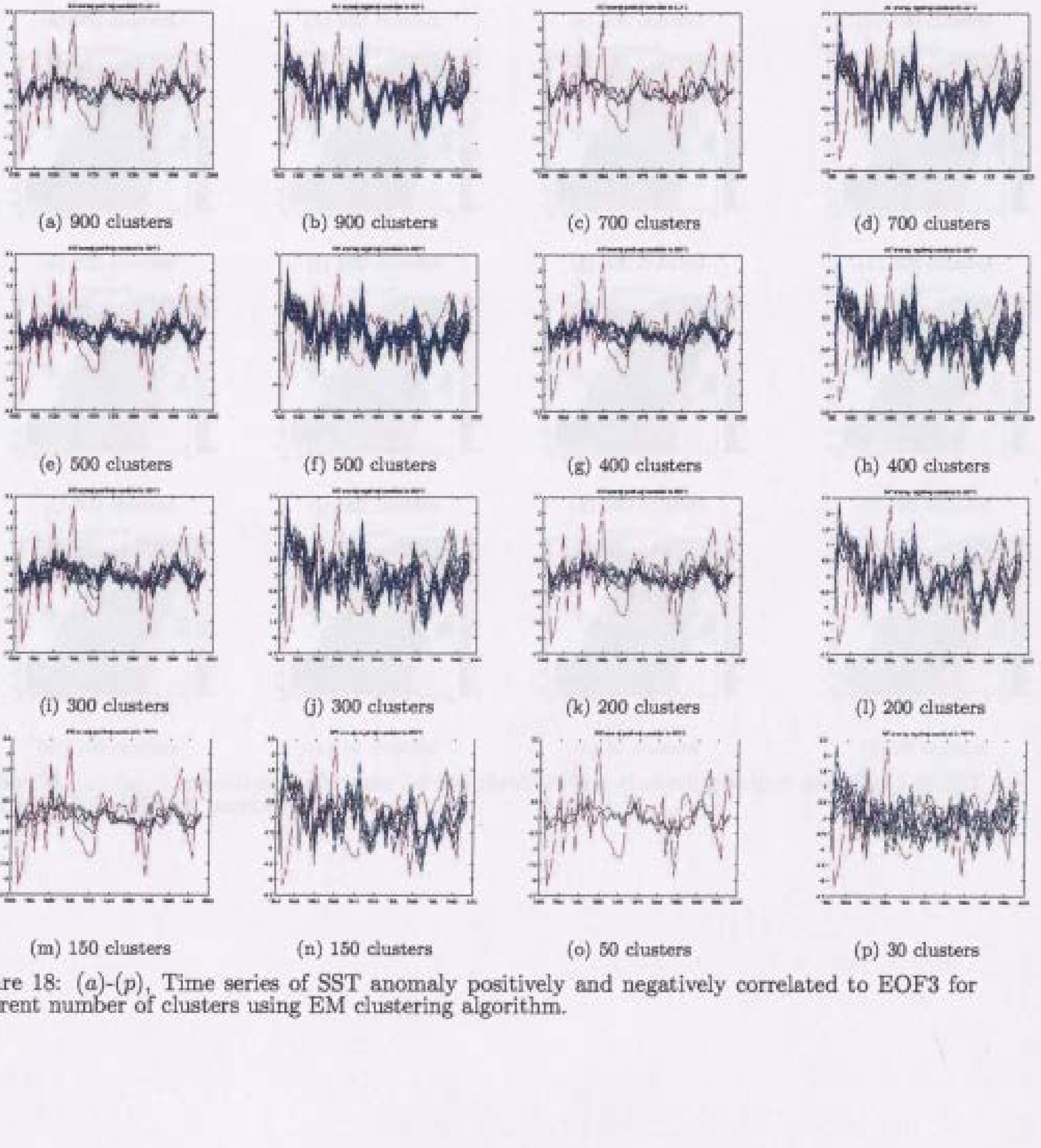
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 17: (a)-(p). Areas of maximum positive and negative correlation for the EM clustering output and EOF3 of the SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

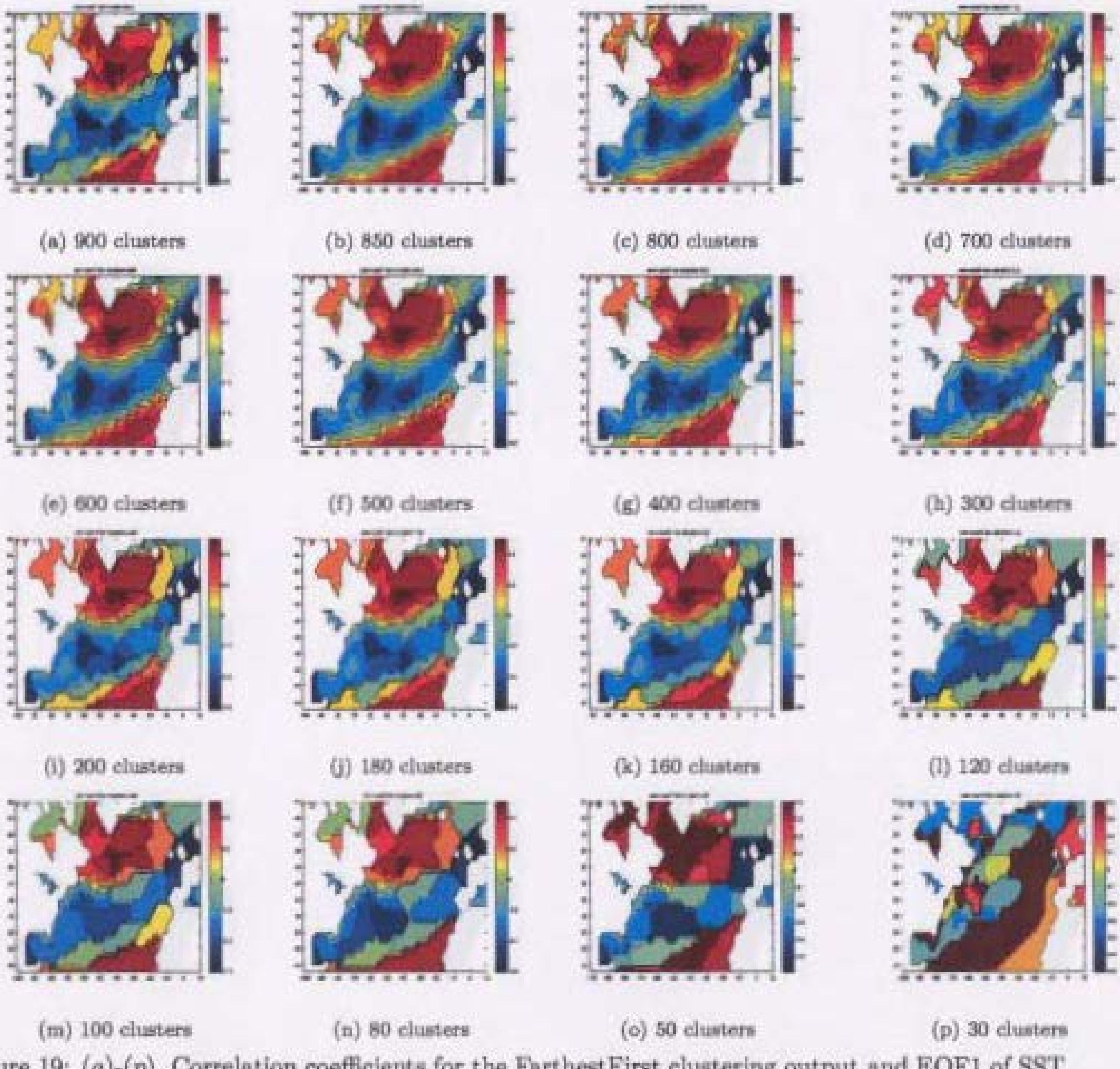
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 18: (a)-(p), Time series of SST anomaly positively and negatively correlated to EOF3 for different number of clusters using EM clustering algorithm.



(a) 900 clusters

(b) 850 clusters

(c) 800 clusters

(d) 700 clusters

(e) 600 clusters

(f) 500 clusters

(g) 400 clusters

(h) 300 clusters

(i) 200 clusters

(j) 180 clusters

(k) 160 clusters

(l) 120 clusters

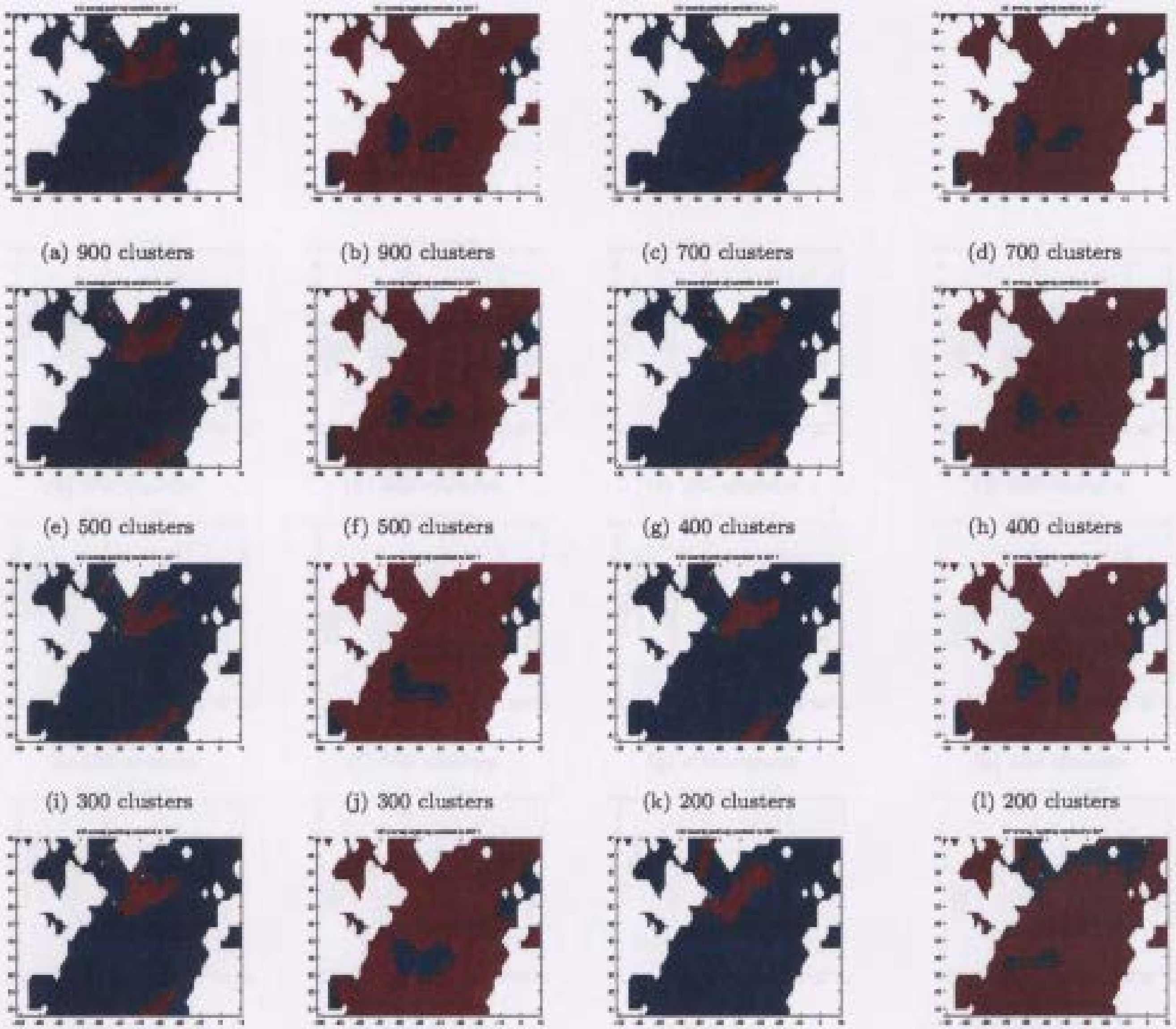
(m) 100 clusters

(n) 80 clusters

(o) 50 clusters

(p) 30 clusters

Figure 19: (*a*)-(*p*). Correlation coefficients for the FarthestFirst clustering output and EOF1 of SST data set using different number of clusters



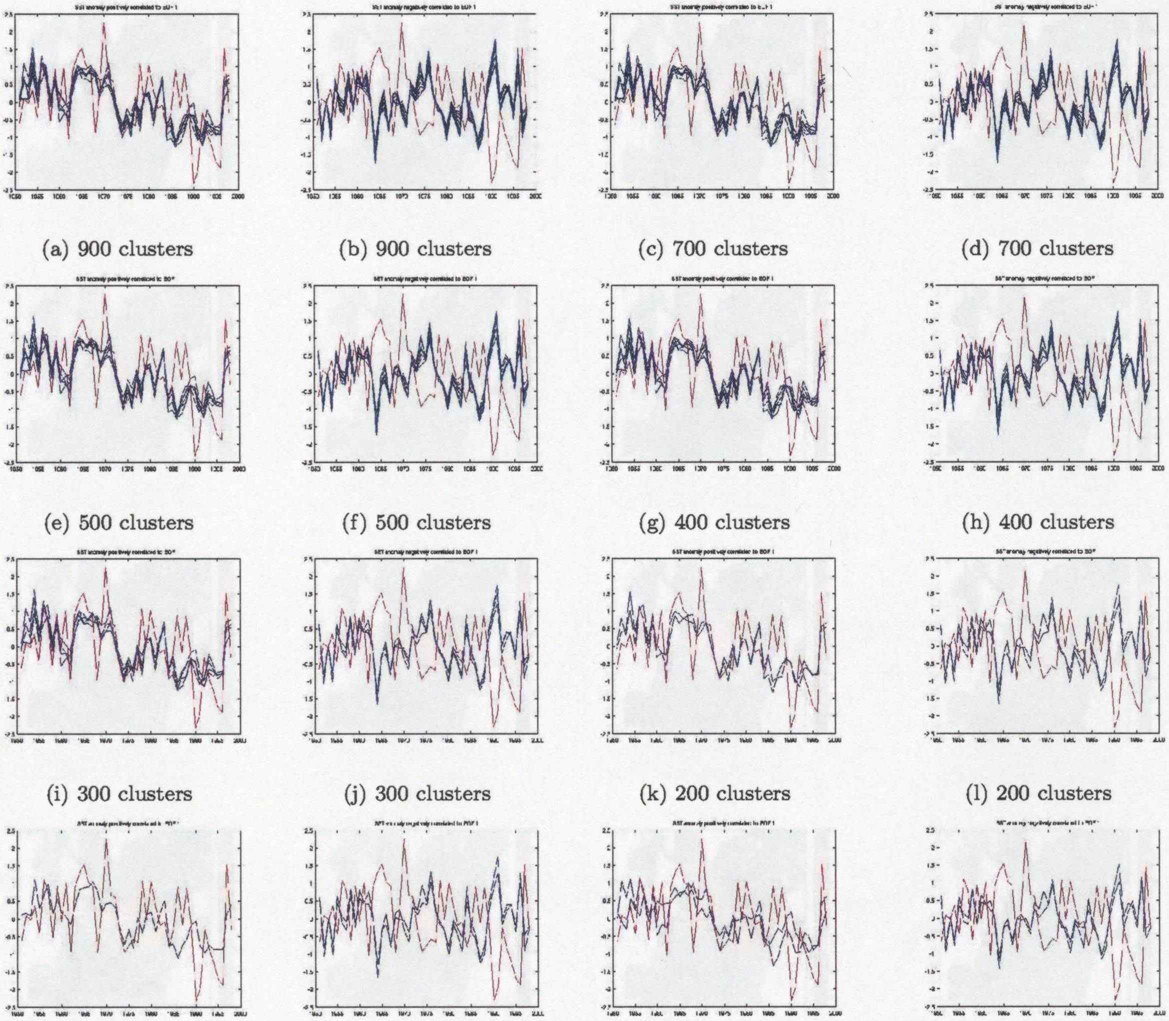
(a) 900 clusters (b) 900 clusters (c) 700 clusters (d) 700 clusters

(e) 500 clusters (f) 500 clusters (g) 400 clusters (h) 400 clusters

(i) 300 clusters (j) 300 clusters (k) 200 clusters (l) 200 clusters

(m) 150 clusters (n) 150 clusters (o) 50 clusters (p) 30 clusters

Figure 20: (*a*)-(*p*), Areas of maximum positive and negative correlation for the FarthestFirst clustering output and EOF1 of the SST data set using different number of clusters



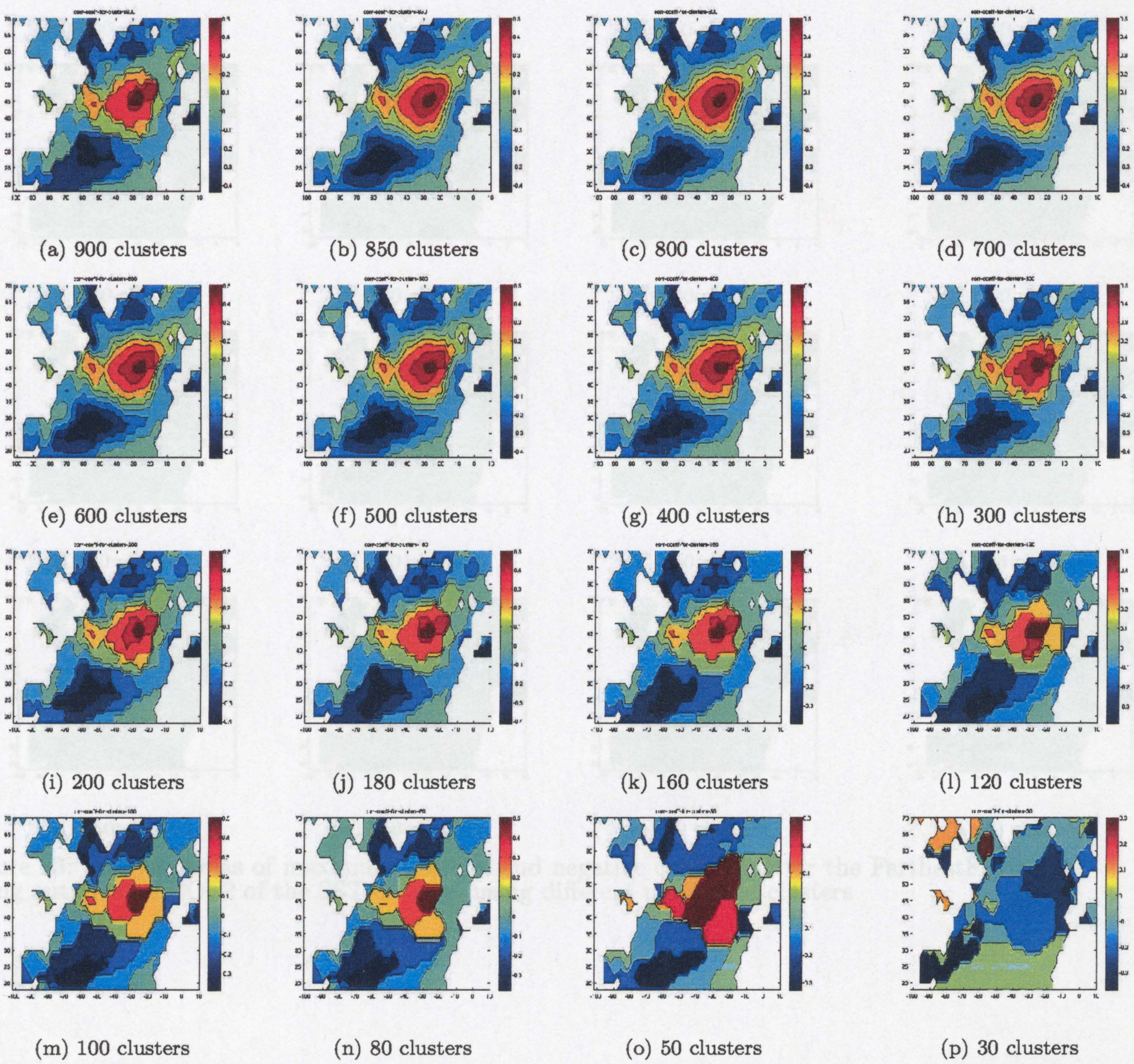
(a) 900 clusters (b) 900 clusters (c) 700 clusters (d) 700 clusters

(e) 500 clusters (f) 500 clusters (g) 400 clusters (h) 400 clusters

(i) 300 clusters (j) 300 clusters (k) 200 clusters (l) 200 clusters

(m) 150 clusters (n) 150 clusters (o) 50 clusters (p) 30 clusters

Figure 21: *(a)*-*(p)*, Time series of SST anomaly positively and negatively correlated to EOF1 for different number of clusters using FarthestFirst clustering algorithm.



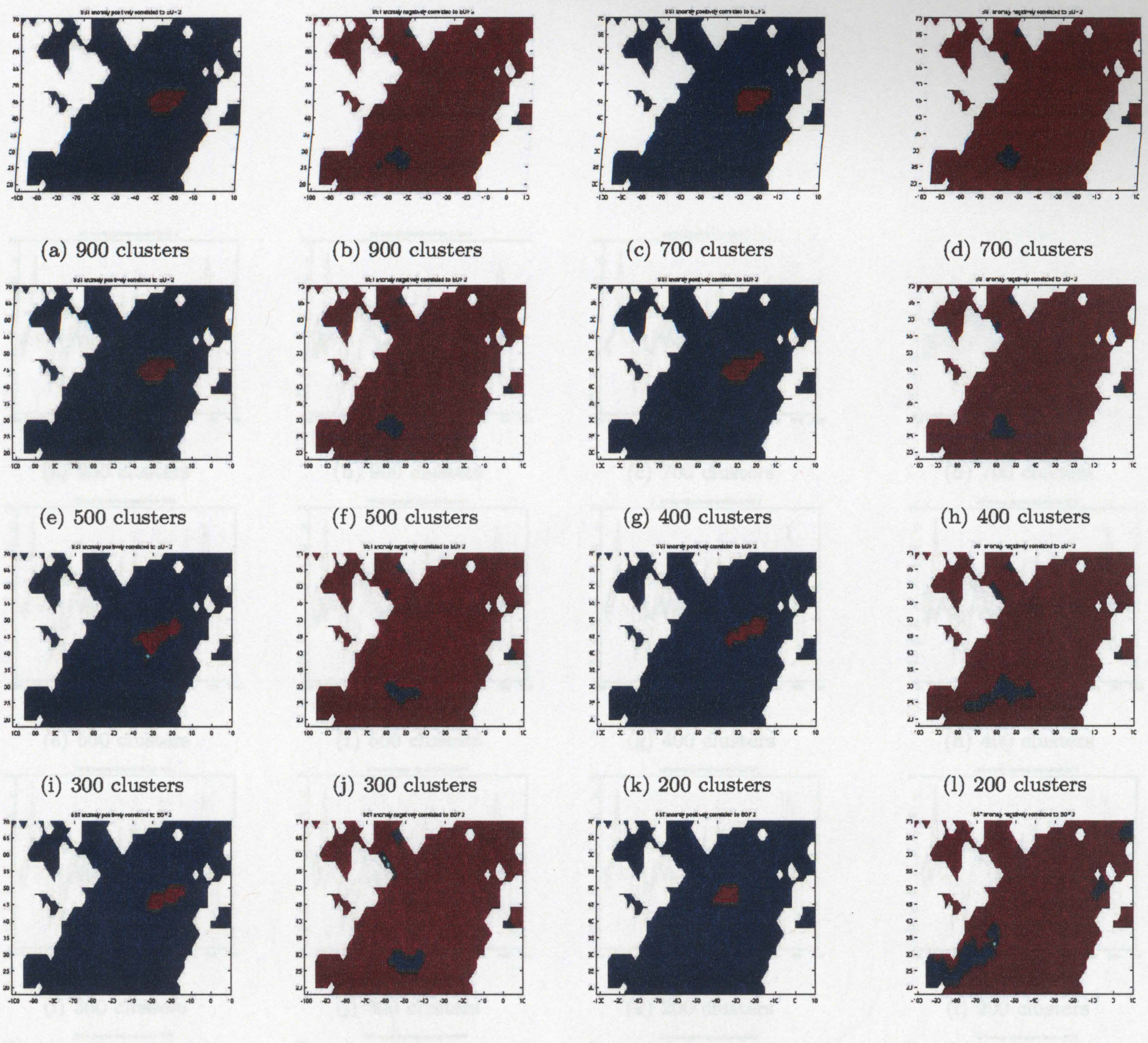
(a) 900 clusters (b) 850 clusters (c) 800 clusters (d) 700 clusters

(e) 600 clusters (f) 500 clusters (g) 400 clusters (h) 300 clusters

(i) 200 clusters (j) 180 clusters (k) 160 clusters (l) 120 clusters

(m) 100 clusters (n) 80 clusters (o) 50 clusters (p) 30 clusters

Figure 22: (a)-(p), Correlation coefficients for the FarthestFirst clustering output and EOF2 of SST data set using different number of clusters



(a) 900 clusters

(b) 900 clusters

(c) 700 clusters

(d) 700 clusters

(e) 500 clusters

(f) 500 clusters

(g) 400 clusters

(h) 400 clusters

(i) 300 clusters

(j) 300 clusters

(k) 200 clusters

(l) 200 clusters

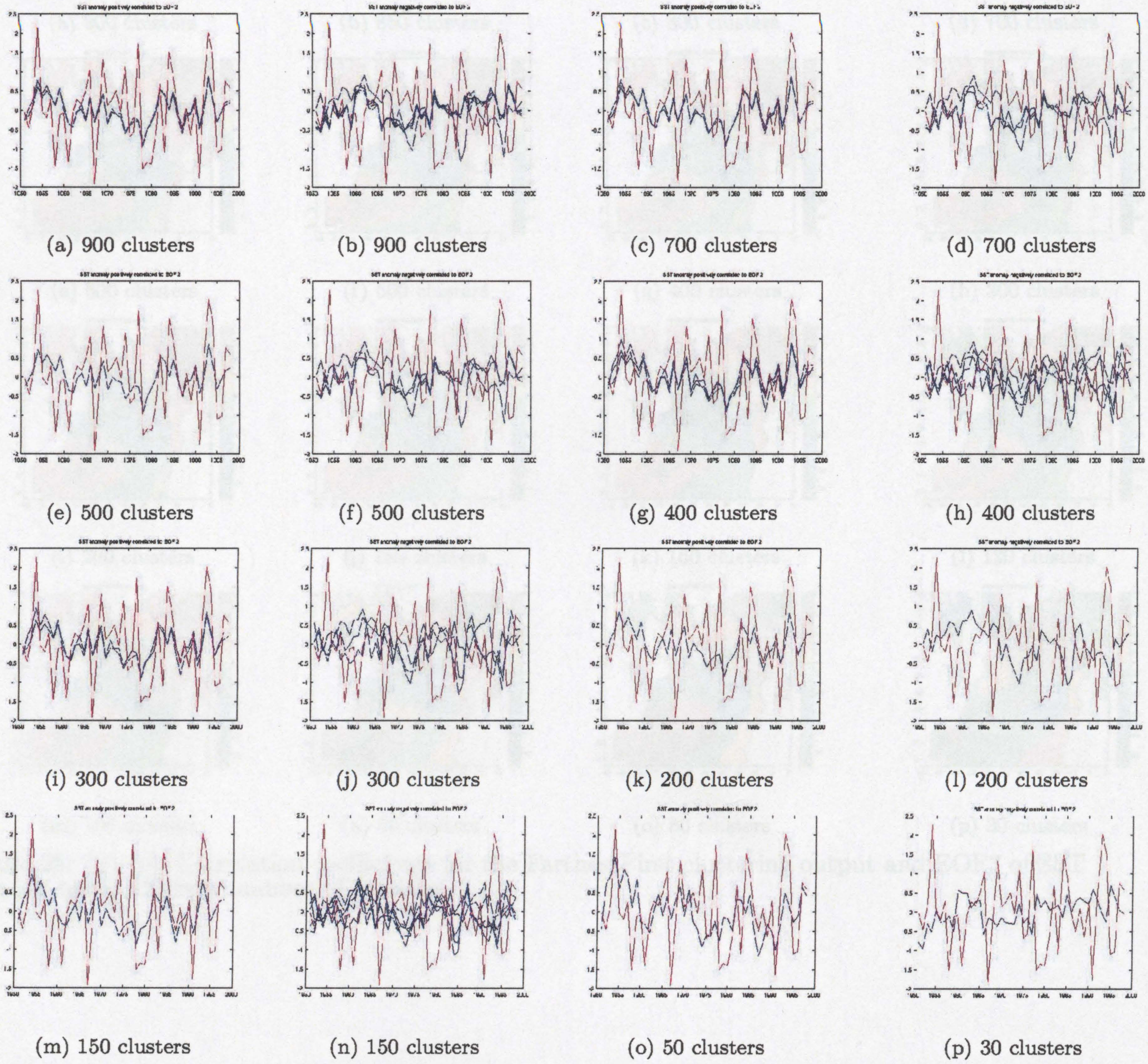
(m) 150 clusters

(n) 150 clusters

(o) 50 clusters

(p) 30 clusters

Figure 23: (*a*)-(*p*), Areas of maximum positive and negative correlation for the FarthestFirst clustering output and EOF2 of the SST data set using different number of clusters



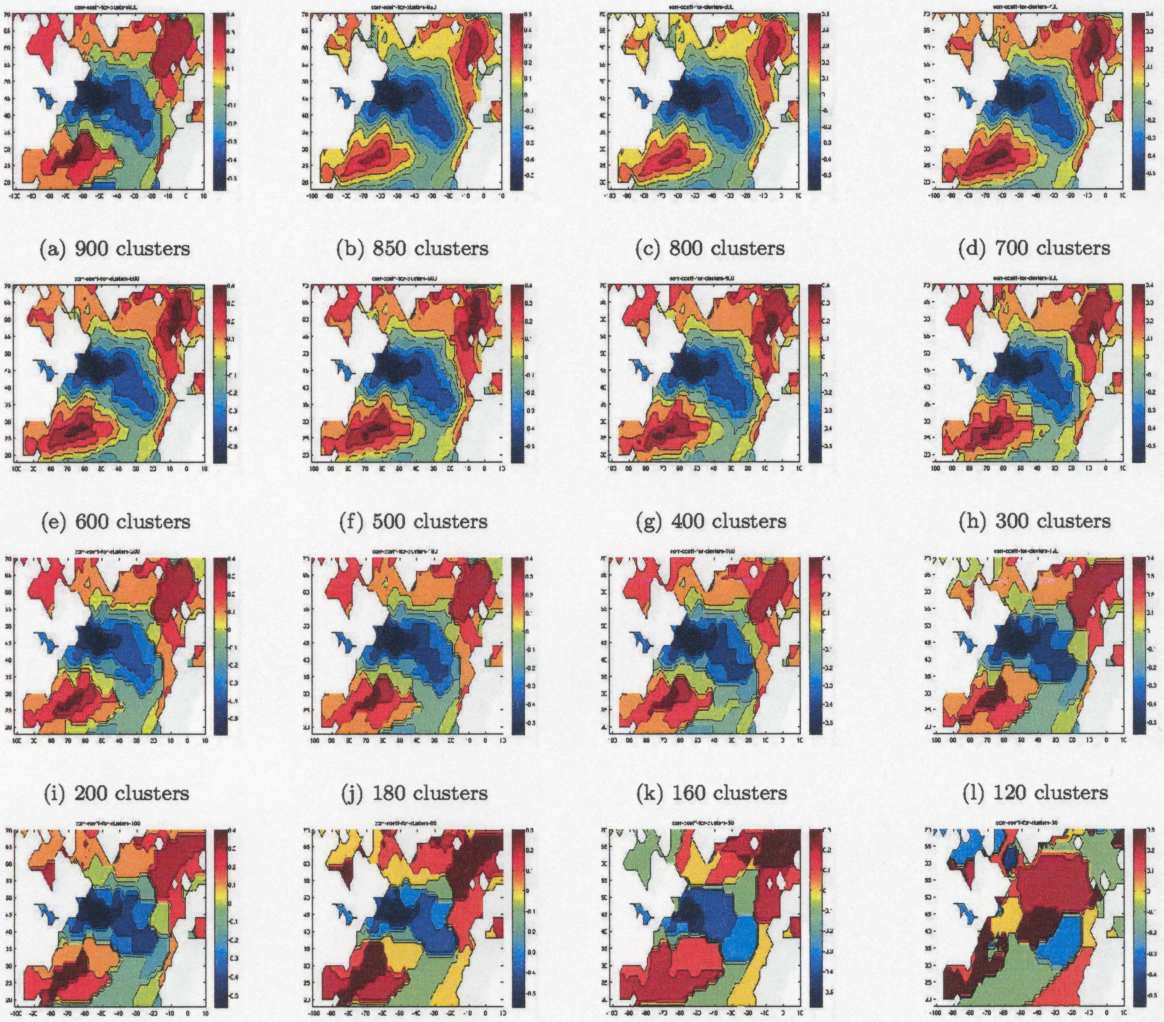
(a) 900 clusters (b) 900 clusters (c) 700 clusters (d) 700 clusters

(e) 500 clusters (f) 500 clusters (g) 400 clusters (h) 400 clusters

(i) 300 clusters (j) 300 clusters (k) 200 clusters (l) 200 clusters

(m) 150 clusters (n) 150 clusters (o) 50 clusters (p) 30 clusters

Figure 24: (a)-(p), Time series of SST anomaly positively and negatively correlated to EOF2 for different number of clusters using FarthestFirst clustering algorithm.



(a) 900 clusters

(b) 850 clusters

(c) 800 clusters

(d) 700 clusters

(e) 600 clusters

(f) 500 clusters

(g) 400 clusters

(h) 300 clusters

(i) 200 clusters

(j) 180 clusters

(k) 160 clusters

(l) 120 clusters

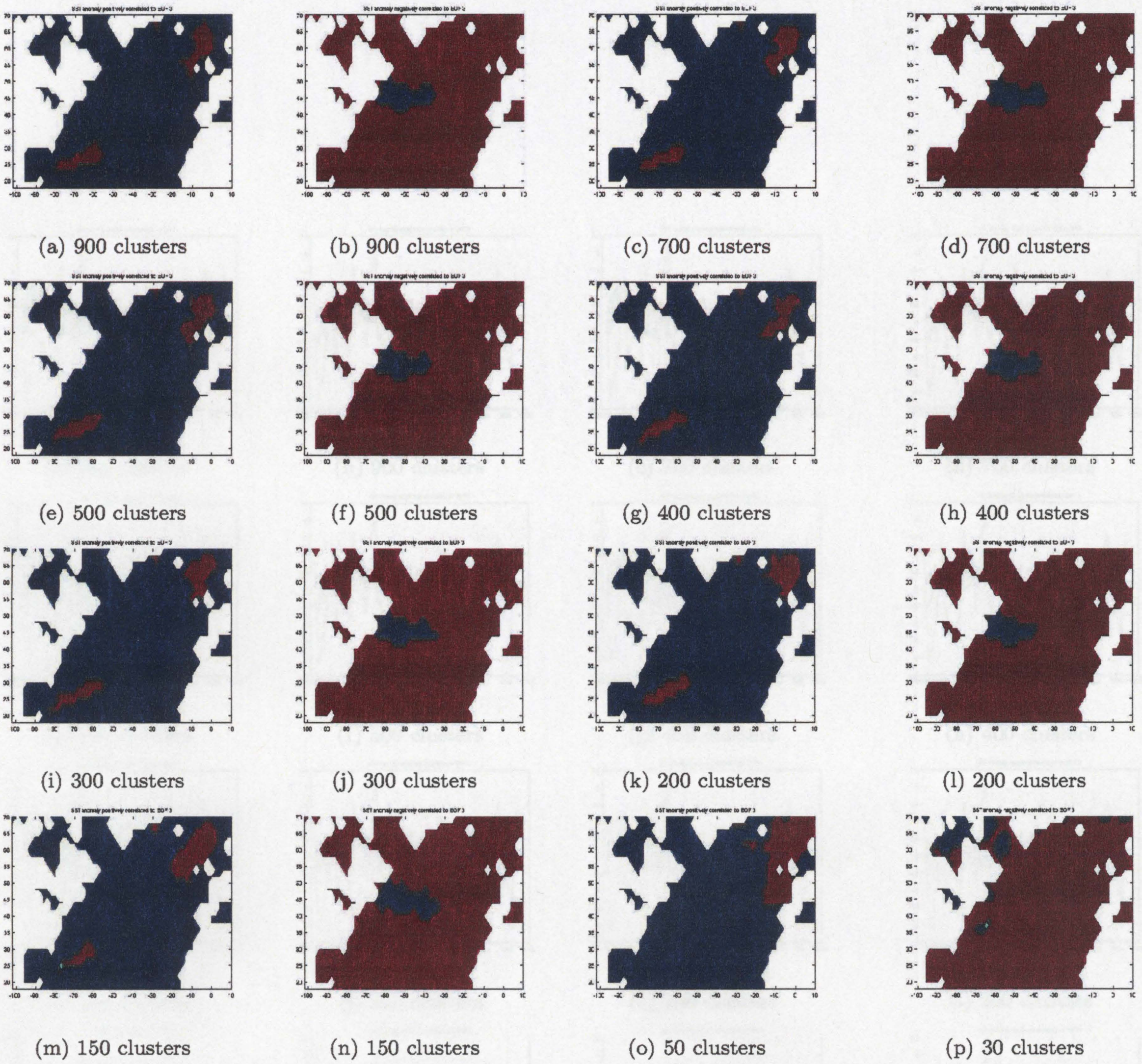
(m) 100 clusters

(n) 80 clusters

(o) 50 clusters

(p) 30 clusters

Figure 25: (a)-(p), Correlation coefficients for the FarthestFirst clustering output and EOF3 of SST data set using different number of clusters


SST anomaly positively correlated to EOF3
SST anomaly negatively correlated to EOF3

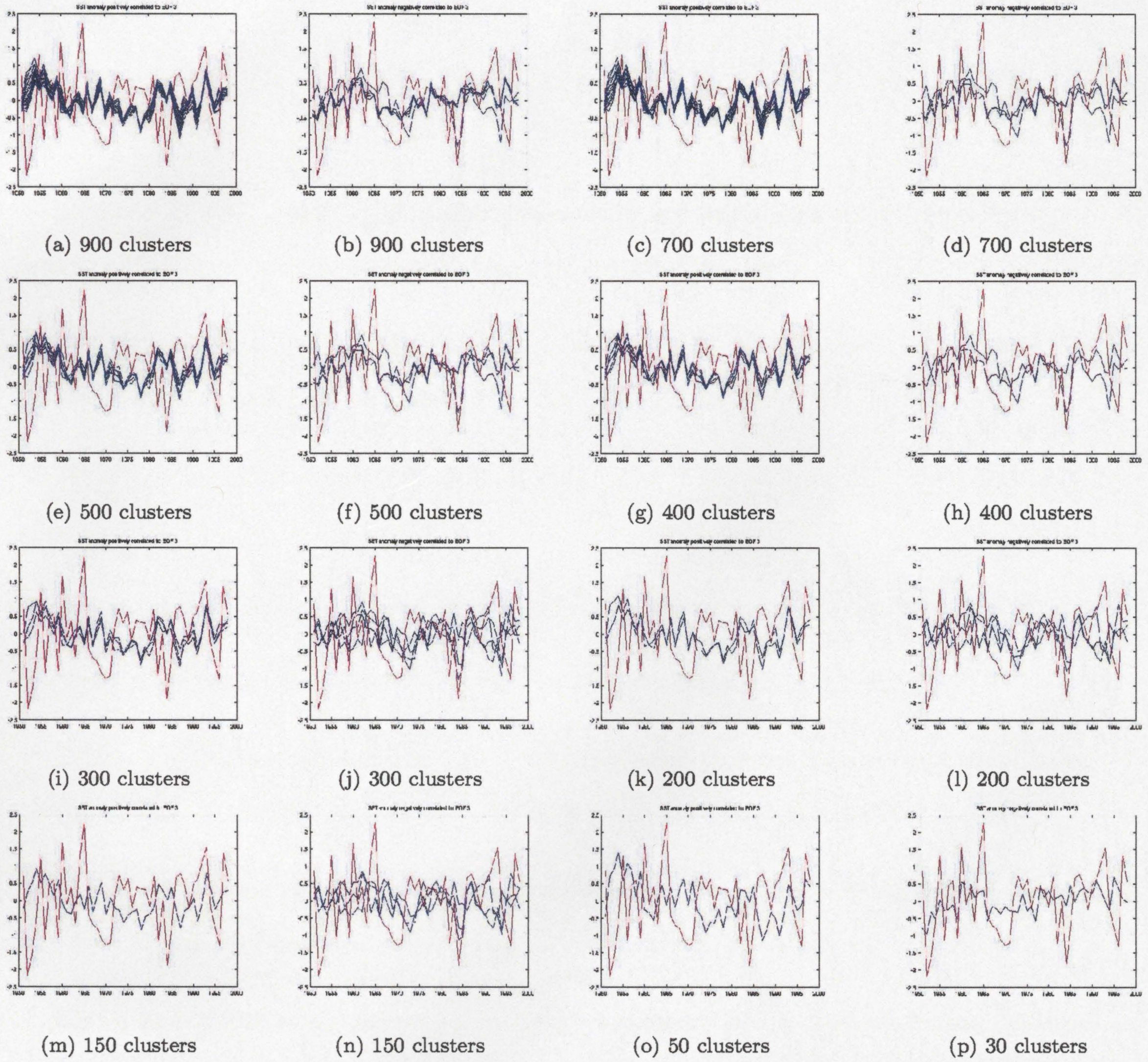
(a) 900 clusters (b) 900 clusters (c) 700 clusters (d) 700 clusters

(e) 500 clusters (f) 500 clusters (g) 400 clusters (h) 400 clusters

(i) 300 clusters (j) 300 clusters (k) 200 clusters (l) 200 clusters

(m) 150 clusters (n) 150 clusters (o) 50 clusters (p) 30 clusters

Figure 26: (a)-(p),Areas of maximum positive and negative correlation for the FarthestFirst clustering output and EOF3 of the SST data set using different number of clusters



(a) 900 clusters (b) 900 clusters (c) 700 clusters (d) 700 clusters

(e) 500 clusters (f) 500 clusters (g) 400 clusters (h) 400 clusters

(i) 300 clusters (j) 300 clusters (k) 200 clusters (l) 200 clusters

(m) 150 clusters (n) 150 clusters (o) 50 clusters (p) 30 clusters

Figure 27: (a)-(p), Time series of SST anomaly positively and negatively correlated to EOF3 for different number of clusters using FarthestFirst clustering algorithm.

Bibliography

[1] A. Barnston and R. Livezey. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea.Rev.*, 115(1):1083–1126, 1987.

[2] D. Battisti, U. Bhatt, and M. Alexander. A modeling study of the interannual variability of the north atlantic ocean. *J. Clim.*, 8(1):3067–3083, 1995.

[3] M. Berry and G. Linoff. *Mastering Data Mining.* John Wiley & Sons, Inc., New York, NY, 1999.

[4] C. Bishop. *Pattern Recognition and Machine Learning.* Springer, LLC, New York, NY, 2001.

[5] J. Bjerknes. Atlantic air-sea interaction. *Advances in Geophysics*, 10(1):1–82, 1964.

[6] W. Bruce, Y. Dandass, and S. Bridges. Scalable knowledge discovery from oceanographic data. In *Proceedings of Artificial Neural Networks in Engineering Conference (ANNIE)*, 1998.

[7] D. Cayan. Latent and sensible heat flux anomalies over the northern oceans: Driving the sea surface temperature. *American Meteorological Society*, 5(1):1–10, 1992.

[8] X. Cheng and J. Wallace. Cluster analysis of the northern hemisphere wintertime 500-hpa height field: Spatial patterns. *Journal of The Atmospheric Science*, 50(16):2674–2696, 1993.

[9] E. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[10] J. Cogswell. *Apache, MySQL, and PHP Web Development.* Wily Publishing Inc., Hoboken, NJ, 2004.

[11] T. Connolly and C. Begg. *Database Systems: Design, Implementation, Management.* McGraw-Hill, Emeryville, CA, 2002.

[12] A. Czaja and C. Frankignoul. Influence of the north atlantic sst on the atmospheric circulation. *Geophys.Res.Lett.*, 26(1):2969–2972, 1999.

[13] C. Darie and K. Watson. *The programmer's guide to SQL.* Springer-Verlag, Inc., New-York, 2004.

[14] S. Dasgupta. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4):555–569, 2005.

[15] T. Delworth. North atlantic interannual variability in a coupled ocean-atmosphere model. *J. Climate*, 9(1):2356–2375, 1996.

[16] C. Deser and M. Blackmon. Surface climate variations over the north atlantic ocean during winter 19001989. *J. Clim*, 6(1):1743–1753, 1993.

[17] B. deYoung. Lecture notes. *Physical Oceanography Class, Physics 3300*, 1:14–15, 2007.

[18] B. deYoung, R. Harris, J. Alheit, G. Beaugrand, N. Mantua, and L. Shannon. Detecting regime shifts in the ocean: data considerations. *Prog. Oceanogr.*, 60:143–164, 2004.

[19] J. Dippner. Recruitment success of different fish stocks in the north sea in relation to climate variability. *Dtsch. Hydrogr.*, 49(1):277–293, 1997.

[20] J. Dippner. Sst anomalies in the north sea in relation to the north atlantic oscillation and the influence on the theoretical spawning time of fish. *Dtsch. Hydrogr.*, 49(1):267–275, 1997.

[21] B. Dong and R. Sutton. Variability in north atlantic heat content and heat transport in a coupled oceanatmosphere gcm. *Climate Dynamics*, 19(1):485–497, 2002.

[22] W. Emery and R. Thompson. *Data Analysis Methods in Physical Oceanography.* Elsevier Science Ltd., Oxford, UK, 1997.

[23] U. Fayyad, G. Piatesky-Shapiro, and P. Smith. From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining*, pages 1–30, 1996.

[24] P. Giudici. *Applied Data Mining – Statistical Methods for Business and Industry.* Wiley Publishing, Inc., London , UK, 2004.

[25] R. Greatbatch. The north atlanctic oscillation. *Stochastic Environmental Research and Risk Assessment*, 1(1):1–20, 2000.

[26] J. Han and M. Kamber. *Data Mining – Concepts and Techniques.* Morgan Kauffman Publishers, Inc., San Francisco, CA, 2005.

[27] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* Massachusetts Institute of Technology, Cambridge, MA, 2001.

[28] D. Hansen and H. Bezdek. On the nature of decadal anomalies in the north atlantic sea surface temperature. *J.Geophys. Res.*, 101(1):9749–9758, 1997.

[29] D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathem. of Operations Research*, 10(2):180–184, 1985.

[30] Y.-P. Huang and F. Sandnes. Efficient mining of salinity and temperature association rules from argo data. *Expert Syst. Appl.*, 35(1):59–68, 2008.

[31] J. Hunter. The national system of scientific measurement. *Department of Statistics and Stanford Linear Accelator Center, Stanford University, Stanford, CA*, 210(21):869–874, 1980.

[32] J. Hurrell. Decadal trends in north atlantic oscillation: regional temperatures and precipitation. *Science*, 269(1):676–679, 1995.

[33] J. Hurrell. Influence of variations in extratropical wintertine teleconnections on northern hemisphere temperature. *Geophys.Res.Lett.*, 23(1):665–668, 1996.

[34] J. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck. *The North Atlantic Oscillation.* American Geophisical Union, Washington, DC, 2003.

[35] A. Jain and R. Dubes. *Algorithms for clustering data.* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[36] F. Jerome. Data mining and statistics: What is the connection? *Department of Statistics and Stanford Linear Accelator Center, Stanford University, Stanford, CA*, 2000.

[37] K. Josef, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):266–239, 1998.

[38] M. Kimoto and M. Ghil. Multiple flow regimes in the northern hemisphere winter. part i: Methodology and hemispheric regimes. *J. Atmos. Sci.*, 50(1):2625–2643, 1993.

[39] S. Klooster and C. Potte. Data mining for the discovery of ocean climate indices. *NASA Ames Research Center*, 2005.

[40] Y. Kushnir. Interdecadal variations in north atlantic sea surface temperature and associated atmospheric conditions. *J. Clim*, 4(1):141–157, 1994.

[41] M. Latif, H. Pohlmann, and W. Park. Predictability of the north atlantic thermohaline circulation. *Predictability of Weather and Climate*, 1(1):160–200, 2006.

[42] S. Levitus, J. Antonov, T. Boyer, R. Locarnini, H. Garcia, and A. Mishonov. Hydrographic changes in the labrador sea, 1960-2005. *Progress In Oceanography*, 73(3-4):242 – 276, 2007.

[43] P. Lionello, P. Malanotte-Rizzoli, and R. Boscolo. Mediterranean climate variability. *Developments in Earth and Environmental Sciences*, 4(1):1–4, 2006.

[44] K. Mann and K. Drinkwater. Environmental influences on fish and shellfish production in the northwest atlantic. *Dtsch. Hydrogr.Environ. Rev*, 2(1):16–32, 1994.

[45] M. Meila and D. Heckerman. An experimental comparison of several clustering and initialization methods. In *Proceedings of the 14th Conf. Uncertainty in Artificial Intelligence*, pages 386–395, 1998.

[46] C. Philip and S. Stolfo. Learning arbiter and combiner trees from partitioned data for scaling machine learning. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 39–44, 1995.

[47] C. Philip and S. Stolfo. Scalable explaratory data mining of distributed geoscientific data. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 2–7, 1996.

[48] L. Pokorna. The correlation of circulation cariabilitiy modes with climate elements in the czech republic. In *Proceedings of the WDS'05 contributed papers, Part III*, pages 480–485, 2005.

[49] S. Rovetta and F. Masulli. Shared farthest neigbor approach to clustering of high dimensionality, low cardinality data. *Pattern Recognition Society*, 39(1):2415–2425, 2006.

[50] R. Seager, Y. Kushnir, M. Visbeck, N. Naik, J. Miller, and G. Krahmann. Cause of atlantic ocean climate variability between 1958 and 1998. *J Clim.*, 13(1):2845–2862, 2000.

[51] M. Steinbach, S. Klooster, P.-N. Tan, C. Potter, and V. Kumar. Temporal data mining for the discovery and analysis of ocean climate indices. *In Proceedings of the KDD Temporal Data Mining Workshop*, 2002.

[52] R. Sutton and M. Allen. Decadal predictability of north atlantic sea surface temperature and climate. *Nature*, 388(1):563–567, 1997.

[53] S. Thompson and J. Wallace. The arctic oscillation signature in wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, 25(1):1297–1300, 1998.

[54] B. Thuraisingham. *Data Mining: Technologies, Techniques, Tools and Trends.* CRS Press, 1999.

[55] A. Timmermann, M. Latif, R. Voss, and A. Grotzner. Northern hemispheric interdecadal variability: A coupled air-sea mode. *J. Climate*, 11(1):1906–1931, 1998.

[56] V. Vaswani. *The Complete Reference to MySQL.* McGraw-Hill, Emeryville, CA, 2004.

[57] I. Witten and E. Frank. *Data Mining - Practical Machine Learning Tools and Techniques.* Morgan Kauffman Publishers, Inc., San Francisco, CA, 2005.

[58] I. Yashayaev. Hydrographic changes in the labrador sea, 1960-2005. *Progress In Oceanography*, 73(3-4):242 – 276, 2007.

[59] X. Zheng, Z. Cai, and Q. Li. An experimental comparison of three kinds of clustering algorithms. In *Proceedings of the International Conference on Neural Networks and Brain*, pages 767–771, 2005.