Statistical Learning To Model Stratospheric Variability

Dissertation

To obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

Submitted to the Department of Earth Sciences
of Freie Universität Berlin

by

Christian Blume

Berlin, 2012

1st Reviewer: Prof. Dr. Katja Matthes

2nd Reviewer: Prof. Dr. Ulrike Langematz

Date of defense: 29th June 2012

"What we observe is not nature itself, but nature exposed to our method of questioning."

- Werner Heisenberg, 1958 -

Abstract

The Variability of the stratosphere is crucial for the evolution of the Earth-climate system as a whole. Stratospheric variability on various time scales is influenced by a number of forcings, such as the Quasi-Biennial Oscillation, the El Niño- Southern Oscillation, the 11-yr solar cycle, or volcanic eruptions, that interact to create a complex system. This link is particularly nonlinear during winter when planetary waves can propagate upward to interact with the stratospheric mean flow. Most commonly, sophisticated chemistry-climate models simulate stratospheric variability, driven by the interactions between dynamics, radiation, and chemistry. However, climate models are computationally expensive and quantifying the importance of forcing factors is difficult. In contrast, statistical methods are mathematically simpler, computationally less expensive, and weight forcing factors according to their importance. Statistical methods learn variability patterns from historical data and can potentially forecast these patterns into the future.

For the first time, a wide class of statistical methods is used in this work to model stratospheric variability in data from observations, reanalyses, and model simulations. The statistical methods are partly nonlinear and nonstationary making them appropriate to cope with the complex feedbacks that govern the stratosphere. These advanced methods, along with a standard linear method, are compared with respect to their ability to model stratospheric variables on different temporal and spatial domains. The considered methods are linear discriminant analysis (LDA), a cluster method based on finite elements (FEM-VARX), a neural network, namely the multi-layer perceptron (MLP), and the support vector machine (SVM). It is shown how an optimal, method-specific set of tuning parameters is estimated using information criteria along with cross-validation.

A prominent example of dynamical wave-mean flow interactions during winter are sudden stratospheric warmings (SSWs). SSWs are dramatic extreme events characterized by a great temperature increase on daily time scales and a breakdown of the polar vortex. While the resulting anomalies can descend downward and provide predictive skill for tropospheric weather conditions, forecasting SSWs themselves remains a difficult task. It is shown in this work that polar stratospheric variability can be modeled and forecasted using nonlinear and nonstationary statistical methods while incorporating all significant forcing factors. Moreover, an approach based on a nonlinear neural network is presented that can classify SSWs in major, minor, and final warmings for the recent climate. The statistical importance of the forcing factors and their nonlinear interrelationships are estimated. In addition, global stratospheric temperature and ozone are statistically modeled due to their specific importance for indicating changes in dynamics and composition. The four statistical methods are used to quantify the natural variability inherent in the stratosphere so that the impact of anthropogenic forcings can be attributed appropriately. Considering various data sets along with the different independent statistical methods makes it feasible to estimate robust uncertainties. Using the statistical methods, variability in temperature and ozone is successfully forecasted up to the year 2100. It is shown in this work that the standard linear method leads to robust results on the monthly scale but is clearly outperformed by the advanced methods on the daily scale.

Zusammenfassung

Die Variabilität der Stratosphäre ist entscheidend für die Entwicklung des gesamten Klimasystems. Stratosphärische Variabilität auf verschiedenen Zeitskalen wird beeinflusst durch eine Anzahl von Antrieben wie der Quasi- zweijährigen Oszillation, der El Niño- Southern Oscillation, dem 11- jährigen solaren Zyklus oder Vulkanausbrüchen welche interagieren und ein komplexes System erzeugen. Diese Verknüpfung ist besonders nicht-linear im Winter, wenn sich planetare Wellen vertikal ausbreiten und mit dem stratosphärischen Grundstrom interagieren. Normalerweise simulieren fortschrittliche Klima- Chemie- Modelle stratosphärische Variabilität, die von den Wechselwirkungen zwischen Dynamik, Strahlung und Chemie bestimmt wird. Klimamodelle sind jedoch rechenintensiv und machen es schwierig die Bedeutung von Antriebsfaktoren zu quantifizieren. Im Gegensatz dazu sind statistische Methoden mathematisch einfacher, rechnerisch weniger anspruchsvoll, und gewichten Faktoren nach ihrer Bedeutung. Statistische Methoden lernen Variabilitätsmuster aus historischen Daten und können diese Muster potentiell vorhersagen.

Zum ersten Mal wird in dieser Arbeit eine breite Klasse von statistischen Methoden verwendet, um stratosphärische Variabilität in Daten aus Beobachtungen, Reanalysen und Modellsimulationen zu modellieren. Die statistischen Methoden sind teilweise nicht-linear und nicht-stationär, so dass sie mit den komplexen stratosphärischen Feedbacks angemessen umgehen. Diese fortschrittlichen Methoden, zusammen mit einem standard linearem Verfahren, werden verglichen mit Bezug zur Fähigkeit der Modellierung von stratosphärischen Variablen auf verschiedenen zeitlichen und räumlichen Domänen. Die verwendeten Verfahren sind die lineare Diskriminantenanalyse (LDA), ein Cluster-Verfahren basierend auf finiten Elementen (FEM-VARX), ein neuronales Netz, nämlich das Multilayer Perceptron (MLP) und die Support Vector Machine (SVM). Es wird gezeigt, wie ein optimaler, methoden-spezifischer Satz von Tuning-Parametern mit Hilfe von Informationskriterien und Kreuzvalidierung bestimmt werden kann.

Ein prominentes Beispiel für dynamische Interaktionen von Wellen mit dem Grundstrom während des Winters sind plötzliche Stratosphärenerwärmungen (SSWs). SSWs sind dramatische Extremereignisse, gekennzeichnet durch einen starken Temperaturanstieg auf täglichen Zeitskalen und einem Zusammenbruch des Polarwirbels. Die daraus resultierenden Anomalien können sich nach unten fortsetzen und bieten Potential für die Vorhersage von troposphärischen Wetterlagen. Prognosen für SSWs selbst sind jedoch besonders schwierig. Es wird in dieser Arbeit gezeigt, dass polare stratosphärische Variabilität mit nicht-linearen und nicht-stationären statistischen Methoden modelliert und prognostiziert werden kann, so lange alle signifikanten Faktoren miteinbezogen werden. Darüber hinaus wird ein Ansatz auf Basis eines neuronalen Netzes vorgestellt, das SSWs in Major-, Minor- und Final- Warmings klassifizieren kann. Dabei wird die statistische Bedeutung der Faktoren und deren nicht-lineare Zusammenhänge abgeschätzt. Stratosphärische Temperatur und Ozon werden aufgrund ihrer spezifischen Bedeutung für Veränderungen in Dynamik und Zusammensetzung statistisch modelliert. Die vier statistischen Methoden werden verwendet, um die inhärente natürliche Variabilität in der Stratosphäre zu quantifizieren. Dadurch kann der Einfluss der anthro-

pogenen Antriebe bestimmt werden. Die Berücksichtigung verschiedener Datensätze zusammen mit den unabhängigen statistischen Verfahren macht es möglich robuste Unsicherheiten abzuschätzen. Schliesslich wird mit Hilfe der statistischen Methoden Variabilität in Temperatur und Ozon erfolgreich bis zum Jahr 2100 vorhergesagt. Es wird in dieser Arbeit gezeigt, dass die standard lineare Methode zu robusten Ergebnissen auf der monatlichen Skala führt. Die fortschrittlichen Methoden sind jedoch deutlich besser für die Modellierung und Vorhersage auf der täglichen Skala geeignet.

Contents

Chapter 1

Introduction

The characteristics of the Earth's atmosphere change depending on spatial and temporal scales. This is referred to as variability. Stratospheric variability, as considered in this work, has become increasingly important over the past 60 years. The destruction of ozone since the 1960s, due to man-made ozone depleting substances (ODS), has not only changed the composition and dynamics of the stratosphere [Solomon, 1999] but also climate of the troposphere, particularly evident in near-surface temperature over Antarctica [Thompson and Solomon, 2002]. The combined effects of the loss of ozone along with the release of greenhouse gases [e.g., Stolarski et al., 2010], most importantly carbon dioxide and methane, into the atmosphere has led to a stratospheric height-dependent cooling trend up to three times stronger in magnitude than the warming trend found at tropospheric levels [Randel et al., 2009]. According to future projections made by various chemistry-climate models, the stratospheric negative temperature trend will persist linearly and well into the second half of the 21st century [SPARC CCMVal, 2010; WMO, 2011]. Due to the ban on ODS instigated by the Montreal protocol in 1987, stratospheric ozone has shown significant signs of recovery over the polar regions in recent years [Randel and Thompson, 2011]. Ozone is expected to reach 1980 levels by the year 2050. From then onwards, ozone could over-recover to values higher than those of 1960, according to future simulations [WMO, 2011]. This expected ozone trend is attributed to the strong increase of greenhouse gases in the 21st century, leading to increased heating and convection in the tropics and a likely acceleration of meridional transport processes in the stratosphere [Eyring et al., 2007].

Aside from long-term trends, as prominently observed in temperature and ozone, caused by anthropogenic emissions many other naturally occurring phenomena are equally important for the stratosphere [e.g., Andrews et al., 1987]. In this work, the natural variability inherent in the stratosphere is quantified. Anthropogenic change can be attributed appropriately, only if the importance of natural variability is known. Natural modes and forcings appear in different facets in the stratosphere. Some are a result of the internal interactions in the stratosphere, such as the Quasi-Biennial Oscillation [Baldwin et al., 2001] or Sudden Stratospheric Warmings [Matsuno, 1971], others are external forcings, such as solar variability [e.g., Gray et al., 2010] or volcanic eruptions [Robock, 2000]. Oceanic variability, as prominently observed in the El Niño- Southern Oscillation [Trenberth,

1997], also plays a major role for stratospheric processes [e.g. Manzini et al., 2006; Randel et al., 2009]. All of these modes and forcings, internal as well as external, interact to create a nonlinear and complex response in the stratosphere [e.g., Calvo et al., 2009; Richter et al., 2011]. This stratospheric response affects processes in other atmospheric layers. A particularly strong coupling is observed during northern hemisphere winter, when the remainders of extreme events in the polar stratosphere tend to descend downward to influence near-surface weather [Baldwin and Dunkerton, 2001; Thompson, 2003].

In order to quantify natural variability in the climate system, chemistry-climate models are commonly used to perform sensitivity experiments. This means that a particular forcing is altered in a certain way, and the consequent change of variability is simulated within the climate model [e.g., SPARC CCMVal, 2010; Richter et al., 2011]. However, this is computationally expensive and is limited by the fact that only a few forcing combinations can be evaluated. Therefore, this work investigates the possibility of modeling stratospheric processes using purely statistical methods. These methods are mathematically simpler and computationally less demanding. Even though statistical methods do not simulate physical processes explicitly, one can use them to learn about underlying relationships to forcing factors. A wide range of statistical methods is applied in this work. The statistical methods are trained with historical data based on reanalyses, observations, and model simulations. Training refers to the process of learning the patterns inherent in the data. Since atmospheric processes are generally nonlinear and nonstationary, statistical methods are applied that are able to cope with these properties. In contrast, standard and widely-used tools of analysis, such as multiple linear regression (MLR), can only model linear and stationary variability. However, because of its robustness and simplicity, MLR is one of the most common tools to analyze stratospheric variability [e.g., Bodeker et al., 1998; Crooks and Gray, 2005; SPARC CCMVal, 2010]. In this work, the modeling and forecast performance of MLR is compared to that of the advanced statistical methods.

Aside from the impact of forcing factors, this work also aims at forecasting stratospheric variability on time scales ranging from days to decades. When the goal is to forecast on the climatological scale, sophisticated chemistry-climate model runs are commonly performed, which is proven to be reliable when combining independent model projections [SPARC CCMVal, 2010; WMO, 2011]. When the goal is to make daily to seasonal weather forecasts in the stratosphere, general circulation model runs consisting of multiple observation constrained ensemble members are performed [e.g., Gerber et al., 2009; Kuroda, 2010]. This is successful on the daily, but looses skill on the seasonal scale. A very different approach is taken in this work. Stratospheric forecasts on various time scales are performed using purely statistical methods. The methods forecast by learning variability patterns with the help of forcing factors known to influence stratospheric processes. The trained model can then be evaluated with factors it has not been trained with. This process is referred to as forecasting in atmospheric science and more generally as statistical pattern recognition [Ripley, 1996].

It is common practice in the stratospheric community to apply a certain statistical method for a specific application. Usually comparisons between different methods are not addressed. For the

first time in this work, a comparison of different and independent statistical methods is performed, applied to the modeling of variability on different stratospheric spatial and temporal scales. Modeling in this sense means the purely data-driven training of a statistical method, often with the help of external factors. The considered methods are linear discriminant analysis (LDA), a cluster method based on finite elements (FEM-VARX), a neural network, namely the multi-layer perceptron (MLP), and the support vector machine (SVM). The training is successful when the statistical model has learned the underlying linear or nonlinear patterns. The modeling performance is assessed with several measures suitable for regression and classification problems. The statistical methods have been selected for their ability to deal with nonlinearity, nonstationarity, multi-dimensional data, as well as regression and classification problems. Another important criterion was whether the corresponding computational libraries are efficiently implemented and freely available. Advanced statistical methods can often not be applied without estimating a set of tuning parameters. A full set of these parameters is referred to as the model architecture and is estimated using information criteria along with cross-validation. The goal of this thesis is to answer three major questions: 1) *What is gained by applying more complex statistical methods, as opposed to simple, linear methods?* 2) *Can statistical methods be used to forecast stratospheric behavior?* 3) *Which of the statistical approaches is best suited to tackle stratospheric problems?*

The variability in stratospheric temperature and ozone is specifically important for indicating changes in dynamics and composition. The considered statistical methods are used to model the large-scale variability in temperature and ozone of the global stratosphere on a zonal mean domain in observations, reanalyses, and model simulations. Despite of the zonally averaged data, the dimensionality is still large. With standard resolutions, the product of the number of latitudes times the number of pressure levels is around one thousand. Therefore, the first step is to reduce the dimensionality of data from various sources, using principal component analysis while still retaining at least 90% of the variance found in the data. The reduced representation is then used to train the statistical models with data representing the recent climate. The different statistical models are compared with respect to their regression performance. Once the variability is modeled, the impact of forcing factors influencing stratospheric processes is calculated. From this, the statistical importance of forcing factors originating naturally, i.e. natural variability, is estimated. With the natural variability being quantified, it is possible to estimate the impact of anthropogenic forcings on recent climate in the stratosphere. Considering various data sets, along with the four independent statistical methods, makes it feasible to estimate robust uncertainties for the impacts of each of the forcing factors that influence stratospheric variability. In addition, long-range statistical forecasts up to the year 2100 are made with reasonable assumptions about the forcing factors. The statistical forecasts are compared to future projections simulated by sophisticated chemistry-climate models [SPARC CCMVal, 2010].

Sudden stratospheric warmings (SSWs) are important phenomena with broad implications for tropospheric weather [Baldwin and Dunkerton, 2001]. They were first observed by scientists at the Free University of Berlin [Scherhag, 1952]. An SSW is a prominent example of the interac-

tion between the stratospheric westerly mean flow during winter and dissipating planetary waves [McIntyre, 1982]. SSWs are characterized by an increase in temperature of up to 70K within a few days in the polar stratosphere, a breakdown of the polar vortex, and a reversal of the zonal flow in the stratosphere [Labitzke and van Loon, 1999]. SSWs are dramatic extreme events and their occurrence is influenced by a number of phenomena, such as the Quasi-Biennial Oscillation, the 11-year solar cycle, the El Niño- Southern Oscillation, volcanic eruptions, etc. Previous efforts investigated the impact of these factors [e.g., Holton and Tan, 1980; Labitzke, 1987; Camp and Tung, 2007a,b; Mitchell et al., 2011] and showed that only the combined effect of all the relevant forcings leads to a consistent picture. However, the statistical analysis of previous work is based on linear and stationary assumptions, which are violated in the polar stratosphere. In this work, the occurrences of SSWs and their classification is investigated with nonlinear statistical methods, while including a range of relevant forcing factors simultaneously [Blume et al., 2012]. SSWs are classified in major, minor, and final stratospheric warmings within this framework, while estimating onset, intensity, and duration of the warming events. The nonlinear interrelationships between the forcing factors are computed. In addition, polar stratospheric variability, as observed in geopotential and temperature anomalies, is forecasted using partly nonlinear and nonstationary statistical methods [Blume and Matthes, 2012]. In a hindcast experiment, the performance of the different methods is assessed. Finally, the polar stratospheric variability for the winter 2011/12 is successfully forecasted using purely statistical methods while making reasonable assumptions about the external factors.

♢ ♢ ♢

This work is arranged as follows. Chapter 2 introduces variability important for the stratosphere. Chapter 3 concisely presents the statistical methods employed in this work and makes first comparisons between them. Data used all throughout this work are introduced in Chapter 4. In Chapter 5, the global stratospheric variability of the recent climate is modeled to then quantify natural variability. In addition, statistical forecasts are made and compared to model simulations. Chapter 6 presents a novel approach to the classification of sudden stratospheric warmings for the time of 1958 to 2010, which is able to distinguish major, minor, and final warmings. Moreover, nonlinear interrelationships between forcing factors are investigated. Chapter 7 aims at forecasting polar stratospheric variability using statistical models and at estimating the importance of external factors. Finally, conclusions and an outlook to further research are given.

Chapter 2

Stratospheric Variability

The Earth's atmosphere can be naturally grouped into different vertical layers, each having certain physical, chemical, and radiative properties. The lowest part of the atmosphere is distinguished between the troposphere up to approx. 10 km (higher in the tropics, lower at high latitudes), followed by the stratosphere up to approx. 50 km, and the mesosphere up to approx. 90 km. This grouping is most commonly made according to the vertical temperature gradient in the atmosphere as indicated in Fig. 2.1. Since the temperature in the stratosphere increases gradually with height due to absorption of solar radiation by ozone, there is little turbulence found making it a stable atmospheric layer. There are different physically consistent simulation approaches that aim at modeling the middle atmosphere and its variability [Andrews et al., 1987]. Historically, General Circulation Models (GCMs) provided the first approach, simulating dynamically and radiatively induced processes. For the stratosphere in particular, chemical processes play a crucial role because of the aforementioned importance of ozone and other gases such as chlorine and bromine. Chemically induced reactions in the atmosphere alter dynamical and radiative processes. Therefore, in order to simulate the interactions between dynamics, radiation, and chemistry, Chemistry-Climate Models (CCMs) are developed since the late 1990s which simulate chemical processes in addition to the physics of dynamics and radiation. The latest generation of CCMs from modeling centers all over the globe were compared in the recent Chemistry Climate Model Validation-2 Activity (CCMVal-2) [SPARC CCMVal, 2010] embedded in SPARC[1]. The most recent advancement are CCMs coupled to ocean, land, and cryosphere models which, as a whole, are referred to as Earth System Models (ESMs). ESMs often simulate interactions with the biosphere.

Almost all ozone, as shown in Fig. 2.1, is found in the stratosphere between 100 and 1 hPa where the so-called ozone layer peaks at around 30 hPa in the tropics and at around 70 hPa over the polar region. Ozone is responsible for the absorption of a great deal of solar radiation in the ultra-violet (UV) band making ozone important for all life on earth because of the damaging effects that UV may cause to living organisms. Ozone is produced in the tropics involving solar radiation and is dynamically transported to the poles via global meridional circulation processes. Atmospheric

[1]SPARC is short for Stratospheric Processes And their Role in Climate. SPARC is a core project of the World Climate Research Program (WCRP).

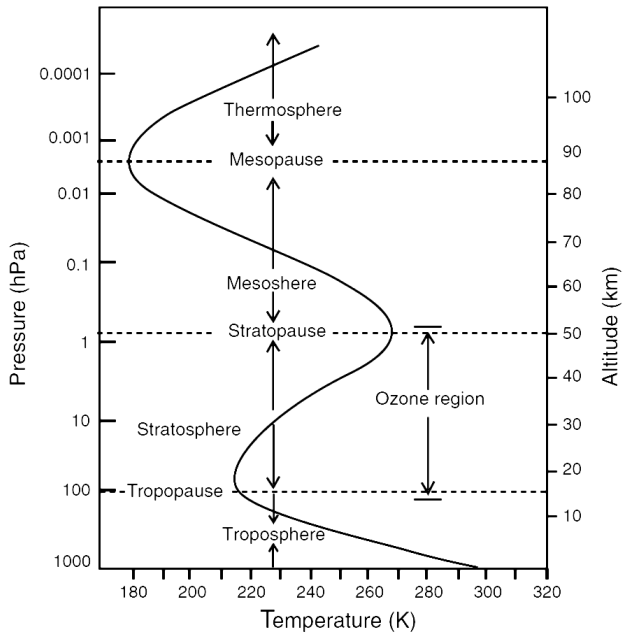

Figure 2.1: The typical vertical evolution of atmospheric temperature in Kelvin depending on height from the troposphere to the thermosphere. Figure from Mohanakumar [2008].

dynamics have a strong seasonal dependence. During summer, the stratosphere is in radiative balance because planetary waves from the troposphere cannot propagate upwards due to an easterly flow in the zonal mean. During winter, however, the westerly mean-flow allows wave propagation and possible dissipation in stratospheric levels causing nonlinear interactions and dynamical feedbacks. Prominent feedbacks are sudden stratospheric warmings as the main manifestation of stratospheric variability during winter.

2.1 Sudden Stratospheric Warmings

The variability of the north-polar stratospheric vortex is a key dynamical feature of the middle atmosphere [Labitzke and van Loon, 1999]. Specifically, its breakdown during winter resulting in a sudden stratospheric warming [Scherhag, 1952; Labitzke, 1972]. Sudden stratospheric warmings (SSW) are characterized by stratospheric temperatures increasing by 25K on average, but up to 70K in extreme cases, within a few days. The polar vortex is a low pressure region in the stratosphere over the pole during winter. The large temperature gradient between low latitudes and the pole in the stratosphere leads to strong thermal westerly winds in mid latitudes during winter which is referred to as the polar night jet [Baldwin and Holton, 1988]. The field of high potential vorticity in the polar stratosphere defines the polar vortex.

The variability of the polar vortex is a prominent example of dynamical wave-mean flow interactions [McIntyre, 1982]. Long planetary waves with wave numbers one or two are induced in the troposphere to then propagate upwards. According to the Charney-Drazin criterion [Charney and

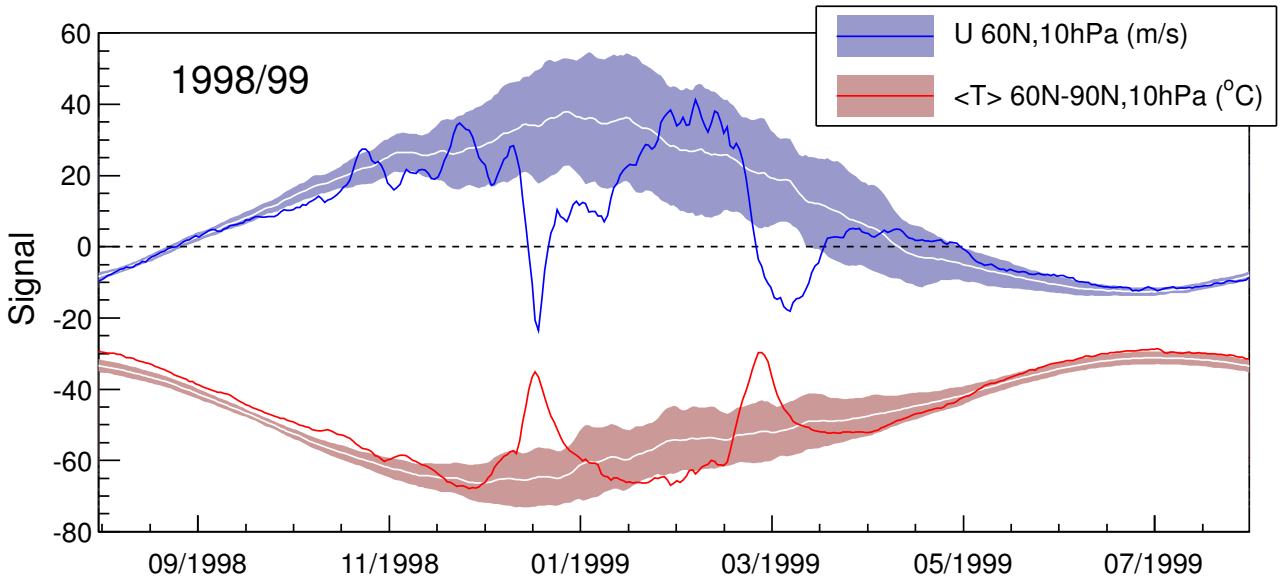

Figure 2.2: The evolution of the zonal mean zonal wind at 60° N, 10 hPa (blue) and the temperature averaged on the polar cap (60° N-90° N) at 10 hPa (red) for the winter 1998/99, taken from the ERA-40 reanalysis. The white line denotes the respective 45-year long-term mean and the shading the corresponding standard deviation. Labeled is the first day of the respective month.

Drazin, 1961], these waves will only reach the stratosphere if the zonal mean flow is westerly (only during winter) and if this flow is not too strong in magnitude. The planetary waves can be induced on, e.g., mountain ranges or blocking situations [Charney and Eliassen, 1949; Labitzke, 1965; Martius et al., 2009; Dörnbrack et al., 2012]. Once these waves reach the upper stratosphere, lower mesosphere they can dissipate and deposit their momentum there which slows down the zonal flow and may even reverse it. If the zonal flow reverses, meaning easterly zonal winds in the zonal mean, the polar vortex is either weakened and largely *displaced* from the pole or *split* into two smaller vortices [e.g., Labitzke and Naujokat, 2000; Charlton and Polvani, 2007]. A reversal of the zonal flow is equivalent to the occurrence of an SSW. Due to the uneven land-sea distribution and the numerous mountain ranges in the northern-hemisphere, planetary waves are constantly induced in the troposphere making a vortex breakdown much more likely in the northern hemisphere than in the southern. There is an SSW taking place every two years on average [Labitzke and Naujokat, 2000] in the northern hemisphere. In contrast, there is only a single SSW in the year 2002 [Varotsos, 2004] on the entire observed record of the southern hemisphere.

To illustrate the dynamics of an SSW in the northern hemisphere, Fig 2.2 shows the progress of the zonal mean zonal wind at 60° N, 10 hPa (blue) and the temperature averaged on the polar cap (60° N-90° N) at 10 hPa (red) for the winter 1998/99, taken from the ERA-40 reanalysis [Uppala et al., 2005]. The white line denotes the respective 45-year long-term mean and the shading the corresponding standard deviation. The zonal flow reversed twice during this winter. First in December and second at the end of February [Labitzke and Naujokat, 2000]. Therefore, two sudden stratospheric warmings took place during the winter 1999/98 which is a very extraordinary situation

since SSWs usually happen only once per winter, if at all. The two SSWs of the winter 1999/98 are naturally accompanied by a strong temperature increase over the polar region as observed in Fig. 2.2. Detecting sudden stratospheric warmings is usually done using the zonal wind from Fig. 2.2. This effective measure was first introduced by the Stratospheric Research Group Berlin [Labitzke and Naujokat, 2000] and incorporated by the World Meteorological Organization (WMO). It was used by Charlton and Polvani [2007] to compile climatologies of sudden stratospheric warmings derived from reanalyses data. However, work made by Blume et al. [2012] and presented in Chapter 6 showed that SSWs can also be classified by only incorporating polar cap temperatures along with external forcings influencing the polar vortex. In fact, another commonly used measure is the meridional temperature gradient at 10 hPa, which is positive (warmer at the pole than at mid-latitudes) during stratospheric warming events. A positive temperature gradient indicates a warming, such as a minor or Canadian Warming, but only the reversal of the zonal mean zonal wind at 60° N indicates a major stratospheric warming. Chapter 6 describes these different events in more detail and presents a novel approach for classifying them.

SSWs are an exciting subject of ongoing research not only because of their impact on stratospheric dynamics and composition but also due to their remarkable influence on the northern hemisphere winter in the troposphere. A large number of extreme vortex events in the stratosphere is shown to propagate downward on time scales of 10 to 20 days [Baldwin and Dunkerton, 2001]. This downward propagation then leads to an increased probability of cold anomalies over north America, Asia, and Europe within 60 days of the SSW central date [Thompson, 2003]. This is likely accompanied by a weakening of the North-Atlantic Oscillation and winter storm tracks that are found more southern than during strong vortex conditions [Baldwin and Dunkerton, 2001]. Since there are at least 10 days needed for anomalies to propagate downward to the surface during weak vortex events, there is predictive skill for the troposphere and near-surface weather. However, forecasting SSWs themselves remains very difficult. When making polar stratospheric forecasts, general circulation model runs are performed that consist of multiple observation constrained ensemble members. These forecasts are reliable on a daily scale but, on a seasonal scale, they quickly become computationally expensive and loose their forecast skill [Gerber et al., 2009; Kuroda, 2010; Dörnbrack et al., 2012].

Another way of approaching seasonal forecasts for the polar vortex, is understanding statistical relationships between external variability factors and the dynamics in the polar stratosphere. The most prominent natural forcing factors include the Quasi-Biennial Oscillation (QBO) [e.g., Baldwin et al., 2001], the El Niño-Southern Oscillation (ENSO) [e.g., Manzini et al., 2006], the 11-year solar cycle [e.g., Gray et al., 2010], and high impact volcanic eruptions [Robock, 2000]. Previous efforts investigated the impact of external factors on the polar vortex. Holton and Tan [1980, 1982] showed that the QBO east phase leads to a generally warmer, more disturbed polar vortex and vice versa for QBO west. This so-called Holton-Tan relationship was later shown to be present during solar minimum but significantly weaker during solar maximum [e.g., Labitzke, 1987; Labitzke and van Loon, 1988]. Labitzke and co-authors showed that sudden stratospheric warmings are most

likely to happen during solar maximum (minimum) and QBO west (east) phase. Accordingly, the work made by Camp and Tung [2007a] found the least-perturbed vortex state to take place during solar minimum and QBO west conditions. Recent studies have shown that positive ENSO phases (El Niño) lead to a more disturbed polar vortex as opposed to negative ENSO phases (La Niña) where the vortex is less disturbed [Camp and Tung, 2007b; Mitchell et al., 2011]. In agreement with previous work, Blume et al. [2012] has found that the QBO and ENSO have the largest impact on the polar vortex, followed by the solar cycle. The least understood forcing factor is aerosols injected into the stratosphere by very strong but rare volcanic eruptions [Robock, 2000] leading to nonlinear feedbacks with other forcings [e.g., Garfinkel and Hartmann, 2007].

The external factors influencing the polar vortex interact with each other, resulting in a complex and nonlinear dynamical response [e.g., Calvo et al., 2009; Richter et al., 2011]. Accordingly, Blume et al. [2012] derived nonlinear patterns important for SSWs, connecting QBO, ENSO, and the solar cycle (see Chapter 6). Blume and Matthes [2012] showed that there is great potential in statistically forecasting extreme vortex events when incorporating all significant factors simultaneously while training a statistical model allowing for nonlinear relationships (see Chapter 7). The natural forcings and modes of variability important for stratospheric processes are introduced in the following.

2.2 Natural Variability Factors

Variability means the significant change of physical properties of a system, such as the stratosphere. This change may happen on many different time scales, from hours to centuries. It may be periodically, a long-term trend, or even just a singular one-time event. Variability in the stratosphere is the result of a complex link between various *forcings* influencing the evolution of the stratosphere. These forcings may be part of the *internal* stratospheric variability (e.g., Quasi-Biennial Oscillation, Northern Annular Mode) or *external* forcings (e.g., solar variability, volcanoes, greenhouse gases, ozone depleting substances). In the following, the most prominent forcings of natural origin are introduced.

2.2.1 The Quasi-Biennial Oscillation

The Quasi-Biennial Oscillation (QBO) is an east-west oscillation in zonal wind in the tropical stratosphere where stronger easterlies and weaker westerlies alternate with an irregular period averaging to approx. 28 months with highest amplitudes in the middle and lower tropical stratosphere [Baldwin et al., 2001]. Figure 2.3 shows observed equatorial zonal mean zonal wind anomalies from 1964 to 1990. The QBO first appears in the upper stratosphere to then propagate downward to the lower stratosphere where it is typically defined at 50hPa ($\approx 21\,\mathrm{km}$). The more chaotic zonal wind regime observed in the upper stratosphere is also influenced by the Semi-Annual Oscillation (SAO) [e.g., Garcia et al., 1997] with a period of approx. 6 months extending well into the mesosphere. The QBO is driven by gravity, inertia-gravity, Kelvin, and Rossby-gravity waves originating from the

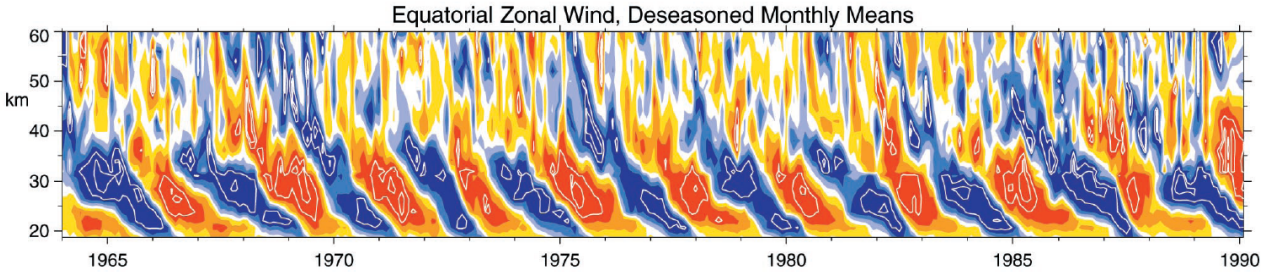

Figure 2.3: Observed monthly equatorial zonal mean zonal wind anomalies in ms^{-1}; figure from Baldwin et al. [2001]. The contour interval is $6\,\mathrm{ms}^{-1}$, with the band between -3 and $3\,\mathrm{ms}^{-1}$ unshaded. Red represents positive (westerly) and blue negative (easterly) winds. The seasonal cycle was removed from these data (deseasoned) making easterly and westerly amplitudes comparable in magnitude.

troposphere and their dissipation in the stratosphere [Holton and Tan, 1982; Baldwin et al., 2001]. Even though the QBO's amplitude decreases very fast when leaving the equator, its influence on the extra-tropical and polar stratosphere along with sudden stratospheric warmings is large [e.g., Holton and Tan, 1980, 1982; Labitzke and van Loon, 1988; Camp and Tung, 2007a]. The north-polar stratospheric vortex is found to be more disturbed and weaker during phases of QBO east whereas QBO west phase lead to a stronger, less disturbed vortex. During QBO east, upward-propagating Rossby waves are more likely to be guided towards the polar vortex to then disturb the westerly flow during winter. This is referred to as the Holton-Tan mechanism [Holton and Tan, 1980, 1982]. There have been also links found that connect the QBO directly with phenomena in the lower troposphere [e.g., Coughlin and Tung, 2001; Thompson, 2003].

In order to quantify the influence of the QBO on extratropical phenomena in the stratosphere, such as polar vortex variability, it is common practice to define time-dependent indices. A common QBO index is the zonal mean zonal wind at 50 hPa averaged between 5° S and 5° N [Holton and Tan, 1980]. However, depending on the specific teleconnection under investigation, a different level may be more appropriate. One efficient way of obtaining a more general representation is computing empirical orthogonal functions (EOFs; see Section 3.1) [Wallace et al., 1993] of the deseasoned zonal mean zonal wind field between 100 and 1 hPa, averaged between 5° S and 5° N. The two leading EOFs represent almost all variance in this region and are often used simultaneously for statistical analysis in the stratosphere. The correlation fields of U (zonal wind) with the first (QBO1) and second EOF (QBO2) are shown in Fig. 2.4 along with their principal components calculated from the ERA-40 reanalysis. Explained variance of each EOF is given in parentheses. It is observed that the QBO1 correlation maximizes at 50 hPa, making its principal component equivalent to the regular QBO index at 50 hPa. The QBO2 correlation maximizes at 20 hPa. QBO2 is referred to as the *orthogonal* QBO because the dot product between QBO1 and QBO2 is zero. Please note that since QBO1 and QBO2 are principal components they have zero mean.

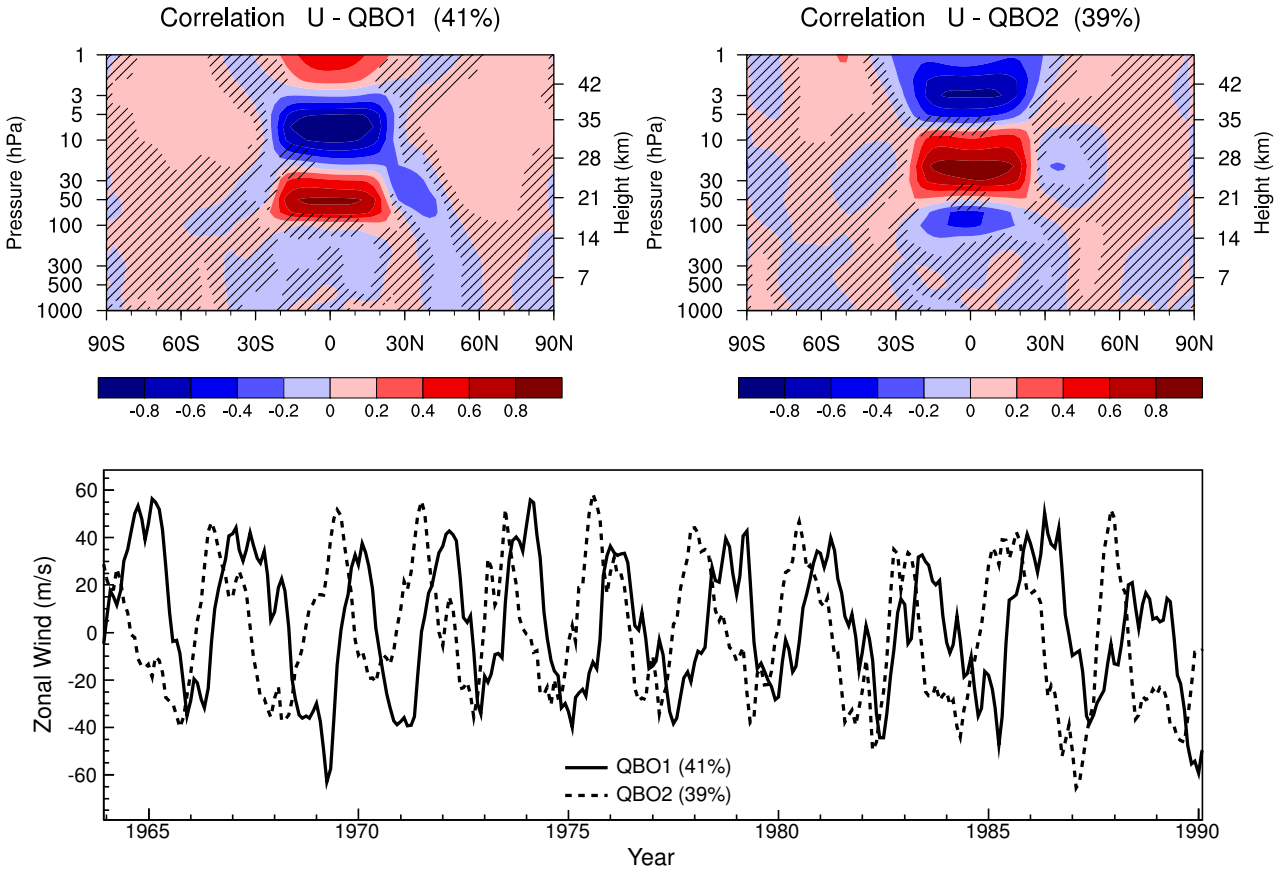

Figure 2.4: *Top*: Correlation fields of U (zonal wind) with the first (QBO1) and second EOF (QBO2) calculated from ERA-40. Areas that are *not* hatched denote correlation coefficients that are statistically significant at the 95% confidence level. Explained variance of each EOF is given in parentheses. *Bottom*: The corresponding principal components for a sample period.

2.2.2 The Annular Modes

The two annular modes are the largest extratropical modes of variability in the stratosphere [Baldwin and Dunkerton, 2001] aside from the seasonal cycle. There is the Northern Annular Mode (NAM), centered around the north pole, and the Southern Annular Mode (SAM), centered around the south pole. A regional manifestation of the NAM in the troposphere is the North-Atlantic Oscillation (NAO) [e.g., Hurrell and Deser, 2009] largely affecting weather in the North-Atlantic region. As the name suggests, NAM and SAM represent ring-like structures around the corresponding pole. They can be imagined as a dipole which redistributes atmospheric mass between the polar and the subtropical regions. For instance, a positive anomaly in geopotential over the north pole leads to a negative anomaly in the subtropics and vice versa, with the zero line somewhere around 60° N. The tropospheric annular modes are thought as to originate from the interaction between baroclinic eddies and the tropospheric mid-latitude jet [Haynes, 2005]. At the same time in the stratosphere, the variability of the polar stratospheric vortex is a result of tropospheric wave forcing and its interaction with the zonal flow [Charney and Drazin, 1961]. However, studies such as Gerber and Vallis [2005] find that annular modes are expected in any momentum and mass conserving rotating

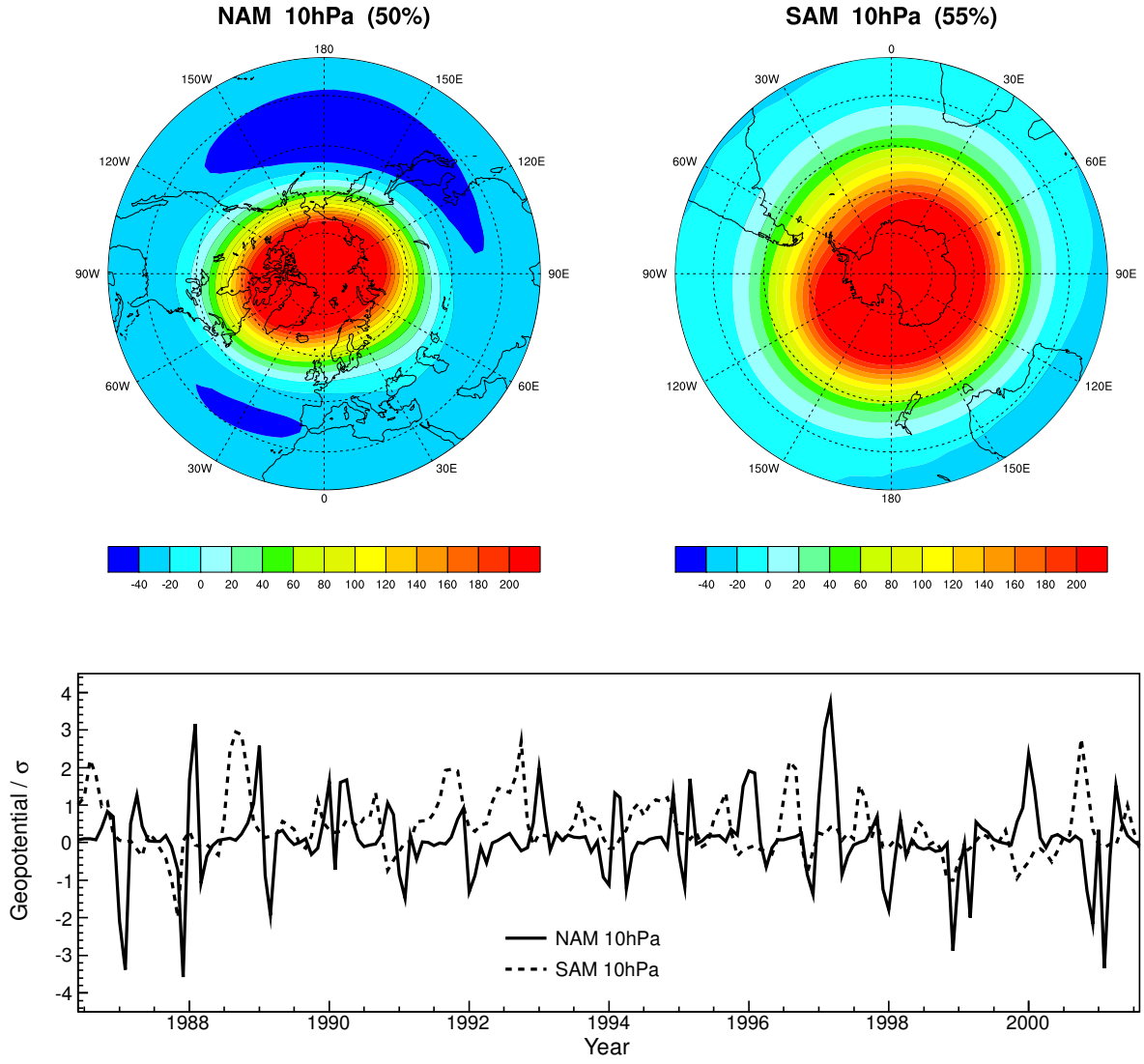

Figure 2.5: *Top*: NAM and SAM patterns in meters at 10 hPa calculated from monthly geopotential anomalies of the ERA-40 reanalysis. Explained variance of each EOF is given in parentheses. *Bottom*: The corresponding principal components in standard deviations for a sample period.

system. The physical causes of the annular modes are still under investigation.

Following Baldwin and Thompson [2009], the NAM can be defined as the leading EOF per pressure level of area-weighted geopotential anomalies from 20° N to 90° N, and accordingly from 20° S to 90° S for the SAM. The result of this procedure is shown in Fig. 2.5 which shows NAM and SAM patterns (the leading EOFs) in geopotential meters at 10 hPa along with the corresponding monthly NAM and SAM indices (principal components). It is observed that the dipole structure between high and middle latitudes is more pronounced on the northern hemisphere. Also, the positive anomaly over the north pole is weaker than that over the south pole reflecting the aforementioned large influence of planetary waves on the northern-hemisphere polar stratosphere (compare Section 2.1). During the respective winter, the north-polar vortex is generally much weaker than the south-polar vortex due to more tropospheric wave disturbances on the northern hemisphere [e.g., Haynes, 2005].

The NAM and SAM indices, as shown in the bottom panel of Fig. 2.5, reflect the strength

of the polar vortex. A positive value indicates a strong vortex whereas a negative value indicates a weak vortex [Baldwin and Dunkerton, 2001]. The daily stratospheric NAM index is commonly used to measure the disturbance of the polar vortex and the occurrence of sudden stratospheric warmings. For instance, the NAM index in Fig. 2.5 (bottom) accurately shows the occurrences of sudden warmings in the late 1980s [Labitzke and Naujokat, 2000] represented by the large negative anomalies (larger than $\approx 2\sigma$ in magnitude) during this time. In contrast, the absence of large negative anomalies during the 1990s represents no occurrences of sudden warmings. Furthermore, the NAM is a powerful way of quantifying coupling between the stratosphere and the troposphere during extreme events in the polar stratosphere. Baldwin and Dunkerton [2001] showed that strong stratospheric anomalies in the polar region can propagate downward to the troposphere within approximately 10 days to then significantly affect tropospheric weather conditions for up to 60 days [Thompson, 2003]. This may be used to improve predicting weather in the troposphere.

2.2.3 Sea Surface Temperatures

The variability in sea surface temperature (SST) and other oceanic parameters play a key role in the development of weather and climate not only in the troposphere, but in the stratosphere as well. In particular, interannual time scales are dominated by phenomena linking ocean and atmosphere. The El-Niño Southern Oscillation (ENSO) [Trenberth, 1997], as the dominant global mode of oceanic variability, is the result of tropical ocean-atmosphere interactions which is therefore observed in oceanic variables such as SST or salinity as well as in atmospheric variables such as air temperature or sea level pressure [e.g., Deser et al., 2010]. ENSO has great implications for other atmospheric phenomena. For instance, it has been shown that ENSO modulates tropical stratospheric ozone and temperature [Randel et al., 2009]. Moreover, it was reported that El-Niño, the warm phase of ENSO, leads to a significantly more disturbed north-polar stratospheric vortex [e.g., Manzini et al., 2006; Camp and Tung, 2007b; Mitchell et al., 2011] than during neutral or negative ENSO phases (La Niña). The resulting vortex anomalies may subsequently propagate downward affecting weather in the troposphere [Cagnazzo and Manzini, 2009].

There are different ways of defining ENSO. Since, in this work, the influence of ENSO on the stratosphere is investigated, the popular Nino3.4 index is used which is reported as having the most significant global teleconnections [Trenberth, 1997]. The Nino3.4 index is the area-weighted average in sea surface temperature anomalies in the box from 170° E to 120° E and from 5° S to 5° N. It is equivalent to the leading EOF of SST anomalies of the mostly ice-free domain between 60° S to 60° N [Deser et al., 2010]. The first four EOFs are shown in Fig. 2.6 along with the corresponding principal components, calculated from observed monthly detrended SST anomalies from the HadISST1 data set [Rayner et al., 2003]. They explain more than 40% of the world-wide SST variance, aside from the seasonal cycle. It is observed that the first EOF corresponds to ENSO with its center of action in the central to western equatorial Pacific. ENSO has direct SST teleconnections mainly in the northern and southern Pacific, but also in the central Indian and Atlantic Ocean. The second EOF

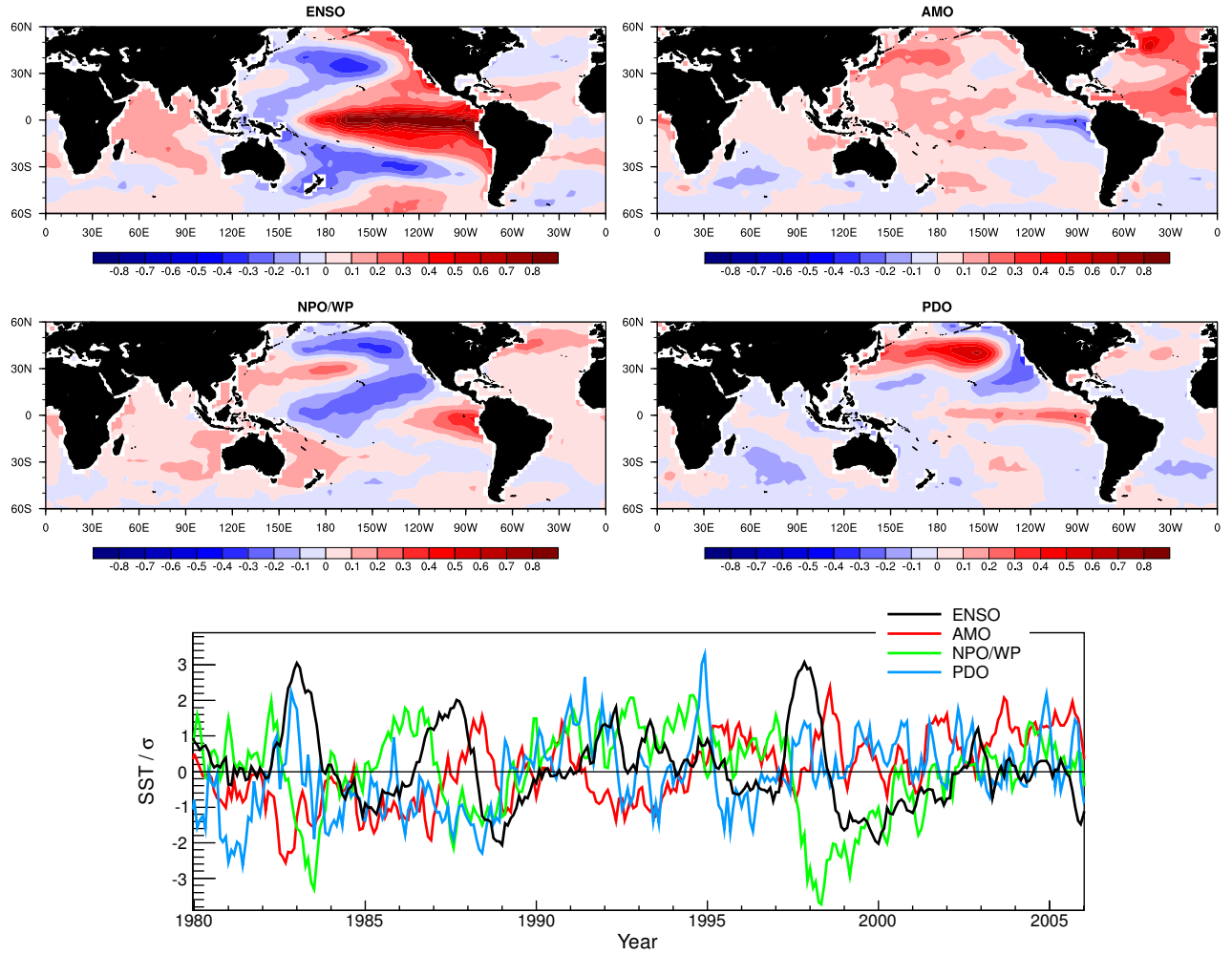

Figure 2.6: *Top*: The first four EOFs in Kelvin of monthly observed detrended SST anomalies between 60° S and 60° N. The first EOF corresponds to ENSO, the second to the AMO, the third to the NPO/WP pattern, and the forth one to the PDO. See text for a description of the different patterns. *Bottom:* The corresponding monthly principal components for a sample period.

reveals the Atlantic Multidecadal Oscillation (AMO) [Schlesinger and Ramankutty, 1994] with its center of action in the North-Atlantic. The AMO has direct SST teleconnections mainly in the northern and central Pacific. The third EOF, as observed in Fig. 2.6, corresponds to the North Pacific Oscillation- West Pacific (NPO/WP) teleconnection pattern [e.g., Linkin and Nigam, 2008] with its center of action in the north-west and central Pacific. The NPO/WP has direct SST teleconnections with the northern Atlantic, the eastern Pacific and the oceanic region surrounding Australia. The fourth EOF clearly reveals the Pacific Decadal Oscillation (PDO) [Bond and Harrison, 2000; MacDonald and Case, 2005] with its center of action in the North-Pacific. The PDO has direct SST teleconnections in the equatorial western Pacific, the southern Indian ocean, and the northern and southern Atlantic ocean. There are other ways to determine indices for these phenomena, e.g., the AMO is more commonly defined as the area-weighted average of SST anomalies over the North Atlantic. However, it is observed that the four main modes of world-wide SST variability are obtained

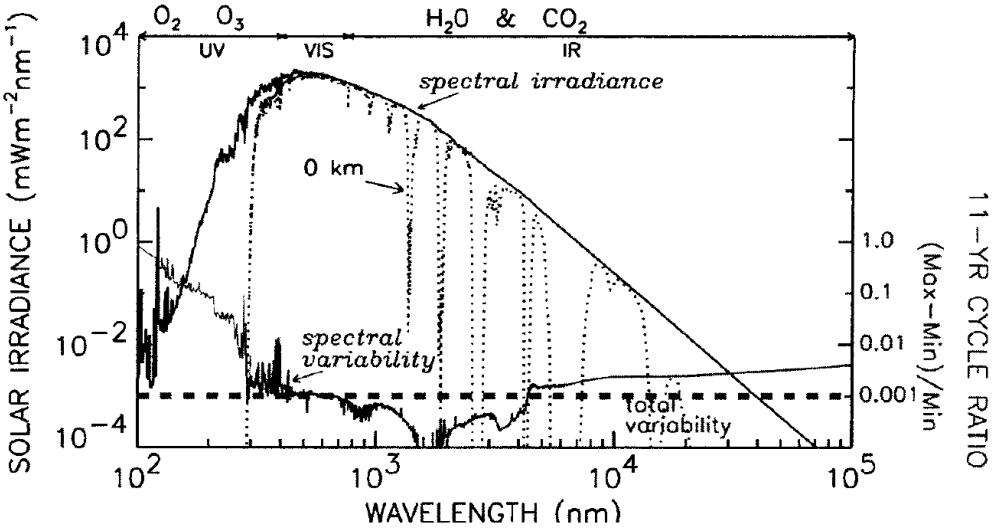

Figure 2.7: The solar irradiance in mW per m^2 and wavelength at the top of the atmosphere (solid spectral irradiance) and at the Earth's surface (dashed spectral irradiance). The spectral variability along with the total variability ($\approx 1\,\mathrm{W/m^2}$) is also shown. Figure from Lean [2000].

consistently with an EOF analysis, making it a robust representation and easy to reproduce.

2.2.4 The 11-Year Solar Cycle

Solar heating is essential for the Earth's weather and climate system. In a good approximation, the sun can be considered a black body emitter with a temperature of 5900 K [Stix, 2004]. The integrated or total solar irradiance (TSI) at the top of the atmosphere is $1366\,\mathrm{W/m^2}$ [Lean, 2000] which is referred to as the solar constant. The TSI is, however, not constant and varies on different timescales from days to centuries [Gray et al., 2010]. The most important mode of solar variation for recent weather and climate evolution is the 11-year solar cycle which is an irregular oscillation in solar irradiance averaging to a period of approx. 11 years. When the solar cycle is referenced in this work, the 11-year solar cycle is meant. The variation in solar activity results in broad implications for processes in the stratosphere, in particular due to changes in the ultra-violet band. There is a significant solar influence, observable mainly in upper and lower equatorial stratospheric levels [e.g., Gray et al., 2010]. This solar signal can be observed in different variables such as ozone, temperature, or wind. In combination with other variability factors, the solar cycle influences the polar vortex and the occurrence of sudden stratospheric warmings [e.g., Labitzke and van Loon, 1988; Camp and Tung, 2007a; Gray et al., 2010]. A possible mechanism for this was proposed by Kodera and Kuroda [2002] which states that an increased heating around the equatorial stratopause during solar maximum leads to a greater temperature gradient between tropics and pole during early winter. This gradient, according to the thermal wind equation, leads to an increased vertical gradient of zonal winds in mid-latitudes and a strengthening of the polar vortex and therefore less interactions between the westerly mean-flow and planetary waves. This proposed mechanism was later verified

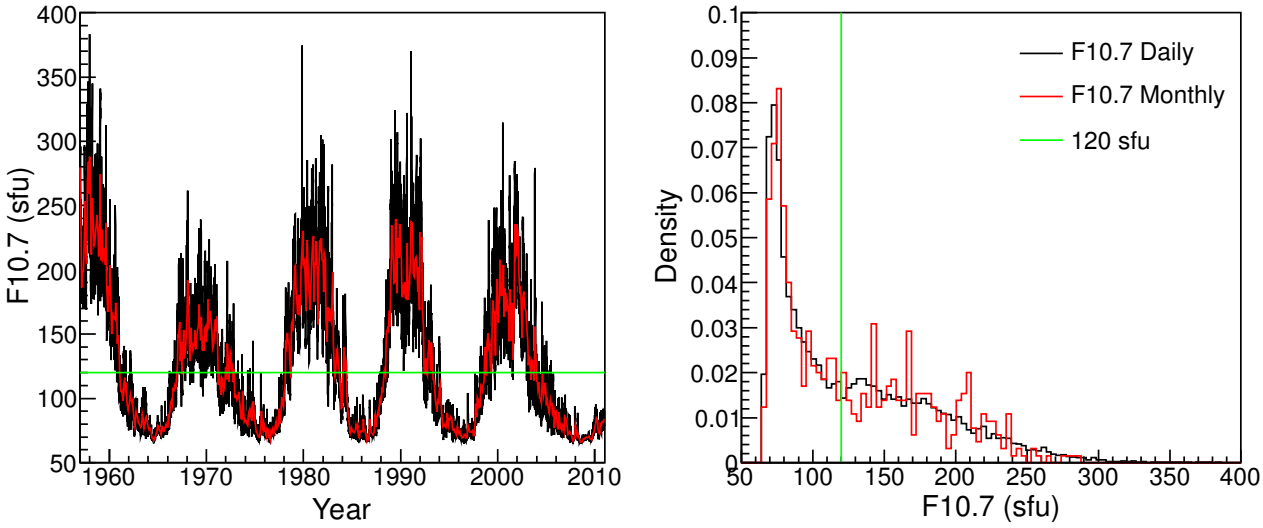

Figure 2.8: *Left:* Daily (black) and monthly (red) F10.7 for the period from 1957 to 2011. *Right:* The corresponding normalized frequency distributions along with the 120 sfu threshold separating solar minimum from solar maximum conditions.

by general circulation model simulations [Matthes et al., 2004]. There are also significant solar impacts reported in the literature concerning, e.g., the QBO [Salby and Callaghan, 2006], tropical SSTs [White and Liu, 2008], and the North-Atlantic Oscillation [Kodera, 2002]. It has been shown that there is a robust solar cycle signal all throughout tropospheric levels [Coughlin and Tung, 2004].

The TSI changes by about $1\,\mathrm{W/m^2}$ ($\approx 0.1\%$) from solar minimum to solar maximum during the solar cycle [Lean, 2000]. This change in solar irradiance is not distributed evenly across the different wavelengths. Fig. 2.7 shows the spectral irradiance along with the spectral variability from 10^2 to 10^5 nm. The spectral variability is much higher in the ultra-violet (UV) range towards smaller wavelengths than in the rest of the spectrum. The UV range is of particular importance for the stratosphere because ozone (O_3) mainly absorbs radiation in this band. Therefore, stratospheric variability in ozone and temperature will mostly be affected by short-wave changes in the solar spectrum [Haigh, 1994, 1996]. As shown in Fig. 2.7, also oxygen (O_2) absorbs in the UV band. Most absorption in the infra-red is done by water vapor (H_2O) and carbon dioxide (CO_2) making them, along with methane, important greenhouse gases. It is observed that only solar radiation in the visible range does not experience a significant absorption in the Earth's atmosphere and can penetrate to the surface.

There are different proxies being used in the atmospheric science measuring the solar variability. In this work, a proxy is used that represents the spectral variability in the UV. It was observed that spectral changes are greatest in the UV and that ozone mainly absorbs in this band (see Fig. 2.7). The utilized proxy is the solar radio flux at 10.7 cm [e.g., Gray et al., 2010], called the F10.7 measured in solar flux units ($1\ \mathrm{sfu} = 10^{-22}\ \mathrm{Wm^{-2}Hz^{-1}}$). The F10.7 data are taken at the observatory in Penticton, British Columbia, Canada and are available at the National Oceanic and Atmospheric Administration (NOAA) website under `ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/`

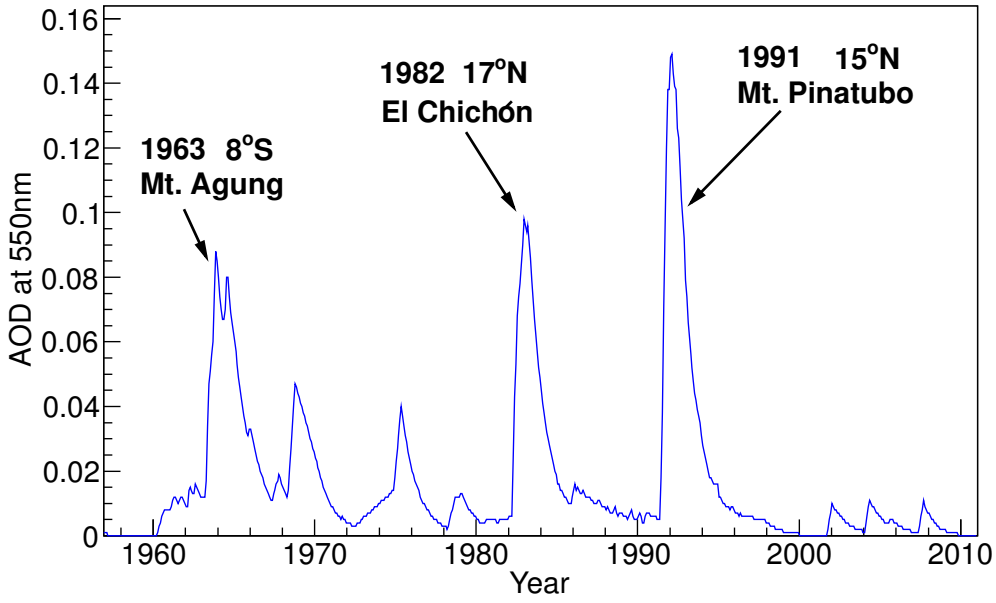

Figure 2.9: The globally averaged aerosol optical depth (AOD) at a wavelength of 550 nm along with the globally most significant volcanic eruptions including year of eruption and corresponding latitude [Sato et al., 1993].

SOLAR_RADIO/FLUX/Penticton_Observed. The F10.7 is highly correlated to other measures of solar variability such as the TSI or the sunspot number [e.g. Gray et al., 2010]. Fig. 2.8 shows the daily and monthly F10.7 for the period of 1957 to 2011 (approx. 4.5 cycles) along with the corresponding normalized frequency distributions. The time series reveal the approximately decadal period as mentioned earlier. In this work, a value of 120 sfu is used to distinguish between solar minimum and maximum conditions as denoted in Fig. 2.8. This threshold divides the frequency distributions in two approximately equally sized parts and is therefore equivalent to the median of the distribution. Other works, such as Labitzke et al. [2006], have considered 110 sfu and 150 sfu for thresholds indicating solar minimum and maximum, respectively. This, however, does not take the solar density distribution (see right panel of Fig. 2.8) into account and potentially neglects events between the two thresholds.

2.2.5 Volcanic Eruptions

Volcanic eruptions are one of the least understood factors influencing the climate system [IPCC, 2007]. To have an effect on the stratosphere in particular, volcanic eruptions have to be powerful enough to inject aerosols well into the stratosphere. This is more likely when the eruption takes place in the tropics since vertical aerosol transport becomes more likely with increased convection [Robock, 2000]. Aerosols from strong tropical eruptions enter the stratosphere and are dynamically spread to higher latitudes and around the globe. In the past 60 years, there have been only three volcanic eruptions that injected significant amounts of aerosols into the stratosphere. Only the eruptions of

Mt. Agung (1963), El Chichón (1982), and Mt. Pinatubo (1991) were powerful enough to have an influence on stratospheric dynamics and composition [Robock, 2000]. Fig. 2.9 shows the globally averaged aerosol optical depth (AOD) at a wavelength of 550 nm [Sato et al., 1993] received from the NASA Goddard Institute for Space Studies under `http://data.giss.nasa.gov/modelforce/strataer`. AOD represents the optical thickness of air at a wavelength where transmission of light is strongly influenced by volcanic aerosols due to absorption and scattering. The aerosol optical depth is the extinction coefficient integrated over a vertical column. Therefore, the globally averaged AOD can be used as a proxy of aerosol density. Fig. 2.9 clearly depicts the three large eruptions. In addition, two smaller ones are charted in the late 1960s and the 1970s. Large volcanic eruptions have significant effects on the Earth's climate. It has been observed that volcanic eruptions lead to global surface cooling by up to 0.1-0.2 K which may last up to 2 or 3 years after the eruption [Robock and Mao, 1995]. On the other hand, there is increased heating observed in the stratosphere [Angell, 1997] due to enhanced absorption in the infra-red. Moreover, depletion of stratospheric ozone is observed following powerful volcanic eruptions [Solomon, 1999].

2.3 Temperature and Ozone Trends

There are significant long-term trends in stratospheric temperature and ozone, observed for the recent climate ($\sim$1960 to present) and simulated for the future climate (present to $\sim$2100). Understanding the causes and evolution of these trends is crucial for understanding stratospheric variability as a whole.

2.3.1 Recent Climate

The recent climate in the stratosphere is governed by a decrease in temperature [Randel et al., 2009]. Fig. 2.10 shows the temperature trends from 1958 to 2007 from the surface to the middle stratosphere averaged for three different latitude ranges and different observational data sets. There is a robust temperature trend in the middle stratosphere of around -0.5 K/decade adding up to a total temperature decrease of -2.5 K over the last 50 years, being almost independent of the considered latitude range. This *decrease* is up to a factor of three stronger in magnitude than the temperature *increase* observed in tropospheric levels [IPCC, 2007] as observed in Fig. 2.10. The stratospheric temperature decrease is considered to be mainly caused by two factors: 1) The increase in tropospheric greenhouse gases (mainly CO_2) causing less back-radiation in the infra-red to the stratosphere; and 2) The increase of ozone depleting substances (mainly chlorine and bromine) leading to less ozone and therefore less absorption of solar radiation in the ultra-violet resulting in less warming. The observed stratospheric temperature trends have also been confirmed by most CCMs during the SPARC CCMVal-2 initiative [SPARC CCMVal, 2010]. The trend in the upper stratosphere (not shown) is even found to be slightly stronger than the trend in lower levels mainly driven by the destruction of ozone.

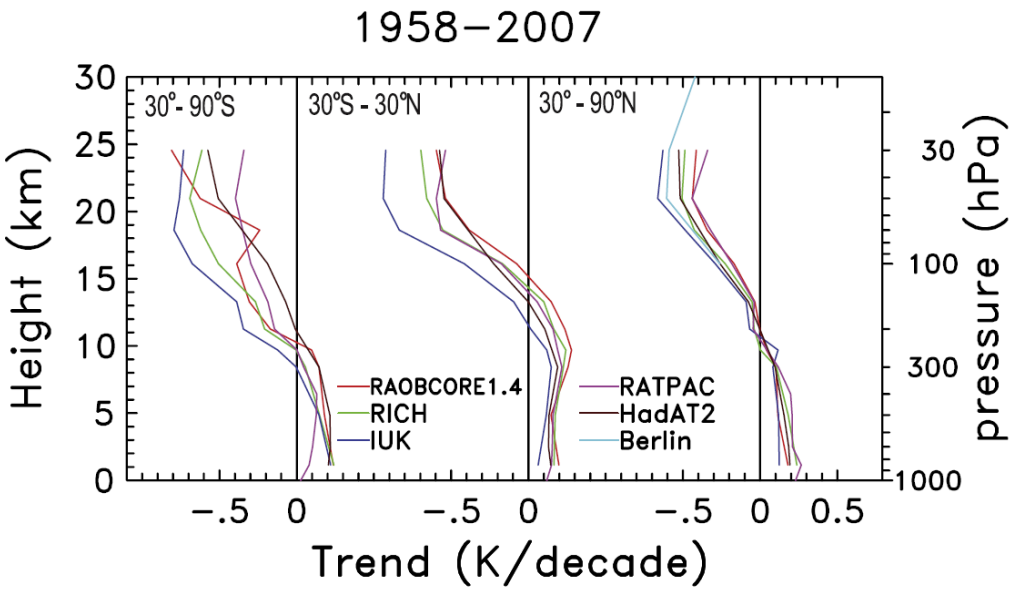

Figure 2.10: Temperature trends from 1958 to 2007 of the annual mean temperature from the surface to the middle stratosphere averaged for three different latitude ranges and different observational data sets. Figure from Randel et al. [2009].

Fig. 2.11 shows observed stratospheric ozone trends from 1979 to 2004, estimated from a number of observational instruments. The trends are shown for the northern and the southern mid-latitudes. It is observed that there is quite some uncertainty in ozone estimation in the stratosphere. Nevertheless, significant trends are observed in the lower stratosphere with up to 4% and in the upper stratosphere with up to 8%. Within the uncertainty, the observed ozone trends in mid-latitudes are similar for northern and southern hemisphere. The ozone trends in the tropics are slightly smaller but retain the same height-dependent structure.

The trends in ozone are unevenly distributed across the latitudes. The largest trends observed for the recent climate are found over Antarctica (∼25%) followed by the Arctic region (∼10%) [WMO, 2007]. These trends have a strong seasonal variation because ozone depletion is favored during times of low temperature. When stratospheric temperatures drop below a certain threshold (approx. -78° C) during winter then polar stratospheric clouds can form which further increase the destruction of ozone [Solomon, 1999]. Hence, ozone loss is highest in the winter hemisphere, especially in the Antarctic region due to its very low temperatures and stable polar vortex [Andrews et al., 1987]. Both the depletion of ozone and the observed ozone trends are highest during spring time. During this time of the year the well-known *ozone hole* appears over Antarctica which denotes a region of very low ozone. Minimum spring-time column ozone has dropped by more than 50% with respect to values of 1980 over Antarctica [WMO, 2007] and by approx. 30% over the Arctic. In addition, the ozone density is highest over the poles [Solomon, 1999]. Ozone depletion implies cooling in the stratosphere. It was reported by Thompson and Solomon [2002] that ozone destruction leads to a strengthening of the polar vortex. The recent climate change observed in the southern hemisphere,

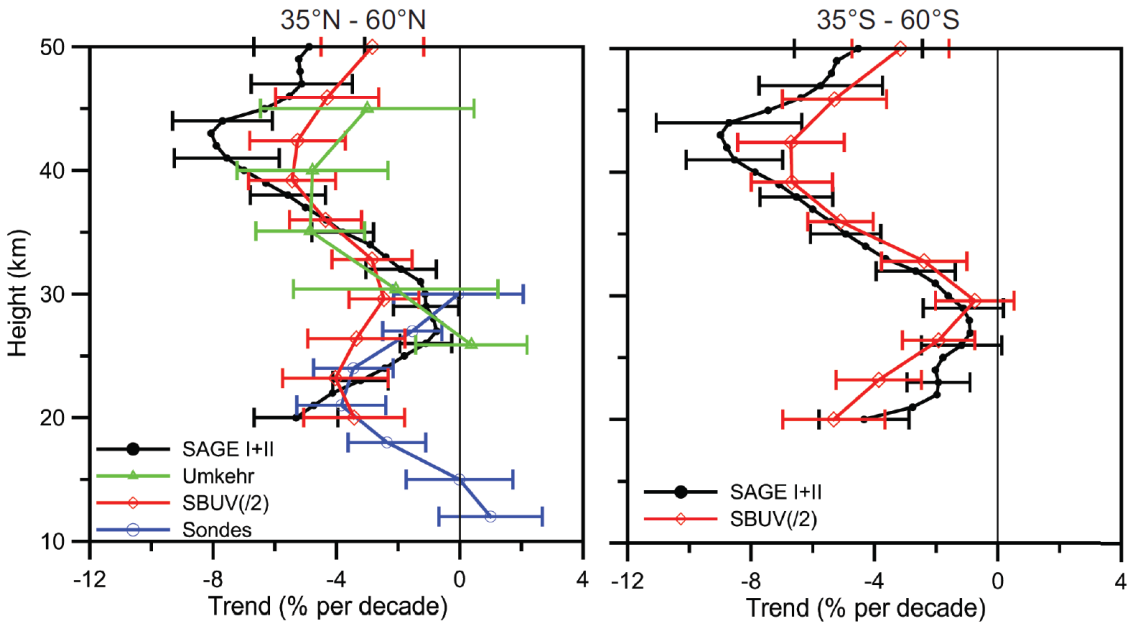

Figure 2.11: Vertical profile of ozone trends in percent with respect to a climatology over the northern (left) and the southern (right) mid-latitudes, estimated from a number of observational instruments for the period of 1979 to 2004. The horizontal bars denote a 95% confidence interval. Figure from WMO [2007].

most evident in near-surface temperature over Antarctica, is attributed to ozone depletion in the stratosphere [e.g., Thompson and Solomon, 2002].

Responsible for the observed ozone depletion are mainly long-lived chlorofluorocarbons (CFCs) released by man into the atmosphere. The most reactive CFCs are chlorine (Cl) and bromine (Br) which effectively destroy ozone in several catalytic cycles with the help of oxygen and solar radiation [Solomon, 1999]. From these two elements an index called *Equivalent Stratospheric Chlorine* (ESC) can be defined as $\mathrm{ESC} = \mathrm{Cl} + 60 \cdot \mathrm{Br}$ [Eyring et al., 2007], reflecting the fact that bromine is 60 times more reactive than chlorine. ESC naturally reflects the observed and simulated global trends in ozone. It maximized somewhere between the years 1995 to 2000 and slowly decreases since. Accordingly, first significant ozone recovery could be observed in recent years over the Arctic and Antarctic regions [Randel and Thompson, 2011]. The decrease in ESC is attributed to the restriction on the emissions of CFCs established by the Montreal Protocol in 1987. However, CFCs will still continue destroying ozone on a high level due to their long residence time. Hence, understanding the evolution of the recent climate is crucial for making reasonable projections into the future.

2.3.2 Future Projections

During CCMVal-2, future projections until 2100 were carried out using CCM simulations performed by several independent modeling groups. Fig. 2.12 shows the simulated evolution of global column ozone for the time of 1960 to 2100 along with the evolution of ESC, normalized to the value of 1960.

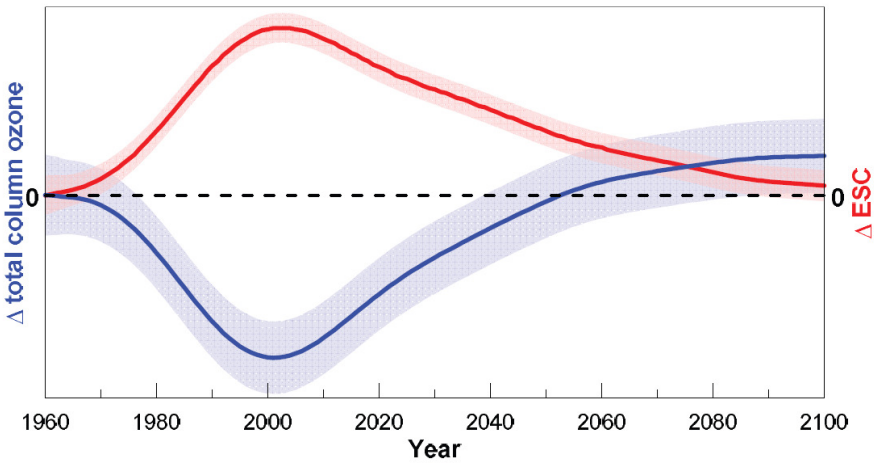

Figure 2.12: The simulated evolution of global column ozone for the time of 1960 to 2100 along with the evolution of ESC (red), normalized to the value of 1960. The projections represent a mean across the different CCMVal-2 future simulations (blue), along with a 95% confidence interval. Figure from WMO [2011].

The projections represent a mean across the different CCM simulations. Ozone change is evidently governed by the change in ESC. Ozone will return to values of 1960 by around 2050. Towards the end of the 21st century, ozone is expected to over recover to values significantly larger than those of 1960 due to increased heating and convection in the tropics caused by the positive temperature trend in the troposphere WMO [2011]. It is speculated that there will be an acceleration of meridional transport processes in the stratosphere further favoring the increase in extra-tropical ozone [Eyring et al., 2007].

Fig. 2.13 shows trends in K/decade of global, annual mean temperature from the upper troposphere to the upper stratosphere from different and independent CCMVal-2 simulations for the first and the second half of the 21st century. The model results agree relatively well and are able to reproduce the observed recent climate trends (see Fig. 2.10). Further, it is observed that the negative stratospheric temperature trend persists into the 21st century, maximizing in the upper stratosphere with about -0.8 K/decade. As previously indicated, the global stratospheric temperature trend is a result of a superposition of the effects of ozone depleting substances and greenhouse gases such as carbon dioxide and methane, resulting in an almost linear stratospheric trend from 1960 to 2100 [Stolarski et al., 2010]. However, in the upper troposphere the temperature trend remains positive with about 0.4 K/decade throughout the 21st century.

2.4 Statistical Analysis

Modeling the atmosphere using climate models is physically self-consistent and effective. During the SPARC CCMVal initiative [SPARC CCMVal, 2010], it was shown that most CCMs are able to *realistically* reproduce behavior and processes in the troposphere and the stratosphere. In order to

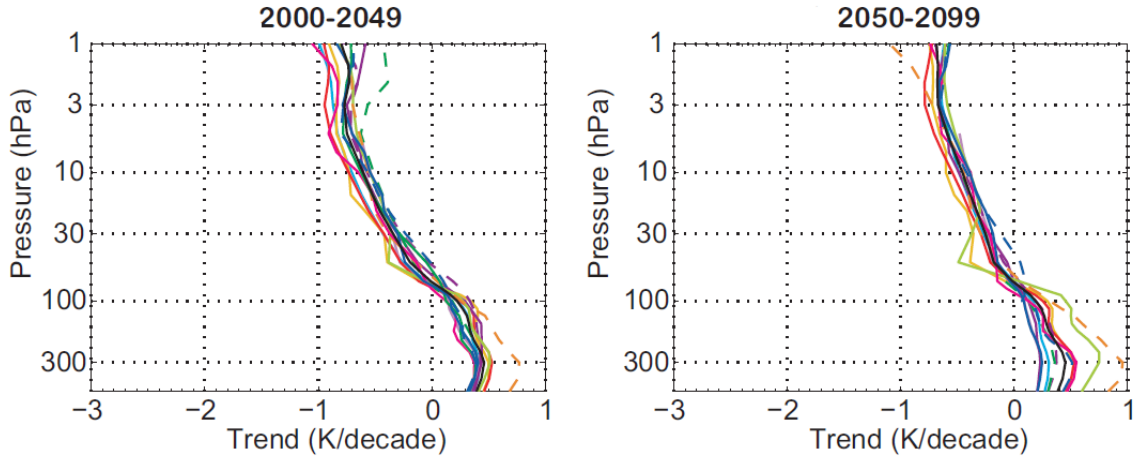

Figure 2.13: Trends in K/decade of global, annual mean temperature from the middle troposphere to the upper stratosphere from different and independent simulation runs for the first and the second half of the 21st century. Figure from SPARC CCMVal [2010].

measure what is realistic, statistical analysis is needed. By using statistical techniques, information about a system can be synthesized and significant information is extracted. Moreover, the importance of variability factors on processes in the atmosphere is commonly attributed using statistical methods. Forecasting, down-scaling, and even data assimilation is also envisaged with statistical models in the atmospheric science community. Statistical analysis in the atmospheric science has become an important and broad field of research [von Storch and Zwiers, 2001]. The introduction of fast processing units along with the availability of free and efficient compilers and statistical libraries, constantly pushes the limit of sophisticated methods in this field.

Different statistical methods are applied in the atmospheric sciences [Wilks, 1995]. It is generally distinguished between *descriptive* and *inferential* statistics. Descriptive statistics aims at finding key measures describing a set of data. This could be something like a simple sum, a correlation coefficient, or a tailor diagram. In contrast, inferential statistics aims at learning inherent patterns from the population that the data is supposed to represent [Jain et al., 2000]. Inferential statistical methods are, e.g., regression analysis [Montgomery et al., 2006], or data clustering [Jain et al., 1999]. A complete statistical analysis will generally make use of both, methods of descriptive *and* inferential statistics. There are two main branches of inference: Frequentist and Bayesian inference. The frequentist [Wilks, 1995] approach aims at drawing repeated samples from a population to infer a statistical model. In contrast, the Bayesian approach [Jaynes, 2003] infers a model by finding a prior degree of belief denoting a probability density. Frequentist methods are usually simpler and more straight forward, which could be the reason for their popularity. As a result, Bayesian methods are not often used in the atmospheric sciences. This work follows the frequentist approach. A number of statistical methods inferring statistical models for pattern recognition are reviewed in the next chapter.

An important statistical method for atmospheric science is the principal component analysis

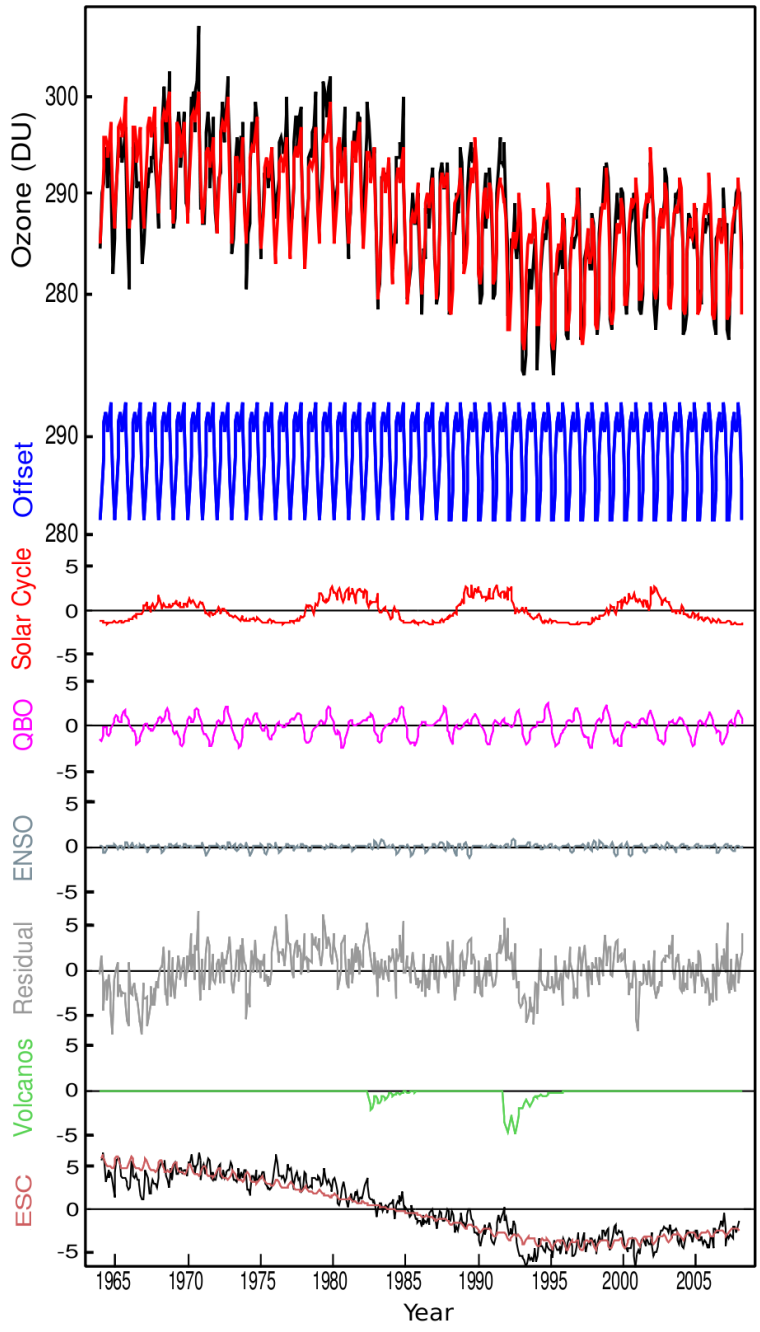

Figure 2.14: The result of a multiple linear regression trained with monthly observed quasi-global (60° S-60° N) column ozone data, subject to several external factors. *From top to bottom*: Observed column ozone (black) and regression result (red); seasonal cycle (blue); solar cycle (red); QBO (purple); ENSO (light blue); Residual (gray); Volcanoes (green); ESC (red). Figure from Chapter 8 of SPARC CCMVal [2010].

(PCA) [Jolliffe, 2002], also called empirical orthogonal function (EOF) analysis. PCA can generally be applied to a variety of problems, specifically when prominent variabilities dominate the system and/or dimension reduction is needed. Another important method is that of multiple linear regression [Montgomery et al., 2006] which is the most commonly used regression technique in stratospheric

research [e.g., Bodeker et al., 1998; Crooks and Gray, 2005; Randel et al., 2009; Gray et al., 2010; SPARC CCMVal, 2010]. To illustrate its use, Fig. 2.14 shows the result of a linear model (red) trained with observed quasi-global (60° S-60° N) column ozone data (black) being subject to external factors (also called features or basis functions), such as the offset (blue), the solar cycle (red), the QBO (purple), or equivalent stratospheric chlorine (ESC). This linear statistical model is able to quite effectively approximate the observed data. The amplitude of each of the external factors indicates its relative importance for the regression result. As observed, most variability is attributed to ESC, the offset (mainly the annual cycle), the solar cycle, and the QBO. However, the residual (gray) is of the size of most external factors or even larger. Hence, more advanced, e.g., nonlinear or nonstationary techniques could be used to improve the modeling.

Many more advanced regression, classification, and cluster techniques are utilized in the atmospheric community. For instance, Walter and Schönwiese [2002] addressed the detection and attribution of observed global climate change in global temperature anomalies using a neural network; Lu et al. [2009] used a nonlinear neural network to receive an alternative representation of the QBO; Coughlin and Gray [2009] use the K-Means cluster algorithm to determine two distinct states in the polar stratosphere, one representing normal, the other weak vortex conditions; Nao et al. [2006] presents a technique based on support vector machines to estimate the surface area of polar stratospheric clouds; and Franzke et al. [2009] used a recently introduced cluster technique (FEM-VARX) based on finite elements to identify large-scale dynamical circulation patterns in GCM simulations. However, to my knowledge no systematic comparison of a range of statistical methods exists, applied to stratospheric variability. A wide class of supervised and unsupervised statistical methods is reviewed in the following. These methods are partly nonlinear and nonstationary making them appropriate for modeling and forecasting the complex variability in the stratosphere.

Chapter 3

Statistical Learning Approaches

Learning refers to the process of obtaining and synthesizing information from a given data set. Once sufficient patterns in a given system have been learned, unknown patterns can be recognized. This process is referred to as statistical pattern recognition [Ripley, 1996]. A statistical model learns these patterns and is, ideally, able to predict information. It is said that a statistical model is *inferred* from the data it is trained with, the training sample. In this work, by the term *statistical model* one incarnation, i.e. one set of parameters that uniquely determine the model is meant.

The statistical methods considered in this work can be grouped into supervised and unsupervised learning procedures [Marques de Sá, 2001]. A schematic picture showing the idea behind supervised and unsupervised learning is shown in Fig. 3.1. Supervised learning is the task of inferring a function (a trained model) from a known training sample consisting of fixed pairs of input and target objects. An object, in the sense of statistical learning, can be any mathematical structure representing characteristics of the system to be analyzed. In this work, an object will usually be a time series which is a sequence of multi-dimensional data depending on time. Input objects are called features and target objects are called truths. Later in this work, features will also be referred to as external factors or forcings. When learning in unsupervised mode, there are only features given to the learning approach and the response is computed by the underlying algorithm. Hence, there is no truth the response can be directly compared with and knowledge about the system must come into play. Please note the difference between target and response: Target is what the statistical model is trained with in supervised mode. Response is what the statistical model "answers" when presented with features, after the model has been trained.

This chapter describes statistical methods that can be used for regression, classification, and clustering. Regression is the task of finding a function mapping from the features to a real-valued vector. Classification is the task of finding a function mapping from the features to a discrete vector holding class labels. This function is called a classifier. Regression and classification are used in supervised mode. Clustering, on the other hand, assigns class labels to certain combinations of features, but there is no truth the statistical model is trained with. Hence, clustering is done in unsupervised learning mode.

Depending on the statistical approach and underlying algorithm, different types of patterns

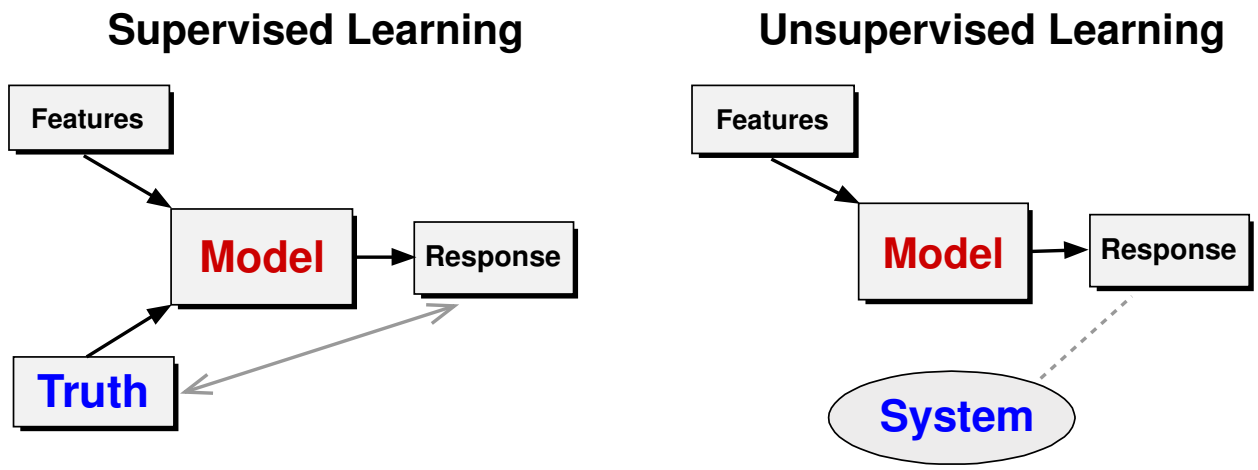

Figure 3.1: A schematic picture showing the principles of supervised and unsupervised learning. In the atmospheric science, features are often referred to as external factors. Truth denotes some target data set and response the model output. In supervised learning, truth and response can be compared to measure the model performance. In unsupervised learning, the model performance can be assessed with knowledge from the considered system.

can be recognized. In general, it is distinguished between methods that learn and recognize linear patterns (see Section 3.2, 3.3, 3.5) and methods that can learn and recognize nonlinear patterns (see Section 3.4, 3.5) in terms of the relationships between features and response. This work introduces both types of pattern recognition algorithms. Most time series models make the assumption that underlying processes that govern the evolution of time series are stationary. This means they do not change their characteristics significantly over time. However, this assumption is often not met in real world situations, especially not in atmospheric science. Therefore, a recently developed multivariate statistical method (FEM-VARX) is used which treats the underlying, generally unknown process as nonstationary (see Section 3.3).

3.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) [Jolliffe, 2002; Hannachi et al., 2007] aims at finding orthogonal modes representing large variability structures in the underlying data. These modes, called empirical orthogonal functions (EOFs), form a complete orthogonal basis in which the first EOF represents the mode of largest variance, the second EOF the mode of second largest variance, and so on. PCA is an unsupervised statistical method clustering a data matrix $\mathbf{X} \in \mathbb{R}^{m \times k}$ where m denotes the number of observations and k the number of variables (dimensions). In atmospheric science, k usually denotes the number of grid points and m the number of steps in time. In the following, it is assumed that none of the column vectors in $\mathbf{X}$ have zero variance.

PCA derives an orthogonal basis $\mathbf{U} \in \mathbb{R}^{k \times k}$ to which $\mathbf{X}$ can be transformed (rotated) so that

$$\mathbf{P} = \mathbf{X} \cdot \mathbf{U} \ \in \mathbb{R}^{m \times k} \, , \tag{3.1}$$

is also orthogonal. The EOFs are the columns of $\mathbf{U}$ and are given by the eigenvectors of the covariance matrix of $\mathbf{X}$. The columns of $\mathbf{P}$ are called the principal components. The data matrix $\mathbf{X}$ consisting of columns that are possibly correlated with each other, is now transformed into a matrix $\mathbf{P}$ consisting of completely uncorrelated columns. In this work, matrix $\mathbf{U}$ is computed using the covariance method. Other methods to compute $\mathbf{U}$ are the singular value decomposition [Jolliffe, 2002] of $\mathbf{X}$ and, especially fast for large k and a small number of desired EOFs, the iterative Lanczos method [Cullum and Willoughby, 2002]. The advantage of the covariance method is that it can be solved using almost any statistical library. In addition, it is possible to compute only a maximum number of EOFs. Inspite of computing only a few EOFs, the sum across all eigenvalues is given through the trace of the covariance matrix.

The covariance matrix of $\mathbf{X}$ is given as $\mathbf{C} = \frac{1}{m}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{k \times k}$ where the columns of $\mathbf{X}$ are mean centered. For large enough m, as usually the case in atmospheric science, $\frac{1}{m}$ approaches the unbiased estimator $\frac{1}{m-1}$. The eigenproblem that determines eigenvectors and eigenvalues can now be formulated by

$$\mathbf{C} \cdot \mathbf{U} = \mathbf{U} \cdot \mathbf{\Lambda} \, , \tag{3.2}$$

where $\mathbf{U}$ denotes the eigenvectors of $\mathbf{C}$ (EOFs of $\mathbf{X}$) and $\mathbf{\Lambda} \in \mathbb{R}^{k \times k}$ the diagonal matrix of the eigenvalues λ_i of $\mathbf{C}$. All λ_i are real and larger than zero since $\mathbf{C}$ is symmetric and positive-definite. Because of large k, the eigenproblem in Eq. 3.2 must be solved numerically. As $\mathbf{C}$ is a quadratic matrix, Eq. 3.2 is solved using the efficient Multishift QR algorithm [Braman et al., 2002] making use of the symmetric structure of $\mathbf{C}$. After solving Eq. 3.2, the eigenvalues are arranged in descending order and the eigenvectors accordingly. It should be noted that PCA makes three major assumptions: a) Linearity imposed by the transformations in Eq. 3.1 and 3.2; b) The columns in $\mathbf{X}$ are jointly distributed according to a Gaussian distribution; and c) Modes representing large variability structures are orthogonal to each other.

The sum of all eigenvalues $E = \sum_{i=1}^{k} \lambda_i$ is called the energy of the eigenvalues which is equal to the trace of $\mathbf{C}$. The variance explained by the i-th EOF is then given by λ_i / E. PCA is mainly applied with two different goals in mind: 1) Finding a small number of high-variance EOFs that help understanding the underlying data and 2) Dimension reduction by retaining only a certain number of EOFs.

When used to understand the underlying data, it needs to be determined if it is possible to statistically distinguish between neighboring EOFs. A rule of thumb was proposed by North et al. [1982] stating that an eigenvalue λ_i possesses an approximate sampling error of

$$\Delta\lambda_i = \lambda_i \cdot \sqrt{\frac{2}{m}}. \tag{3.3}$$

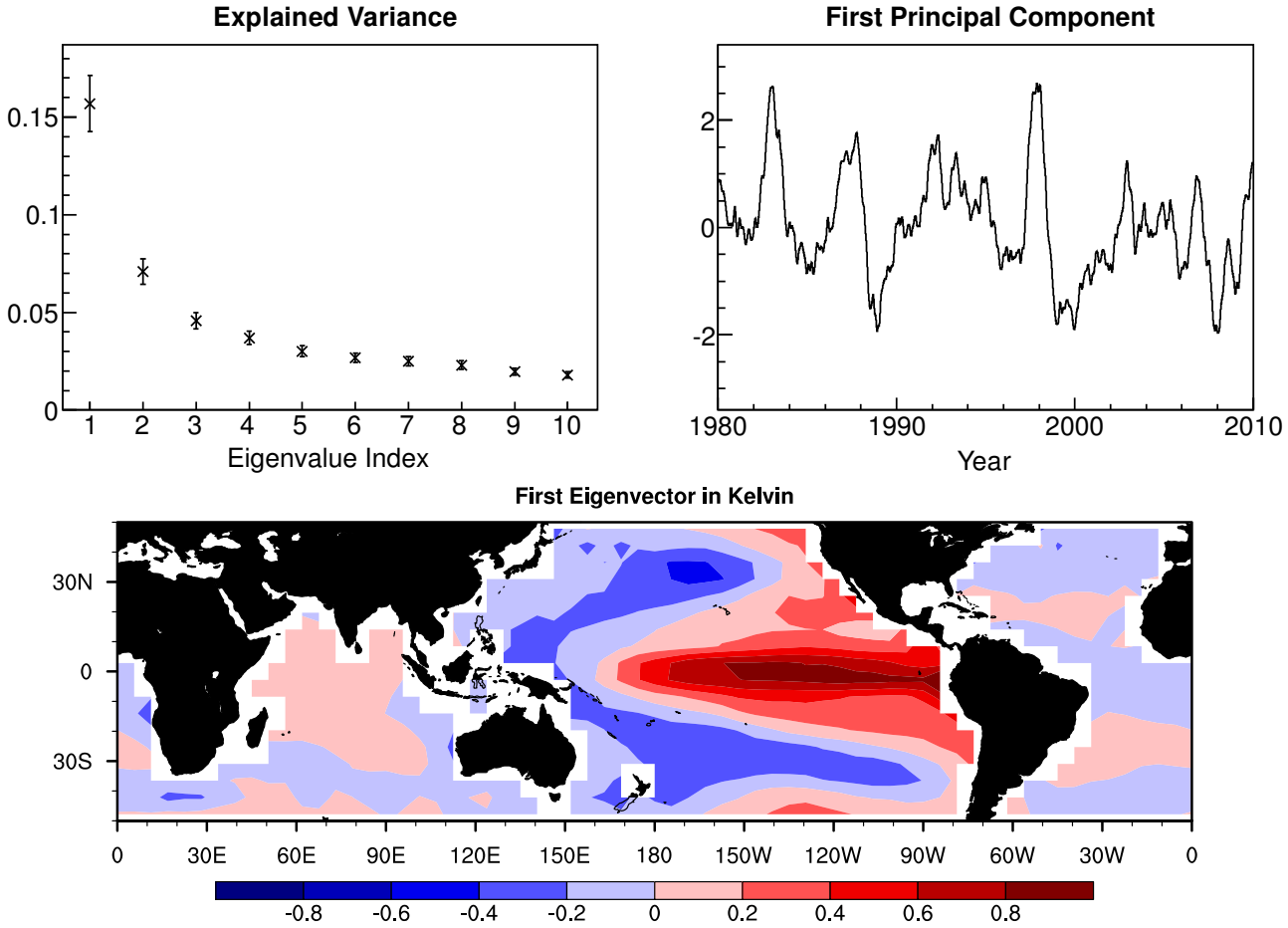

Figure 3.2: An PCA example for observed global sea surface temperature anomalies from 50° S to 50° N on a coarse grid. *Top left*: The first 10 normalized eigenvalues representing the explained variance; *Top right*: The normalized first principal component for a sample period; *Bottom*: The eigenvector (EOF) corresponding to the first eigenvalue.

Two neighboring EOFs with indices $i < j$ are now considered to be statistically independent if $\lambda_i - \Delta\lambda_i > \lambda_j + \Delta\lambda_j$, and otherwise considered to be degenerated and statistically not independent.

An PCA example for observed global sea surface temperature anomalies from 50° S to 50° N is shown in Fig. 3.2. The data are arranged so that grid points denote the columns of $\mathbf{X}$ and time the rows of $\mathbf{X}$. The first EOF explains approximately 16% of the overall variance and is statistically independent from the proceeding EOF as can be derived from the graph of normalized eigenvalues (top left). In fact, eigenvalues and corresponding eigenvectors are independent until the fourth eigenvalue since sampling errors do not overlap up to this eigenvalue. The eigenvector along with the principal component corresponding to the first eigenvalue are also displayed in Fig. 3.2. The first EOF reveals the El Niño- Southern Oscillation (ENSO). The first principal component is equivalent to the Nino3.4 index [Deser et al., 2010] showing, e.g., the prominent El Niño of 1997. It is common practice to weight each grid cell with its area prior to computing a PCA to account for converging longitudes towards higher latitudes. On a Gaussian grid, the weights are given by $\sqrt{\cos(lat)}$ where

lat denotes the current latitude [e.g., Baldwin and Dunkerton, 2001]. The eigenvectors resulting from an area-weighted data matrix should then be scaled back using these weights.

When using PCA for dimension reduction the procedure is the following: At first a fraction of the total variance $s_{\text{tot}} \in (0, 1)$ (accumulated variance) to be retained must be given. For all applications in this work a value of $s_{\text{tot}} = 0.9$ corresponding to 90% total variance is selected as it leads to a reasonable number of EOFs and a small loss of information. Once s_{tot} is selected,

$$p = \min_{p' \leq k} s_{p'} \geq s_{\text{tot}} \ , \tag{3.4}$$

denotes the number of EOFs to be retained. This means that the matrix $\mathbf{U}$ of eigenvectors reduces to a matrix $\mathbf{U}' \in \mathbb{R}^{k \times p}$ containing the first p EOFs. Putting Matrix $\mathbf{U}'$ into Eq. 3.1 leads to the matrix $\mathbf{P}' \in \mathbb{R}^{m \times p}$ holding the first p principal components. The number of dimensions are now reduced from k to p where k/p can be of the order of 10^2 or more, depending on the specific application.

After reducing the dimensions, the p columns of $\mathbf{P}'$ may be, for instance, used as targets for a regression analysis. The regression response $\mathbf{P}'_{\text{fit}}$ may then be rotated back using Eq. 3.1 to receive

$$\mathbf{X}_{\text{fit}} = \mathbf{P}'_{\text{fit}} \cdot \mathbf{U}'^{T} \ , \tag{3.5}$$

where $\mathbf{X}_{\text{fit}} \in \mathbf{R}^{m \times k}$ has the same number of rows and columns as the original data matrix $\mathbf{X}$.

To compute a principal component analysis, a C++ program was written as part of this thesis using functionality from the Intel *Math Kernel Library* (MKL) which is available at `http://www.intel.com/software/products/mkl`. The C++ program was documented and bundled in a package and is available for download at `http://wekuw.met.fu-berlin.de/~ChristianBlume`. At the Institute for Meteorology, FU Berlin, it is pre-installed and can be executed by typing 'pca' in the terminal.

3.2 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) [Wilks, 1995; Montgomery et al., 2006] is a supervised statistical method that can be used for either regression or classification. For regression problems LDA is often referred to as multiple linear regression (MLR) which is one of the most common statistical tools to analyze stratospheric variability [e.g., Crooks and Gray, 2005; SPARC CCMVal, 2010]. In the following, it is assumed that there are m training events, or m steps in time, and a vector of features $\mathbf{x} \in \mathbb{R}^k$. LDA models a set of data by using a function $y : \mathbb{R}^k \to \mathbb{K}$ given as

$$y(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} \beta + \beta_0 \ , \tag{3.6}$$

which is linear in its parameters $\beta \in \mathbb{R}^k$. In Eq. 3.6, $\beta_0 \in \mathbb{R}$ denotes a bias term, also called offset, and $\mathbb{K}$ is a either a set of class labels for classification or $\mathbb{R}$ for regression. The parameters β do not

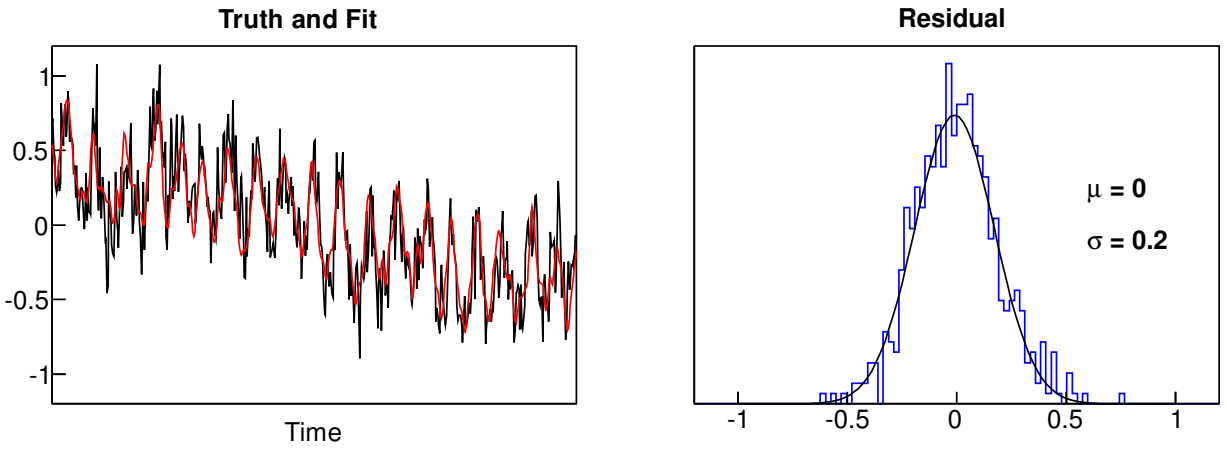

Figure 3.3: An LDA example for a sample time series of equatorial ozone anomalies at 1hPa displayed in black and regression result (fit) displayed in red. The right panel shows the corresponding residual including mean μ and standard deviation σ along with a Gaussian distribution in black fitted onto the residual.

depend on time making LDA a stationary method.

The equation for estimating $\beta \in \mathbb{R}^{k+1}$, which now includes the offset term, is

$$\mathbf{y} = \mathbf{X}\beta \,, \tag{3.7}$$

where $\mathbf{y} \in \mathbb{K}^m$ denotes the vector of targets and $\mathbf{X} \in \mathbb{R}^{m\times(k+1)}$ a matrix of the k features and one offset over m steps in time. For regression problems, the columns of $\mathbf{X}$ are often called *basis functions*. Applying the method of least squares and multiplying Eq. 3.7 with $\mathbf{X}^T$, the *normal equations* are given by

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta = \mathbf{X}^{\mathrm{T}}\mathbf{y} \quad \Longleftrightarrow \quad \beta = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \,, \tag{3.8}$$

where $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ denotes the *Moore-Penrose pseudo inverse* of $\mathbf{X}$, which requires $\mathbf{X}$ to have full rank. The LDA model response is now given by

$$\mathbf{y}_{\text{fit}} = \mathbf{X}\beta = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \,. \tag{3.9}$$

LDA, like every statistical method, will always be left with a residual given by

$$\mathbf{r} = \mathbf{y} - \mathbf{y}_{\text{fit}} \ \in \mathbb{K}^m \,, \tag{3.10}$$

which should be small compared to $\mathbf{Y}$ in terms of variance. The LDA error variance matrix is given by

$$\mathbf{E} = \frac{|\mathbf{r}|^2}{df} \cdot (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \ \in \mathbb{R}^{(k+1)\times(k+1)} \,, \tag{3.11}$$

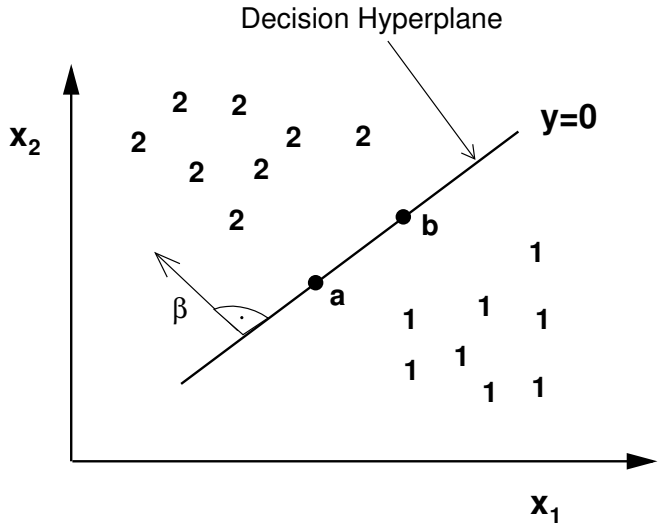

Figure 3.4: An LDA classification example with two features x_1 and x_2 and two classes 1 and 2. The decision hyperplane separates the two classes.

where df are the *degrees of freedom* given by $df = m - k - 1$. The standard errors $\Delta\beta$ of β are given through the diagonal elements of $\mathbf{E}$ in Eq. 3.11 so that

$$\Delta\beta_i = \sqrt{E_{ii}} \quad \forall i = 1, .., k \; . \tag{3.12}$$

LDA makes several assumptions to the data. The main assumptions are: 1) The relationships between the features and the target are linear; 2) The residual is distributed according to a Gaussian distribution with zero mean; and 3) The residual is not auto-correlated.

An LDA example is shown in Fig. 3.3 for a sample time series ($m = 550$) of monthly equatorial ozone anomalies at 1 hPa displayed in black. The LDA regression result (fit) is shown in red and the residual histogram in blue along with a Gaussian distribution in black fitted onto the residual. Six features (e.g., QBO, ENSO) are used in this example leading to a total number of seven free parameters (see Eq. 3.6). The residual is approximately Gaussian with zero mean, a small standard deviation compared to the truth, and an only small auto-correlation (not shown) making the regression results acceptable.

They are several extensions proposed to improve LDA as described above. One of them is the Fourier expansion applied for atmospheric problems by e.g., Bodeker et al. [1998] and in Chapter 8 of SPARC CCMVal [2010]. The idea is to multiply each feature (or basis function) with a sine and cosine term of a certain period. In atmospheric science, this period could be one year so that the Fourier expansion would express the seasonal dependency of a certain feature. However, this work does not make use of that because it significantly increases the parameter space (e.g., factor of 5 for two Fourier pairs per feature) and does usually not reduce the residual significantly. Therefore, it

does not meet the principle of Occam's Razor which favors simpler models while carrying sufficient information (see Section 3.7).

When LDA is used for classification problems, the offset β_0 is usually chosen so that $y < 0$ for class 1 and $y \geq 0$ for class 2 in the case of two classes. Then, a geometrical picture can be drawn: Let $\mathbf{a}$ and $\mathbf{b}$ be two events on the decision boundary, it follows that $y(\mathbf{a}) = y(\mathbf{b}) = 0$ and therefore $(\mathbf{a} - \mathbf{b})^{\mathrm{T}} \cdot \beta = 0$. This is shown schematically for a 2-dimensional example in Fig. 3.4 with two features x_1 and x_2 and two classes 1 and 2. It is observed that LDA is the task of finding a vector β which is orthogonal to the decision hyperplane.

To solve regression and classification problems with LDA, a C++ program was written as part of this thesis using classes from the *Toolkit for Multivariate Data Analysis* (TMVA) which is available at `http://tmva.sourceforge.net`.

3.3 Finite element method plus vector auto-regression (FEM-VARX)

FEM-VARX [Horenko, 2010b] is a supervised *and* unsupervised statistical method. FEM-VARX uses a combination of vector auto-regression with external factors (VARX) [Hamilton, 1994] and a recently introduced clustering procedure [Horenko, 2010a] based on a finite element method (FEM). The FEM clustering approach was recently successfully applied to identify large-scale dynamical circulation patterns in GCM simulations [Franzke et al., 2009]. The VARX model is trained in supervised mode whereas the FEM clustering is trained in unsupervised mode. FEM-VARX, just like LDA, can be used in a nonlinear way by transforming the features with nonlinear functions. The optimal functions performing this transformation are generally unknown. In fact, there is an infinite number of possible transfer functions. Therefore, FEM-VARX along with LDA are used as linear models. However, due to a nonstationary jump process switching between regimes (called *clusters*) within the given system, FEM-VARX is a powerful time series model. Due to this jump process, FEM-VARX is a nonstationary statistical method.

In the following, it is assumed that there is a feature vector $\mathbf{x} \in \mathbb{R}^n$ (n external factors), T steps in time, and a target vector $\mathbf{y} \in \mathbb{R}^d$ in d dimensions over T steps in time. Let us further assume that $\mathbf{y}$ can be partitioned into K clusters with parameters $\Theta = (\theta_1, ..., \theta_K)$ and cluster affiliations $\mathbf{\Gamma}(t) = (\gamma_1(t), ..., .\gamma_K(t))$. The cluster affiliations are subject to

$$\gamma_i(t) \geq 0 \qquad \sum_{i=1}^{K} \gamma_i(t) = 1 \qquad \forall t \in [1, T] \; . \tag{3.13}$$

Since clusters represent persistent states of the analyzed system, a *persistency threshold* C is

introduced which limits the maximum transitions from a given cluster to any other cluster so that

$$\sum_{t=1}^{T-1} |\gamma_i(t+1) - \gamma_i(t)| \leq C \qquad \forall i = 1, ..., K \ . \tag{3.14}$$

The cluster parameters $\theta_i = (\mu^i, \mathbf{A}^i, \mathbf{B}^i)$ contain the parameters of the VARX model response $\mathbf{y}^*_{t,i}$ for cluster i and time t given by

$$\mathbf{y}^*_{t,i} = \mu^i + \sum_{q=1}^{m} \mathbf{A}^i_q \mathbf{y}_{t-q} + \mathbf{B}^i \mathbf{x}_t \qquad \forall i = 1, ..., K \ , \tag{3.15}$$

where $\mu^i \in \mathbb{R}^d$, $\mathbf{A}^i_q \in \mathbb{R}^{d \times d}$, and $\mathbf{B}^i \in \mathbb{R}^{d \times n}$. In Eq. 3.15, m denotes the *memory* determining the order of the auto-regressive process. Including auto-regressive processes can be very useful to forecast time series as it involves the use of information from time $t - q$ about the time series to be forecasted for time t. Unfortunately, it is computationally expensive to compute the optimal value for m. In addition, due to the more complex structure of the VARX model for $m > 0$ in Eq. 3.15, it is much harder to extract an estimate for the importance of external factors. This importance may even be reduced due to the additional auto-regressive term. For these reasons, this work does not consider auto-regressive processes implying $m = 0$.

In order to determine cluster parameters and cluster affiliations, the *average clustering functional* given by

$$L(\Theta, \mathbf{\Gamma}(t)) = \sum_{i=1}^{K} \sum_{t=0}^{T} \gamma_i(t) g(\mathbf{y}_t, \theta_i) \rightarrow \min_{\Theta, \mathbf{\Gamma}(t)} \ , \tag{3.16}$$

is minimized, where $g(\mathbf{y}_t, \theta_i) = |\mathbf{y}_t - \theta_i|_2$ denotes the model distance functional. L is minimized by a discretization of the data in finite elements [Horenko, 2010a] and a subsequent simulated annealing procedure [Kirkpatrick et al., 1983]. Simulated annealing is a numerical technique that aims at finding a global solution to a complex optimization problem. It may, however, only find a local instead of a global optimum. Simulated annealing is an adaptation of the Metropolis-Hastings algorithm [Metropolis et al., 1953], producing random realizations of a thermodynamic system using Monte-Carlo methods. If not stated otherwise, a sufficiently large number of annealing steps of 100 is used to find the near-optimal solution.

Having L minimized and $\mathbf{\Theta}$ and $\mathbf{\Gamma}(t)$ computed, the model response is evaluated by

$$\mathbf{y}^*_t = \sum_{i=1}^{K} \gamma_i(t) \mathbf{y}^*_{t,i} \qquad \forall t \in [1, T] \ , \tag{3.17}$$

where $\mathbf{y}^*_{t,i}$ denotes the local VARX model response for cluster i at time t (see Eq. 3.15). The total number of free parameters N_P in FEM-VARX is given by the number of parameters in the VARX

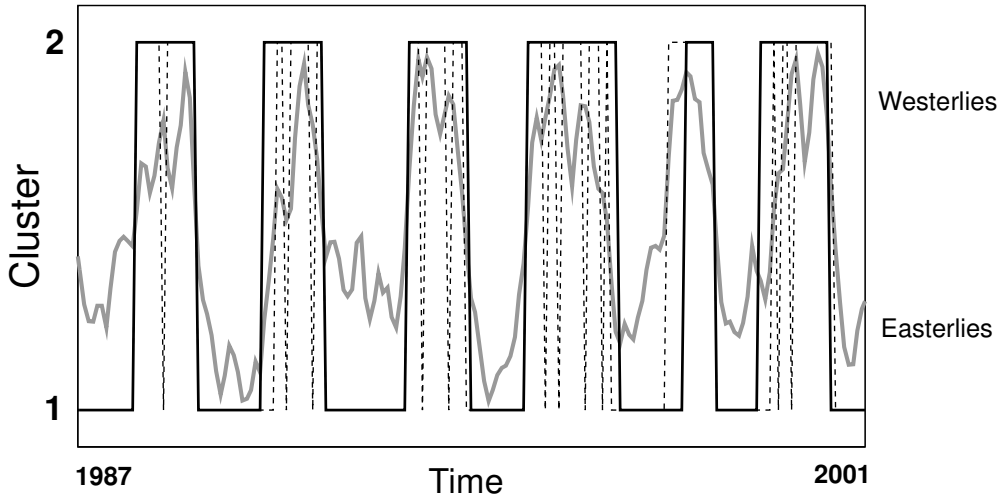

Figure 3.5: FEM-VARX cluster affiliation for $K = 2$ for a sample time series of observed equatorial stratospheric zonal mean zonal wind anomalies at 20hPa (gray) taken from ERA-40 for a representative period. The algorithm was constrained with $C = 100$ (solid black) and $C = 400$ (dashed black).

model (Eq. 3.15) and the cluster affiliation (Eq. 3.13) along with the number of clusters K so that

$$N_P = K \cdot (N_\Gamma + d + n \cdot d + m \cdot d^2) \,, \tag{3.18}$$

where N_Γ denotes the number of cluster transitions (or *jumps*).

In general, FEM-VARX has three tuning parameters to be determined for every application. These parameters are the number of clusters K, the persistency threshold C, and the memory m. Since $m = 0$ in this work, the search is restricted to finding optimal pairs of K and C with the help of information criteria (see Section 3.7).

As an example of the FEM-VARX clustering, a one-dimensional time series of equatorial stratospheric zonal mean zonal wind anomalies at 20hPa is analyzed without incorporating any external factors. The FEM cluster affiliation for two clusters is displayed in Fig. 3.5 with the time series in gray. The algorithm was constrained with two different persistency thresholds: $C = 100$, shown in solid black, and $C = 400$, shown in dashed black. As observed from Fig. 3.5, the clustering for larger C leads to more transitions between the clusters in agreement with Eq 3.13. The result for $C = 100$ leads to a more persistent clustering clearly representing lower values of the time series in cluster 1 (easterly winds) and higher values in cluster 2 (westerly winds).

After the FEM-VARX time series model has been successfully trained, forecasts can be made using FEM-VARX. As forecasting involves the evaluation of Eq. 3.17 with unseen features, a procedure is needed to forecast $\mathbf{\Gamma}(t)$ for $t > T$ in a first step. This can be done by assuming that the jump process $\mathbf{\Gamma}(t)$ is a Markov process. A Markovian system is governed by a process in which the

next state only depends on the current state. The forecast of the jump process can be expressed by

$$\mathbf{\Gamma}(t+1) = \mathbf{\Gamma}(t)\mathbf{P}(\mathbf{x}_t) \qquad \text{for } t > T \,, \tag{3.19}$$

where $\mathbf{P}(\mathbf{x}_t)$ denotes the probabilistic transition matrix at time t depending on features $\mathbf{x}_t$. The computation of $\mathbf{P}(\mathbf{x}_t)$ is described in detail by Horenko [2011]. Once $\mathbf{\Gamma}(t)$ is available, Eq. 3.15 and Eq. 3.17 can be evaluated with external factors from the forecast period to obtain a prediction $\mathbf{y}_t^*$ for $t > T$.

To solve regression and clustering problems with FEM-VARX, a MATLAB program was used written by Illia Horenko and colleagues from the Institute of Computational Science at the University of Lugano. This program was further adjusted for efficient use as part of this thesis. Additionally, a wrapper class in C++ was written for the FEM-VARX time series model.

3.4 Multi-Layer Perceptron (MLP)

A multi-layer perceptron (MLP) [Ripley, 1996] can be trained as a supervised or unsupervised method, depending on the problem to be solved. In this work, an MLP is trained in supervised mode. It is stationary as its parameter space does not depend on the event number or time when analyzing time series. An MLP can be used in a linear or nonlinear fashion depending on the selected transfer function. An MLP is a neural network which is actively being applied in atmospheric science. For instance, Lu et al. [2009] used a neural network to receive an alternative representation of the QBO; Walter and Schönwiese [2002] addressed the detection and attribution of observed global climate change in global temperature anomalies using a neural network; and Elgaali and Garcia [2004] modeled the climate change impacts on irrigation water supplies in Arkansas river basin with a neural network. A review of neural networks is given by Zhang et al. [1998].

In the following, it is assumed that there is a feature vector $\mathbf{x} \in \mathbb{R}^k$ with k elements and m training events (m steps in time). An MLP is a fully connected feed-forward neural network [Bishop, 1995] with one or more hidden layers located between input and output layer. Each layer consists of a certain number of neurons in parallel. Fig. 3.6 shows a neuron with sigmoidal transfer function and a two-layer MLP in a schematic picture. Each neuron calculates a weighted linear combination of its N inputs so that its output y is given by

$$y = f\left(\sum_{i=1}^{N} w_i x_i + \theta\right) \,, \tag{3.20}$$

where $w_i \in \mathbb{R}$ and $\theta \in \mathbb{R}$ denote weights and bias, respectively, and x_i the input to the neuron from synapse i. The weights are given at each synapse (connection between two neurons) and the biases at each neuron. The scalar function f in Eq. 3.20 is called the *transfer function*. It is most commonly chosen to be a sigmoid $f : \mathbb{R} \to (0, 1)$ of the form $f(x) = (1 + e^{-x})^{-1}$ [Kurkova, 1992] which is used in this work and displayed in Fig. 3.7. Other common transfer functions are the radial

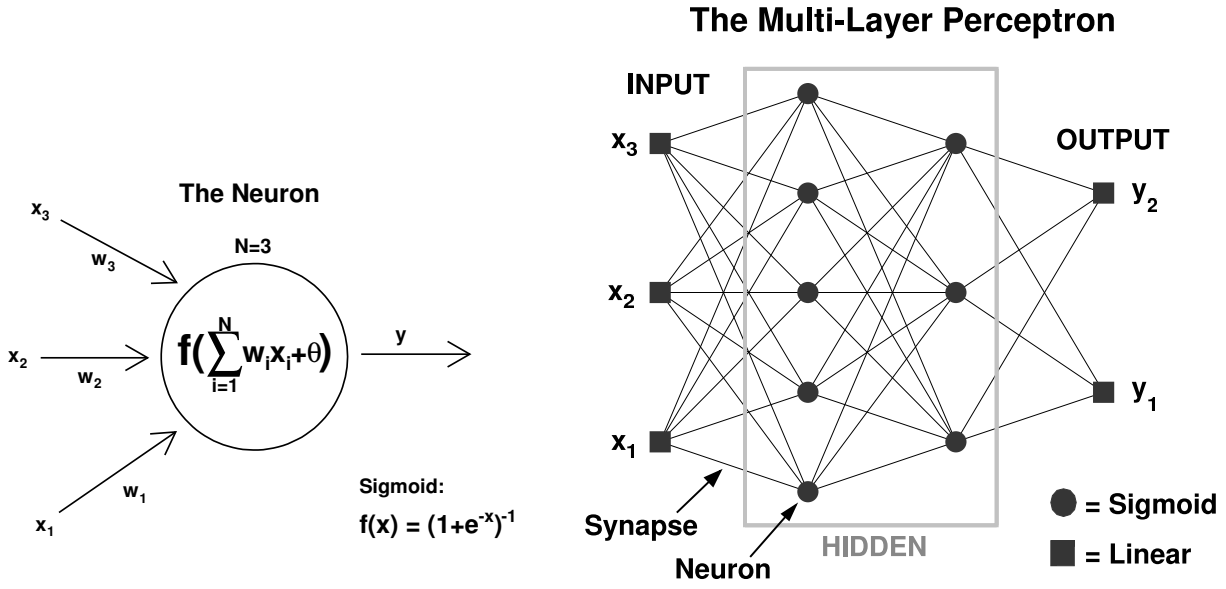

Figure 3.6: Schematic picture of a neuron (left) with sigmoidal transfer function and a multi-layer perceptron (right) with input, output, and two hidden layers.

basis function and the hyperbolic tangent [Bishop, 1995]. The transfer functions at the output layer are linear (compare Fig. 3.6).

Approximation and generalization performance of the MLP stem from the nonlinear transfer functions and the numerous connections within the hidden layer(s). An MLP with a single hidden layer implements a single hyperplane. An MLP with two hidden layers implements arbitrary convex regions containing intersections of hyperplanes. It has been shown that an MLP with sigmoidal transfer functions and two hidden layers can approximate any continuous function [Kurkova, 1992]. For this reason, the analysis is restricted to an MLP with a maximum of two hidden layers and the sigmoid transfer function.

The learning algorithm used to determine the free parameters of the network is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [Avriel, 2003]. It is a faster variation of the standard back-propagation [Bishop, 1995]. In this work, a maximum of 1000 training iterations (epochs) are performed. To avoid overfitting and computational overhead, the training is aborted once the training algorithm converges or once 1000 epochs are reached.

BFGS uses a gradient search technique to iteratively adjust weights $\mathbf{w}$ and biases Θ via minimizing a *cost function* given by

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{q}(\mathbf{y}_j^{(i)} - \mathbf{y}_j^{*(i)}(\mathbf{w},\Theta))^2 \rightarrow \min_{\mathbf{w},\Theta}\,, \tag{3.21}$$

where q denotes the number of output neurons. In Eq. 3.21, $\mathbf{y} \in \mathbb{K}^q$ denotes the desired output and $\mathbf{y}^* \in \mathbb{K}^q$ the actual MLP response over m training events with $\mathbb{K}$ being a set of class labels for classification problems or $\mathbb{R}$ for regression. The minimization problem is unconstrained and generally

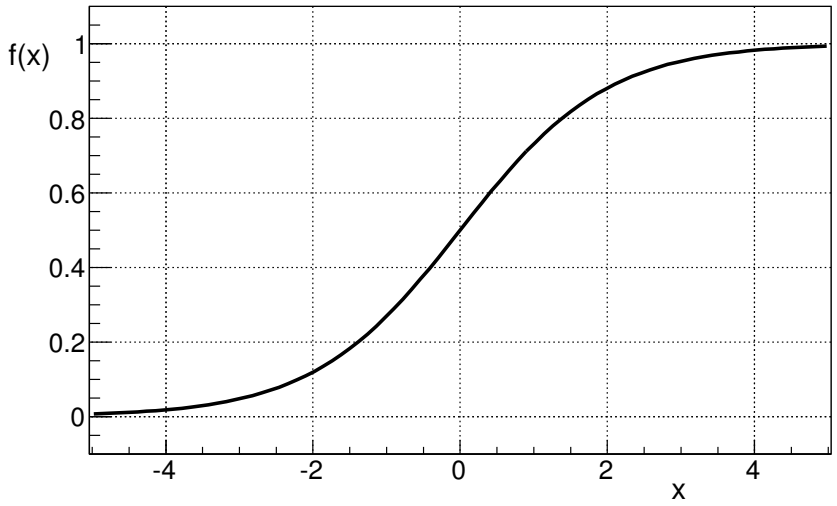

Figure 3.7: The sigmoid transfer function given by $f(x) = (1 + e^{-x})^{-1}$ as used in the multi-layer perceptron.

non-convex [Bishop, 1995]. The effect of resulting local minima can be reduced by performing several training realizations ($\mathcal{O}(10)$) with different initial values for weights and biases. If not stated otherwise, 10 different MLP training realizations are performed in this work while retaining the realization with the smallest training error.

There are two tuning parameters that have to be determined: 1) The number of neurons ($L_1 > 0$) in the first hidden layer and 2) The number of neurons ($L_2 \geq 0$) in the second hidden layer. The MLP tuning parameters can be determined with information criteria or cross-validation as described in Section 3.7. From L_1 and L_2 along with the number of features L_0 (No. of input neurons) and targets L_3 (No. of output neurons), the MLP's total number of free parameters is given by

$$N_P = N_\mathrm{w} + N_\Theta \quad \text{with} \quad N_\mathrm{w} = L_0 L_1 + L_1 L_2 + L_2 L_3 \ , \ N_\Theta = L_0 + L_1 + L_2 + L_3 \ , \tag{3.22}$$

where N_w denotes the total number of weights and N_Θ the total number of biases of the MLP. Please note that Eq 3.22 can be easily extended to MLPs with more than two hidden layers.

When an MLP is used to solve classification problems, the network response needs to be post-processed in order to obtain conditional probabilities. This is done via the *softmax* function given by

$$p_i = \frac{\exp(\mathbf{y}_i^*)}{\sum_{j=1}^{q} \exp(\mathbf{y}_j^*)} \ , \tag{3.23}$$

where q denotes the number of output neurons and p_i the probability of output i given the remaining probabilities p_j for all $j \neq i$.

To solve regression and classification problems with an MLP, a C++ program was written as part of this thesis using classes from the *Toolkit for Multivariate Data Analysis* (TMVA) which is

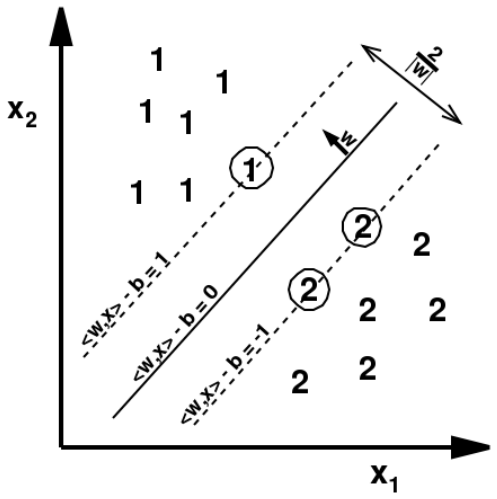

Figure 3.8: Schematic picture of the principles of support vector classification for two classes $\{1, 2\}$. The encircled class members denote the support vectors. The decision hyperplane (solid line) is shown along with the classification margin (dashed lines).

available at `http://tmva.sourceforge.net`.

3.5 Support Vector Machines (SVMs)

Support vector machines (SVMs) [Vapnik, 1995] are statistical methods that can be applied for regression and classification problems. For regression, it is referred to as support vector regression (SVR) whereas for classification, it is referred to as support vector classification (SVC). SVMs are stationary as their parameter space does not depend on the event number, or when analyzing time series, on time. They can be used in a linear or nonlinear fashion. A detailed SVM tutorial is given by Burges [1998]. There were also extensions developed that allow SVMs to be used for unsupervised data clustering [Ben-Hur et al., 2001]. In this work, SVMs are exclusively applied to supervised regression and classification problems. In the following, it is assumed that there is a feature vector $\mathbf{x} \in \mathbb{R}^k$ in k dimensions and m training events (m steps in time).

3.5.1 Support Vector Classification

Support vector machines aim at classifying a data set by maximizing a margin that defines a separating hyperplane. In general, the larger this margin the better the generalization performance. Let us assume that there is a target $y \in \{1, 2\}$ over m training events that is not perfectly separable, which means that there will be a certain amount of misclassification. Then, a vector $\mathbf{w} \in \mathbb{R}^k$, a parameter $b \in \mathbb{R}$, and a so called "slack" variable $\xi_i \geq 0$ can be found so that

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - b) \geq 1 - \xi_i \quad \forall i \in \{1, ..., m\} \; , \tag{3.24}$$

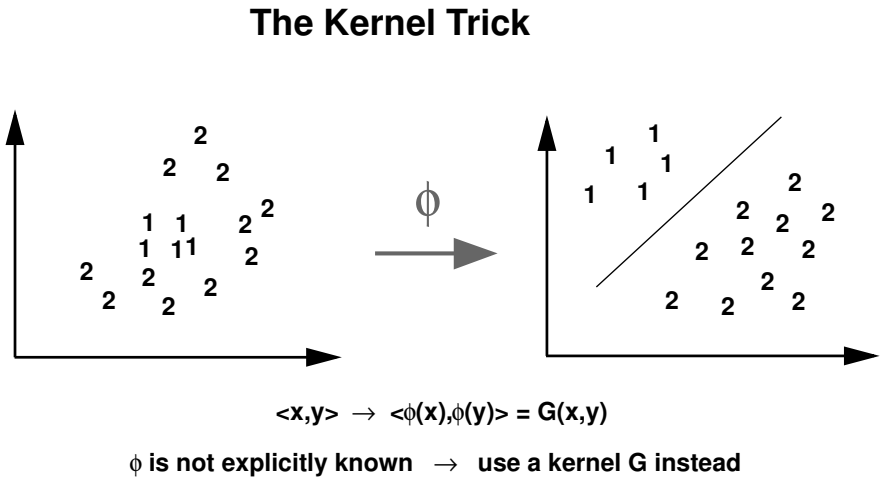

Figure 3.9: A schematic picture of the kernel trick in a two-dimensional feature space and two classes $\{1, 2\}$. The kernel trick aims at transforming into a linearly separable space without ever computing ϕ explicitly.

where $\langle \cdot, \cdot \rangle$ denotes a dot-product. The principles of support vector classification are shown in Fig. 3.8 for two classes $\{1, 2\}$ and a two-dimensional feature vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$. It is observed that the pair $(\mathbf{w}, b)$ defines the separating hyperplane where $\langle \mathbf{w}, \mathbf{x} \rangle - b = 0$. The encircled feature vectors on the margin (dashed lines) are called the support vectors.

The classification margin defined by $M = 2/|\mathbf{w}|$ is to be maximized with respect to the constraints given in Eq. 3.24. Hence, maximizing the margin is equivalent to minimizing the cost function

$$\frac{1}{2}|\mathbf{w}|^2 + C \sum_{i=1}^{m} \xi_i \rightarrow \min_{\mathbf{w}, b, \xi} , \tag{3.25}$$

being subject to Eq. 3.24. In Eq. 3.25, $C > 0$ denotes a parameter describing the trade-off between maximizing the margin and minimizing the training error. Introducing slack variables is equivalent to support vector machines with *soft margins*. Eq. 3.25 is a constrained quadratic optimization problem that has a unique solution. It is solved by translating into Lagrangian formalism [Burges, 1998]. The resulting non-zero Lagrangian multipliers define the support vectors.

In order to use SVMs in a nonlinear way, the *kernel trick* is used which transforms the feature space using a nonlinear kernel function into a higher dimensional space (SVMs are kernel machines). The dot-product in Eq. 3.24 is replaced by a kernel function. Ideally, the kernel trick makes the transformed feature space linearly separable. A schematic picture of the kernel trick in a two-dimensional feature space and two classes $\{1, 2\}$ is provided in Fig. 3.9. The kernel trick aims at transforming the feature space into a linearly separable space of features without ever computing the transferring function ϕ explicitly. This principle is applicable wherever there are only dot-products and a suitable kernel can be found.

The most common kernels are the polynomial, the radial basis, and the hyperbolic tangent kernel.

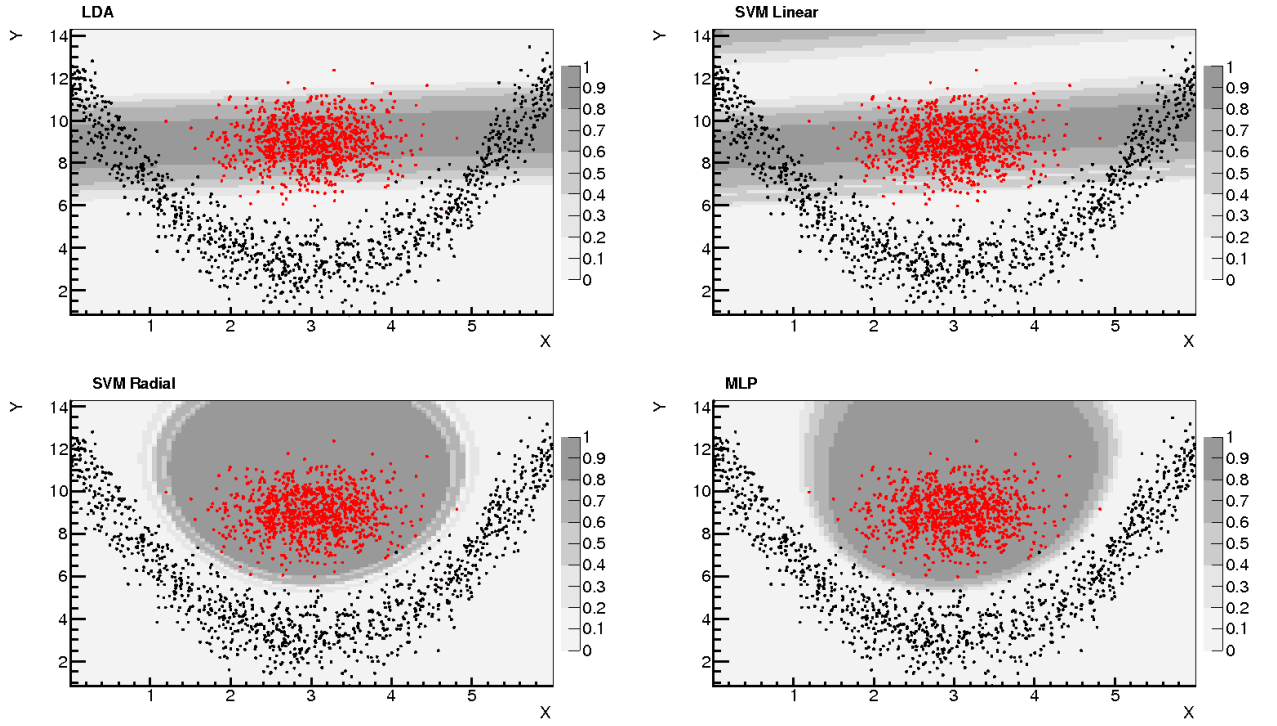

Figure 3.10: A nonlinear, two-dimensional classification example with two classes (red, black) in which LDA, MLP, SVM with linear kernel, and SVM with a radial basis kernel are compared. Shading denotes the probability of being in class red.

It was shown that the radial basis kernel is the most efficient while being simple and including the linear case [Keerthi and Lin, 2003]. Therefore, this work uses the radial basis kernel when aiming at the nonlinear statistical modeling with SVMs. It is given by

$$G(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma|\mathbf{x}_i - \mathbf{x}_j|^2) \,, \tag{3.26}$$

where $\gamma > 0$ denotes the kernel parameter. To apply SVMs in practice, the cost parameter C and the radial basis parameter γ have to be selected. This is done using cross-validation (see Section 3.7.2) on a grid of pairs of (C, γ).

A nonlinear two-dimensional classification example with two classes (red, black) in which LDA, MLP, SVM with linear kernel, and SVM with a radial basis kernel are compared is displayed in Fig. 3.10. The shading denotes the probability of being in class red. It is observed that LDA and SVM linear ($C = 1$) are not able to discriminate between the classes. In contrast, the nonlinear classifiers, SVM radial ($C = 1,\ \gamma = 1$) and MLP ($L_1 = 10,\ L_2 = 5$), are able to efficiently classify this data set.

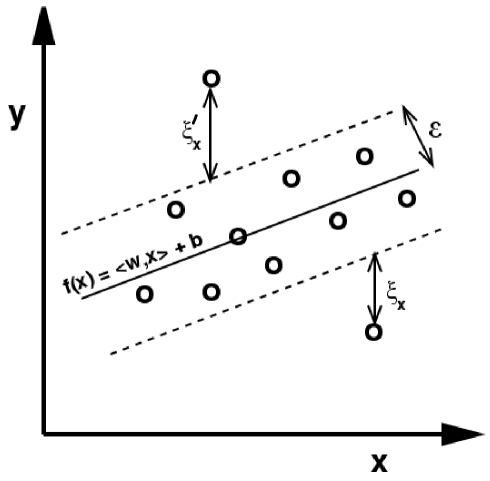

Figure 3.11: Schematic picture of the principles of support vector regression. The circles denote the data points that are supposed to be approximated using SVR. The regression result (solid line) is shown along with the "wiggle" room around it (dashed lines) given by ε. It is observed that $|y - f(x)| - \varepsilon$ denotes the maximum loss.

3.5.2 Support Vector Regression

Let there be a target $y \in \mathbb{R}$ over m training events. The support vector algorithm for regression is similar to that for SVC. Analogous to soft margins by introducing slack variables, SVR is concerned with using Vapnik's ε-*insensitive loss function* [Vapnik, 1995]. The cost function in Eq 3.25 transforms to

$$\frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i') \rightarrow \min_{\mathbf{w},b,\xi,\xi^*}, \tag{3.27}$$

which is to be minimized subject to

$$f(\mathbf{x}_i) - y_i \leq \varepsilon + \xi_i \tag{3.28}$$

$$y_i - f(\mathbf{x}_i) \leq \varepsilon + \xi_i', \tag{3.29}$$

for $\xi_i, \xi_i' \geq 0$ and all $i = 1, ..., m$, where $f(x) = G(w, x) + b$ with the radial basis kernel G (see Eq. 3.26) and $b \in \mathbb{R}$. A schematic picture of SVR for a one-dimensional regression problem is shown in Fig. 3.11. For simplicity, this is presented without using the kernel-trick. The circles denote the data points that are supposed to be approximated and the solid line the actual regression result. SVR aims at approximating y leading to a maximum loss of $|y - f(x)| - \varepsilon$. Through the introduction of slack variables ξ_i, SVR is generally more robust against outliers than LDA. Similar to SVC, Eq. 3.27 is minimized by translation into Lagrangian formalism [Vapnik, 1995].

In SVR, one faces the problem of selecting another parameter ε representing the maximum loss. The work made by Smola et al. [1998] shows that an optimum ε scales with the random noise level of y. It is often unclear what the noise level might be. For the applications in this work, it turns

out that $\varepsilon = 0.1\sigma_y$ is a reasonable value leading to acceptable training and forecast results. σ_y denotes the standard deviation of the target time series y. SVR results presented in this work, are also rather insensitive to choices of ε. Altering ε by a factor of 2 does not change regression results notably. After ε is selected, one has to determine optimal pairs of (C, γ) using cross-validation (see Section 3.7.2).

To solve regression and classification problems with SVR and SVC, respectively, a C++ program was written using *LIBSVM* (A Library for Support Vector Machines) which is available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

3.6 Performance Measures

After a statistical method has been successfully trained it can be evaluated with features (external factors) it has been trained with (seen data) to receive a *fit*; or with features it has *not* been trained with (unseen data) to receive a *forecast*. In order to quantify how well the method performs, performance measures have to be computed for fit and forecast assuming that there exist a true representation, called *truth*, for both regimes. In the following, y denotes the truth, y^* the response of the statistical method, and m the number of events.

When considering performance measures, it must be generally distinguished between classification and regression problems. However, there are two popular measures that can be used in both cases:

1. The *mean prediction error* (MPE) is defined as

$$\text{MPE} = \frac{1}{m}\sum_{i=1}^{m} |y_i - y_i^*| \,, \tag{3.30}$$

   which measures the averaged deviation of the response from the truth.

2. The *sum of squared errors* (SSE) is defined as

$$\text{SSE} = \sum_{i=1}^{m} (y_i - y_i^*)^2 \,. \tag{3.31}$$

   from which error variance $s_e = \frac{1}{m}\text{SSE}$ and error standard deviation $\sigma_e = \sqrt{s_e}$ are derived assuming that the residual is mean-centered.

3.6.1 Classification

For the following let there be two classes, namely the signal (class 1) and the background (class 0), and let y_S and y_B be the respective area-normalized response distributions. Three important performance measures for classification problems can be defined:

1. The *separation* S between signal and background is given by

$$S = \frac{1}{2}\int_{-\infty}^{\infty} \frac{(y_S(x) - y_B(x))^2}{y_S(x) + y_B(x)} \mathrm{d}x \ . \tag{3.32}$$

   which quantifies how well the method is able to discriminate signal from background events.

2. The *signal efficiency* ε_S at a given *background efficiency* ε_B is defined by

$$\varepsilon_S = \int_a^{\infty} y_S(x) \mathrm{d}x \ . \tag{3.33}$$

   where a is given by $\varepsilon_B = \int_a^{\infty} y_B(x) \mathrm{d}x$. A representative background efficiency of 0.01 has been selected. ε_S also measures the power of discriminating signal from background.

3. The integral of the *receiver operating characteristic* (ROC) curve is given by

$$I_{\mathrm{ROC}} = \int_0^1 (1 - \varepsilon_B) \mathrm{d}\varepsilon_S \ , \tag{3.34}$$

   where $1 - \varepsilon_B$ is called the *background rejection*. I_{ROC} measures the integrated true positive rate of a classification problem.

The three performance measures S, ε_S, and I_{ROC} are bounded between 0 and 1, where 0 means the worst and 1 the best performance achievable. Detailed overviews of signal analysis can be found in e.g., Spackman [1989] and Fawcett [2006].

3.6.2 Regression

Aside from MPE and SSE, two additional performance measures are suitable for regression problems:

1. The *correlation coefficient* R given by

$$\mathrm{R} = \frac{\mu_{yy^*} - \mu_y \mu_{y^*}}{\sigma_y \sigma_{y^*}} \ , \tag{3.35}$$

   where μ denotes the mean and σ the standard deviation of the distribution of the truth y or the response y^*, respectively. R measures the linear dependence between two variables and is bounded between -1 and 1 where 1 denotes the best performance achievable.

2. The *explained variance* EV is given by

$$\mathrm{EV} = 1 - \frac{\sigma_e^2}{\sigma_y^2} \ , \tag{3.36}$$

   where $\sigma_e^2 = \frac{1}{m}\mathrm{SSE}$ denotes the error variance and σ_y^2 the sample variance of the truth y. EV does not have a lower bound but its upper bound is 1. The closer EV to 1 the better the regression result.

3.7 Optimal Model Architecture

When applying statistical methods to solve classification or regression problems, a number of *tuning* parameters must be selected. A full set of these parameters for a specific method and underlying algorithm is called an *optimal model architecture* [Burnham and Anderson, 2002]. An example would be the number and kind of features, the number of clusters in FEM-VARX, the number of hidden layers in the MLP, or the value of an optimal trade-off parameter in SVM.

There are two major approaches that tackle the problem of model selection, both aiming at meeting the principle of Occam's Razor [Ariew, 1976]. Occam's Razor states that the simplest model is the preferred model if it contains as much information as any of the more complicated models. A model represents one incarnation or one specific architecture of a statistical method.

The two approaches are 1) *information criteria* [Burnham and Anderson, 2002] that aim at minimizing

$$\text{training error} + \text{number of parameters}, \tag{3.37}$$

and 2) the method of *cross-validation* [Kohavi, 1995] that aims at minimizing the

$$\text{forecast error}. \tag{3.38}$$

A more complex model, i.e. a model with more degrees of freedom (e.g. more parameters), is generally able to more efficiently approximate the training data than a simpler model. However, as the model complexity increases, the forecast error will eventually start decreasing which is denoted as *overfitting*. It is the goal of Occam's Razor to find the optimal trade-off between approximating the training data and generalizing to unseen data (forecasting).

Both approaches presented here, information criteria and cross-validation, are very different and one is more suitable than the other for certain situations and methods. Cross-validation is generally not suitable for non-stationary methods. Hence, it cannot be used with FEM-VARX since cross-validation does not retain the temporal order of the cluster affiliation (see Section 3.3). On the other hand, only cross-validation is suitable for methods, such as SVM (see Section 3.5), where the number of tuning parameters is constant but their optimal values need to be estimated. Therefore, information criteria can be applied to LDA, MLP, and FEM-VARX whereas cross-validation can be applied to LDA, MLP, and SVM. Both model selection approaches will be applied in this work and reviewed in the following sections.

3.7.1 Information Criteria

There are two major information criteria [Liddle, 2008]. The first to be reviewed is the *Bayesian Information Criterion* (BIC) [Schwarz, 1978] which is defined as

$$\text{BIC} = -2\ln(L_{\max}) + \ln(n)k \tag{3.39}$$

where $L_{\max}$ denotes the maximized value of the model likelihood, k the number of model parameters, and n the number of training events. Under the assumption that the residuals are independent and identically distributed according to a Gaussian distribution [Priestley, 1983], Eq. 3.39 becomes

$$\boxed{\text{BIC} = n\ln(\sigma_e^2) + \ln(n)k} \tag{3.40}$$

where σ_e^2 denotes the error variance.

The second criterion to be reviewed is the *Akaike Information Criterion* (AIC) [Akaike, 1974] which is defined as

$$\text{AIC} = -2\ln(L_{\max}) + 2k \tag{3.41}$$

where $L_{\max}$ denotes again the maximized value of the model likelihood and k the number of model parameters. Corrected for small sample sizes [Liddle, 2008] and under the assumption that the residuals are independent and identically distributed according to a Gaussian distribution, Eq. 3.41 becomes

$$\boxed{\text{AIC} = n\ln(\sigma_e^2) + 2k + \frac{2k(k+1)}{n-k-1}} \tag{3.42}$$

where σ_e^2 denotes the error variance and n the number of training events.

The simplified versions of BIC and AIC given in Eq. 3.40 and Eq. 3.42 respectively, can now be easily applied in practice having in mind the assumptions about the residuals. The difference between AIC and BIC is essentially that the BIC penalizes the parameter space more strongly than the AIC does [McQuarrie and Tsai, 1998], meaning that a model selected by the BIC is often less complicated. This can be seen when comparing the seconds summands in Eq. 3.39 and Eq. 3.41 since $\ln(n)k$ (BIC) is usually larger than $2k$ (AIC) for the same number of parameters k.

When applying the information criteria to LDA and MLP, the simplified versions given in Eq. 3.40 and Eq. 3.42 will be used. The SVM tuning parameters are estimated using cross-validation, since the number of parameters in SVM is fixed (C and γ), but their optimal values need to be determined.

For FEM-VARX, there are no assumptions being made about the residuals as they are fitted explicitly for each cluster by choosing from a wide set of likelihood functions (such as, Gaussian, exponential, etc.). Hence, BIC and AIC are calculated in its general form given by Eq. 3.39 and Eq. 3.41 respectively, incorporating the number of FEM-VARX model parameters (see Eq. 3.18).

The procedure of model selection using information criteria is the following: Firstly, all models are trained and evaluated that could be a potential candidate for the optimal model which depends highly on the specific application and the statistical method to be used. Secondly, BIC and AIC are evaluated for every model, where the model with the lowest value of each BIC and AIC is preferred. The result will be two models because two independent criteria are applied. A favorable but rare situation is when both BIC and AIC point to the same model. Generally, the models will be different and the model used must be selected heuristically. For instance, the more complicated model or the model not including an important feature would be rejected. Therefore, the more that is known about the given system, the better the guess can be on which criterion to use.

3.7.2 Cross-Validation

The method of cross-validation [Kohavi, 1995] aims at selecting the model with the smallest forecast error or, in other words, the model with the best generalization performance. The data set with which this performance is evaluated can be independent or an unseen part of the data set that the model is trained with. To avoid systematic errors, the evaluation data set must be drawn from the same population as the training data set.

They are many different approaches to cross-validation [Picard and Cook, 1984]. One very common approach is dividing a data set into k subsets where the subset members may be randomly picked or assigned with knowledge of the system. If certain events are logically necessary to be included during training, then they should not be used for forecasting. In this work, a time series is divided into k contiguous subsets of equal length, ensuring that k is not too small so that there are always sufficient training events. The model is then trained on all $k-1$ subsets where the subset being left out is used to compute the forecast error. This is done for all k subsets, so that each subset is forecasted exactly one time. All forecasts are combined and the overall cross-validation error can be computed represented by the mean prediction error (see Eq. 3.30) or the correlation coefficient (see Eq. 3.35).

With cross-validation, each model has to be trained and evaluated k times, meaning that cross-validation is approximately a factor of k computationally more expensive than information criteria. However, if the main goal is to statistically forecast, cross-validation should be used since it aims at minimizing the forecast error. If the goal is not to forecast but to select the model with the best descriptive power, information criteria should be used since information criteria optimize on the full training period. As previously mentioned, the optimal model architecture of certain statistical methods can only be estimated using *either* cross-validation *or* information criteria.

3.7.3 Model Averaging

Instead of selecting a single model, a set of models can be selected which are subsequently averaged [Burnham and Anderson, 2002]. This approach often leads to better explanatory power and better forecasts. Moreover, an uncertainty imposed by the different model realizations can be estimated. This work follows a simple but efficient model averaging procedure summarized by Turkheimer et al. [2003]. Let there be N models coming from a model selection procedure as described in the previous subsections. Depending on the application, there are n models with $n \leq N$ that are used further. Each of these n models gets a weight ω_i assigned to it with

$$\omega_i \in [0,1] \quad \text{and} \quad \sum_{i=1}^{n} \omega_i = 1 \; . \tag{3.43}$$

The simplest assumption would be to set $\omega_i = \frac{1}{n} \quad \forall i = 1, ..., n$. Since the weights in Eq. 3.43 somewhat represent the *performance* of a particular model, the weights may be received from

measures like the cross-validation errors, information criteria, or a correlation coefficient.

The model average Θ is now given by

$$\Theta = \sum_{i=1}^{n} \omega_i \Theta_i \ , \tag{3.44}$$

where Θ_i denotes the response from model i. The uncertainty estimate of this average is given by

$$\sigma_\Theta = \sum_{i=1}^{n} \omega_i \sqrt{\sigma_i^2 + \beta_i^2} \tag{3.45}$$

where $\beta_i = \Theta_i - \Theta$. In Eq. 3.45, σ_i denotes the uncertainty of the response from model i which can in general be estimated using bootstrap (see next section). The strength of Eq. 3.45 is that it incorporates not only the uncertainties (σ_i) of each model but also their response variation (β_i) with respect to the model average.

3.8 Bootstrapping

Bootstrapping [Efron, 1979] is a method aiming at evaluating uncertainties independently from underlying distributions. Assumptions, such as normality, do not have to be made. Its simplicity and general applicability make it a powerful method used in many fields of research [Davison and Hinkley, 1997]. It is based on randomly sampling from a given distribution and evaluating the quantity of interest. Each of these random estimates is called a *bootstrap*. Bootstrapping is considered a Monte Carlo Method. This work makes use of two important bootstrapping procedures reviewed in the following.

3.8.1 Case Resampling

Let there be a random variable X which is distributed according to an arbitrary discrete distribution F of size N_F. Let λ be the quantity of interest calculated from X; λ might be a sample average or correlation coefficient. The non-parametric bootstrap method of *case resampling* [Efron and Tibshirani, 1986] randomly draws N_B independent samples with replacement of size N_F from F. It is non-parametric because no parameters need to be chosen prior to the sampling and no density function is fitted onto F. When sampling with replacement, F is the same for all sampling iterations. It ensures that the samples are independent and identically distributed. For each bootstrap, the quantity of interest $\tilde{\lambda}$ is calculated. Since the bootstraps are normally distributed for many bootstraps (large N_B) for most estimators λ according to the central limit theorem, the bootstrap standard error is given by

$$\sigma_e = \left(\frac{1}{N_\mathrm{B}} \sum_{i=1}^{N_\mathrm{B}} (\tilde{\lambda}_i - \langle \tilde{\lambda} \rangle)^2 \right)^{\frac{1}{2}} \ , \tag{3.46}$$

where $\tilde{\lambda}_i$ denotes one bootstrap and $\langle\tilde{\lambda}\rangle$ the mean of all N_B bootstraps. A typical uncertainty estimate is now given by $\lambda \pm \sigma_e$. If not stated otherwise, a sufficiently large value of $N_\mathrm{B} = 1000$ is used in this work. For unbounded estimators λ, such as the sample average, $\langle\tilde{\lambda}\rangle$ will be zero. For bounded estimators, such as a correlation coefficient, $\langle\tilde{\lambda}\rangle$ will usually denote a noise level not equal to zero.

3.8.2 Gaussian Process Sampling

When dealing with time series models, ordinary bootstrapping, such as case resampling, would destroy the inherent correlation of the time series. Therefore, this work uses the parametric bootstrap method of *Gaussian process sampling* [Kirk and Stumpf, 2009] to estimate the uncertainty of the response y^* of a time series model. To do so, N_B bootstraps are drawn from all n features of the time series model so that

$$\tilde{x}_t^k \sim \mathcal{N}(x_t^k, \sigma^k) \quad \forall t \ , \tag{3.47}$$

where x_t^k denotes the value of feature k at time t, $\tilde{x}_t^k$ the corresponding bootstrap, σ^k the sample standard deviation of x^k, and $\mathcal{N}$ a Gaussian distribution with mean x_t^k and standard deviation σ^k. The time series model is now evaluated N_B times to obtain $\tilde{y^*}_t$ for each time t. The bootstrap standard error is then given by

$$\sigma_{e,t} = \left(\frac{1}{N_\mathrm{B}} \sum_{i=1}^{N_\mathrm{B}} \tilde{y^*}_{i,t}^2 \right)^{\frac{1}{2}} \quad \forall t \ , \tag{3.48}$$

so that a typical uncertainty estimate is $y_t^* \pm \sigma_{e,t}$. Also here and if not stated otherwise, a sufficiently large value of $N_\mathrm{B} = 1000$ is used in this work. Bootstrapping time series models can be seen as testing empirically how sensitive the model response is to changes in the features. Another approach would be to case resample the residual distribution after training and use the resampled residual to train the statistical model. Each training on randomized residuals is then called a bootstrap. As this is computationally expensive for many bootstraps, it has not been done in this work.

3.8.3 Confidence Interval

The confidence interval is computed under the valid assumption that the bootstraps are normally distributed for most estimators according to the central limit theorem. This yields

$$\kappa \pm \sigma_e \Phi^{-1}(\alpha) \ , \tag{3.49}$$

where κ denotes the quantity of interest λ or the response of a time series model y^* at a given time. In Eq. 3.49, σ_e denotes the bootstrap standard error and $\Phi^{-1}(\alpha)$ the probit function at a confidence level of α, e.g., $\Phi^{-1}(0.95) = 1.96$. The probit function is the inverse of the standard

normal cumulative density function defined by

$$\Phi^{-1}(\alpha) = \sqrt{2}\,\mathrm{erf}^{-1}(\alpha)\ , \tag{3.50}$$

where erf^{-1} is the inverse error function implemented in almost any computing library. The error function is given by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt\ . \tag{3.51}$$

3.9 Normalization and Impact of the Features

3.9.1 Normalization

Features generally have different physical units and may tremendously differ in magnitude. Therefore, prior to using features for solving classification or regression problems, the features need to be normalized appropriately, making their magnitudes comparable. There are different approaches to the problem of normalization [e.g., Bishop, 1995]. Let x be a given feature on m training events, two simple and popular methods are given by

1. $$\hat{x}_i = \frac{x_i - \mu_x}{\sigma_x} \qquad \forall i = 1, ..., m\ , \tag{3.52}$$

   where μ_x denotes the mean of x and σ_x its standard deviation. The normalized feature $\hat{x}$ has zero mean and a variance of one.

2. $$\hat{x}_i = 2\frac{x_i - \max_x}{\max_x - \min_x} + 1 \qquad \forall i = 1, ..., m\ , \tag{3.53}$$

   where $\max_x$ denotes the maximum of x and $\min_x$ its minimum. The normalized feature $\hat{x}$ has its maximum at 1 and its minimum at -1.

Which of these methods to be used in practice is mostly arbitrary. However, there are cases for which Eq. 3.53 will lead to unwanted results, such as when a feature contains outliers. It is recommended to try both normalizations and finally select the one leading to the best performance.

3.9.2 Impact

An important part of statistical modeling is the quantification of statistical importance or *impact* of each of the features on the model response. For complex statistical models, the feature to be analyzed is altered and the statistical model is re-evaluated. For this, the model response after training y^* is considered an equilibrium state. The perturbed model response y_k^* is then the result of altering feature k. Hence, the impact I_k of feature k averaged over a considered period of time can be defined by

$$I_k = \sqrt{\mathrm{var}(y^* - y_k^*)}\ , \tag{3.54}$$

Method	LDA	FEM-VARX	MLP	SVM
Nonlinear built-in	no	no	yes	yes
Nonstationary built-in	no	yes	no	no
Multivariate output	no	yes	yes	no
Unique solution	yes	no	no	yes
Regression	yes	yes	yes	yes
Classification	yes	no	yes	yes
Supervised learning	yes	yes	yes	yes
Unsupervised learning	no	yes	yes	yes
Data clustering	no	yes	yes	yes
Time for training (s)	$\mathcal{O}(1)$	$\mathcal{O}(100)$	$\mathcal{O}(100)$	$\mathcal{O}(10)$
Information criteria	yes	yes	yes	no
Cross-validation	yes	no	yes	yes

Table 3.1: Comparison of the different learning approaches. 'built-in' denotes that nonlinearity or nonstationarity are part of the underlying algorithm of the specific method. 'Multivariate output' denotes that the model response can be multi-dimensional. 'Unique solution' denotes that the training leads to results that are independent of the initialization of the training algorithm. 'Regression' and 'Classification' denotes that a model can be used to solve problems with regression and classification, respectively. 'Supervised' and 'unsupervised' learning denotes if a model can be trained in supervised or unsupervised mode. 'Data clustering' denotes if it is possible to cluster a given data set which is typically done using unsupervised learning. 'Time for training' denotes the approximate time in seconds that is needed for one training realization on a typical desktop machine on one core, which may vary depending, e.g., on the number of features or the number of training events. 'Information criteria' denotes if information criteria, such as AIC and BIC, can be used for model selection. 'Cross-validation' denotes if cross-validation can be used for model selection.

where $\mathrm{var}(\cdot)$ denotes the sample variance, y^* the original model response, and y_k^* the perturbed model response for feature k held constant at its median, keeping all other features at their original values as used during training. The median was selected as an average state because it is more robust than the mean when dealing with non-Gaussian distributions that may be largely skewed. The impact, as defined in Eq. 3.54, represents the averaged response deviation from the equilibrium state given by y^*. The uncertainties of this estimate are generally computed with a bootstrap procedure (see Section 3.8). The impact I_k can be simplified for the special case of LDA. From Eq. 3.54 and Eq. 3.6, one obtains the estimate

$$I_k = |\beta_k| \sqrt{\mathrm{var}(x_k)} \,, \tag{3.55}$$

where β_k denotes coefficient k and x_k the corresponding feature (or basis function).

In order to be able to compare impacts throughout different statistical models, the impact needs to be normalized. This is done by dividing by the sum of all impacts to obtain the *relative impact*

$$I_k^{\mathrm{rel}} = \frac{I_k}{\sum_{i=1}^{n} I_i} \,, \tag{3.56}$$

where n denotes the number of all features.

3.10 Comparing the Learning Approaches

This section aims at providing an overview of the similarities and differences of the different learning approaches previously introduced, namely LDA, FEM-VARX, MLP, and SVM. The method of principal component analysis (PCA) is not included here since it cannot be used as a time series model. When necessary, PCA is used for dimension reduction before and after the training of the above learning approaches.

Table 3.9.2 presents important characteristics of the different statistical learning approaches. It should be noted that all considered methods can be used in a linear and stationary way. Aside from comparing methods and their performance for a specific application, this compilation will help to make the initial decision of which method(s) to use. It is observed in Table 3.9.2 that the perfect method is not any of the methods considered here. The perfect statistical method does not exist. However, certain methods are better for certain applications than others. For instance, when it is known that the process to be modeled is nonstationary, then a nonstationary method, such as FEM-VARX, should be used. If the relationships between the features and the process to be modeled are nonlinear, then a nonlinear method, such as MLP, should be used.

There are no modeling examples given in this section because it highly depends on the specific application which method will perform best as observed in the next chapters. There may be even applications for which the simplest of all considered methods, namely LDA, performs most efficiently. In the upcoming chapters, the four statistical learning approaches will be compared with respect to their ability to model and forecast stratospheric variability on different stratospheric domains, variables, and time scales.

Chapter 4

Data

This chapter introduces the data sets that the upcoming statistical investigations are based on. This includes data from observations, reanalyses, and climate model simulations.

4.1 Observations

Making stratospheric observations using radiosondes, satellites, or lidar is essential for understanding the evolution of weather and climate in the stratosphere. The higher the quality and resolution of the observations, the better the past is understood and the higher the chance to make reliable predictions into the future. Stratospheric observations are generally considered as having a higher quality with the introduction of satellite measurements (satellite era) in 1979 [e.g., Randel et al., 2009; Blume et al., 2012]. In this work, observations from different sources for ozone and sea surface temperature are utilized.

Ozone

Two observational data sets for ozone are used in this work:

- The **NIWA-3D** (referred to as NIWA hereafter) ozone data set [Hassler et al., 2009] is a gridded monthly data set based on observations. Where observations were not available, data points were computed by a regression model. NIWA has a horizontal resolution of 5° and 70 pressure levels from approx. 878 hPa to 0.05 hPa. NIWA is available from January 1979 to December 2007.

- The **Randel&Wu** (referred to as RANDEL hereafter) ozone data set [Randel and Wu, 2007] is a gridded monthly data set based on observations and an ozone climatology by Fortuin and Kelder [1998]. Where observations were not available, data points were computed by a regression model. RANDEL has a horizontal resolution of 5° and 31 pressure levels from approx. 300 hPa to 0.1 hPa. RANDEL is available from January 1979 to December 2005.

Sea Surface Temperature

Observations of sea surface temperature (SST) are taken from the **HadISST1** data set [Rayner et al., 2003], available at the Met Office Hadley Centre, UK. HadISST1 is a gridded data set, has a horizontal resolution of 1°, and is available from 1870 to present day.

4.2 Reanalyses

Reanalysis data sets play an important role when trying to understand climate and weather development over the past sixty years. Reanalyses provide world-wide gridded data sets on a high-resolution temporal domain for many different variables (e.g., temperature, wind, geopotential). They are often available up to the stratopause or even higher.

Reanalyses are produced in a data assimilation system [Kalnay, 2003], meaning that available observations are assimilated into a model simulation. Therefore, a reanalysis data set is a *best guess* incorporating observations and model simulations. Naturally, the higher the quality of the observations and the more realistic the model simulation, the better is the reanalysis. The considered reanalyses are generally more reliable during the satellite era from 1979 onwards. Also, the data assimilation scheme (e.g., 4D-Var, Ensemble Kalman) plays a crucial role as it ultimately combines model and observations. There are many different reanalyses from all over the globe. In this work, two European and two American reanalyses are utilized:

- The **ERA-40** (referred to as ERA40 hereafter) reanalysis [Uppala et al., 2005] was computed by the European Centre for Medium-Range Weather Forecasts (ECMWF) for the period of September 1957 to August 2002. It is based on a spectral model with T159L60 and a model top at 0.1 hPa (lower mesosphere). A 3D-Var assimilation system was used for ERA40. In this work, ERA40 is utilized in a horizontal resolution of 1.1° on 23 pressure levels from 1000 to 1 hPa. ERA40 was originally intended to only cover 40 years. ERA stands for **E**CMWF **r**e-**a**nalysis.

- The **ERA-Interim** (referred to as ERAInt hereafter) reanalysis [Simmons et al., 2006] is also being computed by the ECMWF for the period of January 1979 to present day. It is based on a spectral model with T255L60 and a model top at 0.1 hPa (lower mesosphere). ERAInt uses the more powerful 4D-Var assimilation system. In this work, ERAInt is utilized in a horizontal resolution of 0.7° on 31 pressure levels from 1000 to 1 hPa. ERAInt is intended to eventually replace ERA40.

- The **NCEP/NCAR** (referred to as NCEP hereafter) reanalysis I [Kalnay et al., 1996] is being computed by the National Center for Environmental Prediction (NCEP), USA in collaboration with the National Center for Atmospheric Research (NCAR), USA for the period of January 1948 to present day. It is based on a spectral model with T255L28 and a model top at 3 hPa (upper stratosphere). A 3D-Var assimilation system was used for NCEP. In this study, NCEP

is used in a horizontal resolution of approx. 2.5° and 17 pressure levels from 1000 to 10 hPa. There is also the NCEP/NCAR reanalysis II beginning in 1979 with higher horizontal and vertical resolution (not used in this study).

- The **MERRA** reanalysis [Rienecker et al., 2011] is being computed by the National Aeronautics and Space Administration (NASA), USA for the period of January 1979 to present day. The grid used for MERRA is 0.5° (lat) x 2/3° (lon) with 72 vertical levels up to 0.01 hPa (mesosphere). A 3D-Var assimilation system was used for MERRA. In this study, MERRA is used in a horizontal resolution of approx. 0.5° and 42 pressure levels from 1000 to 0.1 hPa. MERRA stands for **M**odern **E**ra **R**etrospective Analysis for **R**esearch and **A**pplications.

There are a few studies carried out in the past that systematically compared reanalyses and observational data sets for the stratosphere. Randel et al. [2004] compared middle atmosphere climatologies in temperature and zonal winds in observations and reanalyses and found significant differences for temperatures near the tropical tropopause and in the polar lower stratosphere, temperatures near the global stratopause, and zonal winds throughout the tropics. More differences appeared when comparing oscillations in zonal wind such as the tropical semiannual oscillation and the quasi- biennial oscillation. Labitzke and Kunze [2005] compared stratospheric north-polar temperature trends and found a good agreement between historical hand-analyzes prepared at FU Berlin, ERA40, and NCEP. Randel et al. [2009] show that stratospheric trends in temperature and ozone agree well between different observational data sets during the satellite era (1979-2007) but deviate for the pre-satellite era (1958-1978). Studies carried out by Charlton and Polvani [2007] and Blume et al. [2012] compared stratospheric warming climatologies between ERA40 and NCEP and found very similar monthly frequency distributions but significant differences during the pre-satellite era.

4.3 Chemistry-Climate Models

During the Chemistry-Climate Model Validation Activity 2 (CCMVal-2) [SPARC CCMVal, 2010] instigated by the SPARC initiative and the World Climate Research Program (WCRP), numerous chemistry-climate models (CCMs) simulated the recent past and the future climate while incorporating common forcing fields (e.g., SSTs, solar cycle, greenhouse gases, etc.). The CCMVal website (`http://www.pa.op.dlr.de/CCMVal`) states that "the goal of CCMVal is to improve understanding of CCMs and their underlying GCMs (General Circulation Models) through process-oriented evaluation, along with discussion and coordinated analysis of science results". The output of the different model simulations is stored on the British Atmospheric Data Centre (BADC) which can be accessed by registered users. This work makes use of several simulation runs of three different CCMs [Morgenstern et al., 2010] which participated in CCMVal-2 (MRI, WACCM) or with which equally constrained runs were performed (EMAC-FUB):

- **EMAC-FUB** - The **E**CHAM5/**M**ESSy **A**tmospheric **C**hemistry model [Roeckner et al., 2003, 2004; Jöckel et al., 2006] is a spectral chemistry-climate model, which is simply referred to

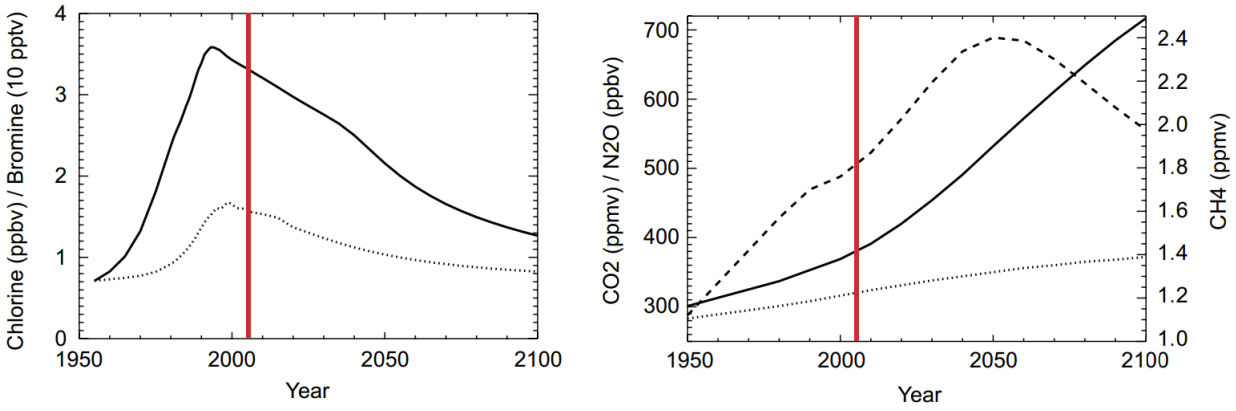

Figure 4.1: *Left:* Surface total chlorine (solid) and total bromine (dotted) as defined in the A1 scenario [WMO, 2007]. *Right:* Surface CO_2 (solid), N_2O (dotted), and CH_4 (dashed) as defined in the A1b scenario [IPCC, 2001]. Figure from Chapter 2 of SPARC CCMVal [2010].

as **EMAC** in the following[1]. It was run with a vertical resolution of 39 layers (L39) up to 0.01 hPa (≈80 km) and a horizontal resolution of 2.8° (T42). The model results are available on a Gaussian grid in a resolution of 2.8° and 31 pressure levels from 1000 to 0.1 hPa. The QBO in EMAC is nudged in a latitude range from 7° S to 7° N and from 70 to 12 hPa. Nudging refers to the gentle pushing of simulated values towards observed values. For the QBO, this is done with a weighting function decreasing exponentially away from the equator. Nudging is necessary when important modes of variability cannot be generated internally.

- **MRI** - The **M**eteorological **R**esearch **I**nstitute model [Shibata et al., 1999, 2005; Shibata and Deushi, 2008] is a spectral chemistry-climate model. It was run with a vertical resolution of 68 layers (L68) up to 0.01 hPa (≈80 km) and a horizontal resolution of 2.8° (T42). The model results are available on a Gaussian grid in a horizontal resolution of 2.8° and 31 pressure levels from 1000 to 0.1 hPa. The QBO in MRI is self-consistent and internally generated. This is possible due to the dense vertical layering in MRI.

- **WACCM** - The **W**hole **A**tmosphere **C**ommunity **C**limate **M**odel [Garcia et al., 2007; Kinnison et al., 2007; Richter et al., 2010], version 3.5, is a finite-volume chemistry-climate model. It was run with a vertical resolution of 66 layers up to $5.96 \cdot 10^{-6}$ hPa (≈140 km) and a horizontal resolution of 1.9°(lat) x 2.5°(lon). The model results are available on a Gaussian grid with 31 pressure levels from 1000 to 0.1 hPa. The QBO in WACCM is nudged in a latitude range from 22° S to 22° N and from 90 to 3 hPa [Matthes et al., 2010].

Using these CCMs, the different research groups performed reference simulations from January 1960 to December 2005 (REF-B1) as well as future projections from 1960 to 2100 (REF-B2)

[1] The name EMAC-FUB is historically grown and used for the L39 version as run at the Institute for Meteorology, Freie Universität Berlin (FUB). There is also a L90 version developed and run at the Max-Planck-Institute for Chemistry in Mainz, Germany, generally referred to as EMAC.

that are analyzed in this work. The reference runs, REF-B1 and REF-B2, are the most realistic simulations performed during the CCMVal-2 initiative [Eyring et al., 2008]. The REF-B1 runs were constrained with observed forcings in SSTs, solar cycle, volcanic eruptions, greenhouse gas emissions, and ozone depleting substances. The REF-B2 runs were forced with simulated SSTs, constant solar irradiance (except EMAC where observed solar cycles where repeated into the future) and no volcanic eruptions. The QBO was repeated into the future for CCMs nudging the QBO (EMAC, WACCM). The greenhouse gas (GHG) emissions were prescribed according to the A1b scenario [IPCC, 2001] and ozone depleting substances (ODS) according to the A1 scenario [WMO, 2007]. The most important GHGs and ODSs, as used for the aforementioned reference simulations, are shown in Fig. 4.1 for the time period of 1950 to 2100. Chlorine and bromine are shown on the left, carbon dioxide, nitrous oxide, and methane on the right. This figure is taken from Chapter 2 of SPARC CCMVal [2010] where it states: "The ODSs increase sharply during the 1970s and 1980s, resulting in an approximate 6-folding of organic chlorine and a doubling of organic bromine at peak abundances in the 1990, relative to pre-industrial times. For the 21st century, a continuous decline, in accordance with the Montreal Protocol, is anticipated. The decline is substantially slower than the increase in the 20th century. By contrast, for the leading greenhouse gas CO_2 a steady increase is anticipated, leading to a more than doubling by 2100, compared to 1950. N_2O follows a similar trend, albeit with smaller growth rates. CH_4, by contrast, is anticipated to undergo a trend reversal around 2050." The red bar in Fig. 4.1 separates the REF-B1 to the left from the REF-B2 period to the right. REF-B2 actually denotes the full period of 1960 to 2100 but only REF-B2 results from 2006 onward are considered in this work. In the following, the large-scale stratospheric variabilities in temperature and ozone in the CCM simulations and the previously introduced reanalyses and observations are statistically modeled and the natural contribution is quantified.

Chapter 5

Quantification of Natural Variability

The statistical methods introduced in Chapter 3 are applied to the large-scale statistical modeling of the global stratosphere on a zonal mean domain. Considering zonally averaged data, i.e. averaging along each latitude across all longitudes, is a good approximation because most systematic stratospheric variations are in latitude and height [e.g., Haynes, 2005]. The statistical methods considered in this regression study are linear discriminant analysis (LDA), a cluster method based on finite elements (FEM-VARX), a neural network, namely the multi-layer perceptron (MLP), and support vector regression (SVR). LDA is equivalent to multiple linear regression for regression problems.

The variables to be modeled are monthly temperature and ozone because of their specific importance for indicating variability in the radiative budget, dynamics, and chemistry (see Chapter 2). All temperature and ozone time series considered in this study are deseasoned, i.e. anomalies with respect to their long-term mean. Once variability is modeled, the impact of forcing factors is calculated. From this, the statistical importance of forcing factors originating naturally, i.e. natural variability, is estimated. The impact of anthropogenic forcings on recent climate in the stratosphere can then be estimated. Considering various data sets along with the four independent statistical methods makes it feasible to estimate robust estimates and uncertainties for the impacts of each of the forcing factors that influence stratospheric variability.

The statistical methods are trained with data representing the recent climate for the time period from 01/1980 to 12/2005 which is the longest common period among the different data sets. Selecting this period ensures comparability between different data sets and garanties the highest available quality of observations and reanalyses due to the introduction of satellites in 1979 (satellite era). The reanalyses and observations considered in this chapter are ERAInt, MERRA, RANDEL, and NIWA (see Sections 4.1, 4.2). The SPARC CCMVal-2 model simulations considered are EMAC-FUB, WACCM, and MRI where statistical training is performed with the REF-B1 runs (see Section 4.3). In addition, long-range statistical forecasts up to the year 2100 are made with reasonable assumptions about the forcing factors. These statistical projections are compared with simulated projections within CCMVal-2, namely the REF-B2 runs (see Section 4.3).

The statistical methods are trained by incorporating a range of forcings and modes of variability that are known to influence the stratosphere. These so-called external factors, as described in

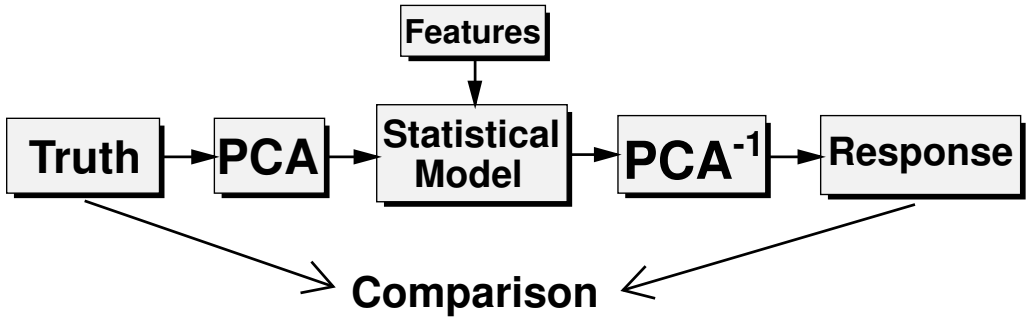

Figure 5.1: A schematic picture showing the principal of dimension reduction in combination with a statistical model incorporating features (external factors). The high-dimensional truth is decomposed by a PCA and only the first n principal components are retained, explaining at least 90% of the overall variance. See text for explanation.

Chapter 2, are variabilities in sea surface temperature (ENSO, AMO, NPO/WP, PDO), in equatorial zonal mean zonal wind (QBO1, QBO2), the annular modes (NAM, SAM), the solar cycle (SFL), and volcanoes (AOD). In addition, a sine and a cosine with a period of one year are included, representing the seasonal dependency of the stratosphere. For the temperature modeling, a linear trend term (TRE) is included, whereas for ozone, ESC is used to represent long-term behaviour (see Section 2.3). During selection and computation of the external factors, as described in Chapter 2, it was ensured that none of the hemispheres were favored a priori by, e.g., using principal components to obtain suitable indices. In order to make the different external factors comparable in magnitude, all indices are normalized between -1 and $+1$ according to Eq. 3.53. Factors for sea surface temperatures, equatorial zonal mean zonal winds, and annular modes were detrended before being presented to a statistical model.

As one part of this chapter, the four different statistical methods are compared with respect to their regression performance. Prior to the actual training, the optimal model architectures need to be determined with the help of information criteria and cross-validation (see Section 3.7). Due to the naturally high-dimensional space (latidude, level, time), an efficient procedure for dimension reduction is applied in a first step, namely principal component analysis (see Section 3.1), while still retaining almost all variance found in the data.

5.1 Dimension Reduction

This analysis consideres the stratosphere on a zonally averaged domain. Therefore, the variables depend on latitude, pressure level, and time. All latitudes are considered from 90° S to 90° N and 11 pressure levels at 100, 70, 50, 30, 20, 10, 7, 5, 3, 2, and 1 hPa. Hence, with a typical horizontal resolution of 2° the spatial dimension has a size of approximately $90 \times 11 \approx 1000$. Hence, every statistical method would have to model 1000 dimensions. Assuming that the training for one dimension takes about a minute and considering that this work deals with many data sets and four statistical

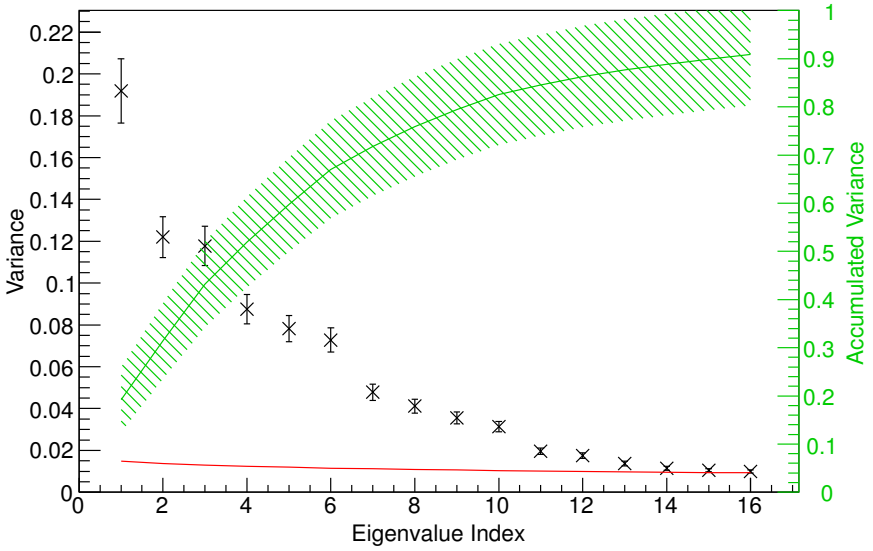

Figure 5.2: The result of a PCA calculated for temperature in ERAInt. This figure displays the variance (black) along with the accumulated variance (green) explained by the first 16 eigenvalues and corresponding EOFs. The error bars and the hatched area denotes the standard sampling error calculated with North's rule of thumb (see Eq. 3.3). The graph in red denotes the level above which an eigenvalue can be considered as not being caused by a noise process at a 99% confidence, estimated with bootstrap based on case resampling (see Section 3.8).

methods, one realizes that analyzing the entire raw domain is computationally time consuming. Therefore, principal component analysis (see Section 3.1) is used to reduce dimensionality, prior to the actual statistical modeling. This is displayed schematically in Fig. 5.1. The high-dimensional raw domain (truth) is decomposed by a PCA and only the first n principal components (PCs) are retained, explaining at least 90% of the overall variance. These n PCs are then modeled by the statistical model with the help of the external factors (features). The result is then transformed back by inverting the PCA (PCA^{-1}; see Eq. 3.5). The resulting response can be directly compared to the truth.

The idea of dimension reduction is exemplarily presented for the temperature in ERAInt but performed on all data sets considered in this chapter. Fig. 5.2 displays the variance (black) along with the accumulated variance (green; see Eq. 3.4) explained by the first 16 eigenvalues and corresponding EOFs. The error bars and the hatched area denotes the standard sampling error calculated with North's rule of thumb (see Eq. 3.3). It is observed that already 16 EOFs explain more than 90% of the overall variance. All 16 eigenvalues are significant on a 99% confidence level with respect to a noise process simulated by bootstrap based on case resampling as described in Section 3.8.

With a spatial dimension size of 704 (#latitudes x #levels) in ERAInt, the number of dimensions are therefore reduced by a factor of $704/16 = 44$ (!), only loosing about 10% of the overall variance. Fig. 5.3 shows the first two EOFs of this PCA, representing the large variability in the north-polar stratosphere and its connection to the equatorial stratosphere. The first EOF repre-

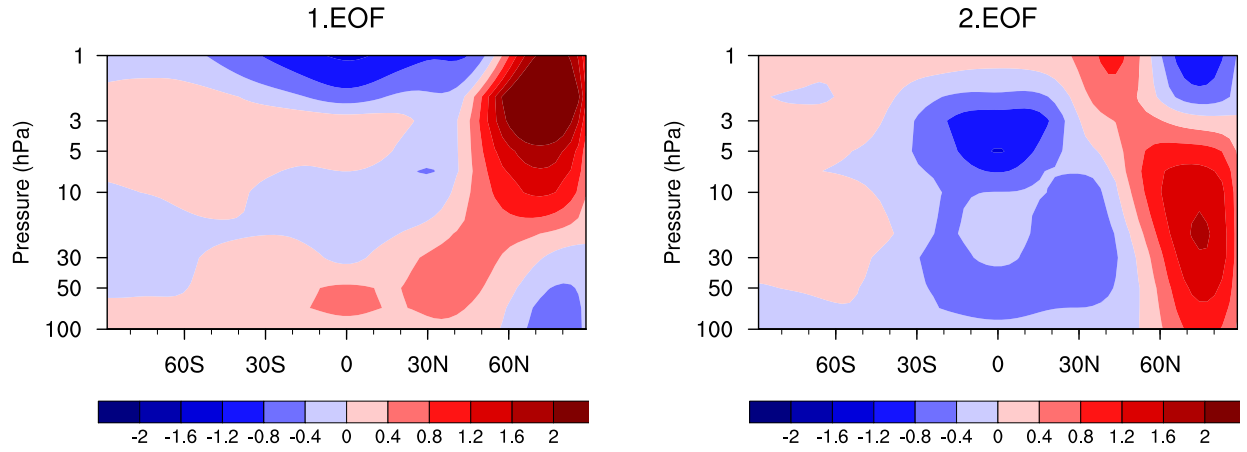

Figure 5.3: The two leading EOFs in Kelvin of a PCA calculated from temperature anomalies in ERAInt. Both represent the large variability in the north-polar stratosphere and its teleconnection to the equatorial stratosphere.

sents how positive (negative) temperature anomalies in the polar upper stratosphere (north-polar variability) coincide with negative (positive) anomalies at the equatorial stratopause (semiannual oscillation). The second EOF represents how positive (negative) anomalies in the polar middle and lower stratosphere (polar vortex) coincide with negative (positive) anomalies over the equator in the upper and lower stratosphere. Both EOFs represent the fact that warm/cold polar vortex conditions result in an opposite temperature anomaly in the tropics [e.g., Kodera, 2006; Yoshida and Yamazaki, 2011]. The first two leading EOFs have centers of action in the northern hemisphere due to the much higher variability there compared to the southern hemisphere (compare Chapter 2). On one hand, it can be seen how PCA is used for dimension reduction, on the other hand, it is used to obtain insights into high-dimensional data sets.

Naturally, it is interesting to investigate how much variance is retained regionally using the PCA dimension reduction. This is calculated by simply omitting the statistical model as displayed in Fig. 5.1 and directly comparing truth and response. Fig. 5.4 shows the resulting explained variance (EV) according to Eq. 3.36. It is observed that EV is generally larger than 70% except for the low-latitude tropopause where higher-order EOFs play a larger role. For the most part, EV is even larger than 80%, and larger than 90% for the polar regions and the equatorial stratosphere. There is slightly more variance explained on the northern hemisphere due to the higher variability present in this region (see Section 2.1). When averaging the entire EV field, a resulting value of 84% is well within the uncertainty range of the accumulated variance at the 16th EOF in Fig. 5.2. It can be concluded that the dimension reduction is very efficient while retaining most variance. The regional explained variance is similar for ozone and the other data sets (not shown). The number of retained EOFs for temperature, ozone, and the different data sets are given in Table 5.1. For temperature, the CCMs underestimate variability in the stratosphere represented by the lower number of retained EOFs in comparison to ERAInt and MERRA. For ozone, the number of EOFs are similar between observations and CCM simulations, except for WACCM which may overestimate variability. In the

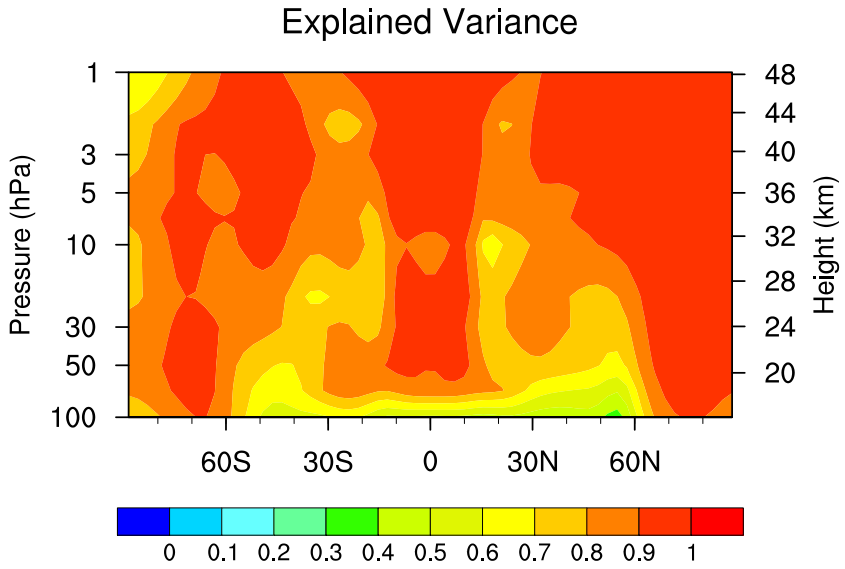

Figure 5.4: The explained variance (EV) according to Eq. 3.36 as a result of a PCA and a subsequent back-transformation while retaining 90% of the overall variance calculated for ERAInt.

Table 5.1: Number of retained EOFs for temperature, ozone, and the different data sets.

	ERAInt	MERRA	RANDEL	NIWA	EMAC	WACCM	MRI
T	16	15	n/a	n/a	10	10	11
O_3	n/a	n/a	9	7	10	17	9

following, the reduced representation, as presented in Table 5.1, is used to compute optimal model architectures.

5.2 Optimal Model Architecture

Each of the statistical methods considered in this chapter, except LDA, has free tuning parameters. A full set of these parameters is called the model architecture (see Section 3.7). In order to determine the optimal model architecture, a k-fold cross-validation is carried out (see Section 3.7.2) for MLP and SVR. Cross-validation is used because a long-term forecast is conducted for each statistical model at the end of this chapter. A value of $k = 5$ is selected to ensure that there is sufficient training data (20 years) and a forecast period on interannual scales (5 years). As described in Section 3.7, the optimal FEM-VARX architecture is determined with the help of information criteria. FEM-VARX is referred to as FVX in this chapter. FVX and MLP are used in a multivariate mode, meaning that their target object is a vector carrying all target dimensions (the EOFs given in Table 5.1). SVR and LDA can only be used in univariate mode, meaning that a separate training for each target dimension needs to be performed (compare Table 3.9.2).

The procedure of estimating the optimal model architecture is exemplarily presented for the

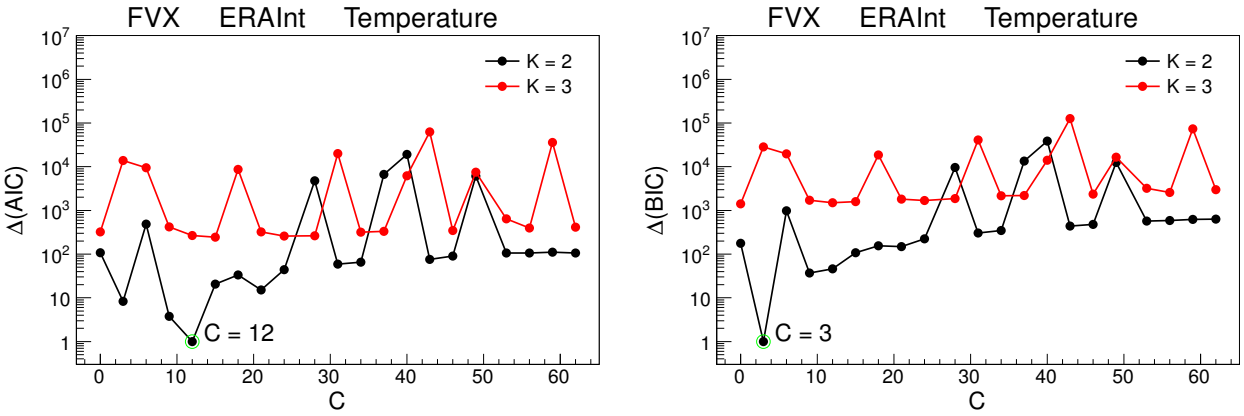

Figure 5.5: FVX information criteria received from the temperature in ERAInt. Displayed are $\Delta(\text{AIC})$ and $\Delta(\text{BIC})$, respectively, calculated for two clusters K and a varying persistency threshold C.

temperature in ERAInt. Fig. 5.5 shows $\Delta(\text{AIC})$ and $\Delta(\text{BIC})$ where $\Delta(\cdot)$ denotes a simple function, called *information difference*, that subtracts the global minimum minus one from each value. This was calculated for FVX for $K = 1, 2$ and a varying number of the persistency thresholds C. Please note that FVX with $K = 1$ would be equivalent to LDA. The smallest Δ across all combinations then points to the optimal model architecture (compare Section 3.7). Therefore, both criteria, AIC and BIC, clearly favor two clusters. Also higher numbers of clusters were tested which were consistently rejected by the information criteria (not shown).

AIC and BIC select $K = 2$ but the results differ for the persistency threshold C. The AIC result is given more trust because $C = 12$ (see Eq. 3.14) allows clusters to form on the 25-year time series that resemble the QBO which is one of the largest sources of natural variability in the stratosphere (see Section 2.2.1). In contrast, with $C = 3$, as favored by the BIC, there would be only very few transitions possible between clusters, making it impossible to cluster QBO variability and its influence. Following equal reasoning, all FVX results presented in this chapter are based on AIC results.

Fig. 5.6 presents the 5-fold cross-validation result for MLP and SVR received from the temperature in ERAInt, showing the mean prediction error (MPE) in Kelvin (see Eq. 3.30). The smaller the MPE the better the cross-validation result. Hence, the minimum value denotes the optimal model architecture. For MLP, 8 neurons in the first and 0 neurons in the second hidden layer are optimal. Therefore, the cross-validation selects a network with only one hidden layer in this case. Please note that the white areas in the MLP plot denote regions where the training algorithm failed (certain number of neurons) or where the number of parameters exceeded the number of steps in time (to avoid overfitting). There is one other local minimum found for MLP with 7 neurons in the second hidden layer as observed in Fig. 5.6. However, the number of free parameters according to Eq. 3.22 is much larger than that of the global minimum and is therefore rejected. For SVR, $C = 10^2$ and $\gamma = 3 \cdot 10^{-2}$ is the optimal setting found by the cross-validation for the temperature in ERAInt.

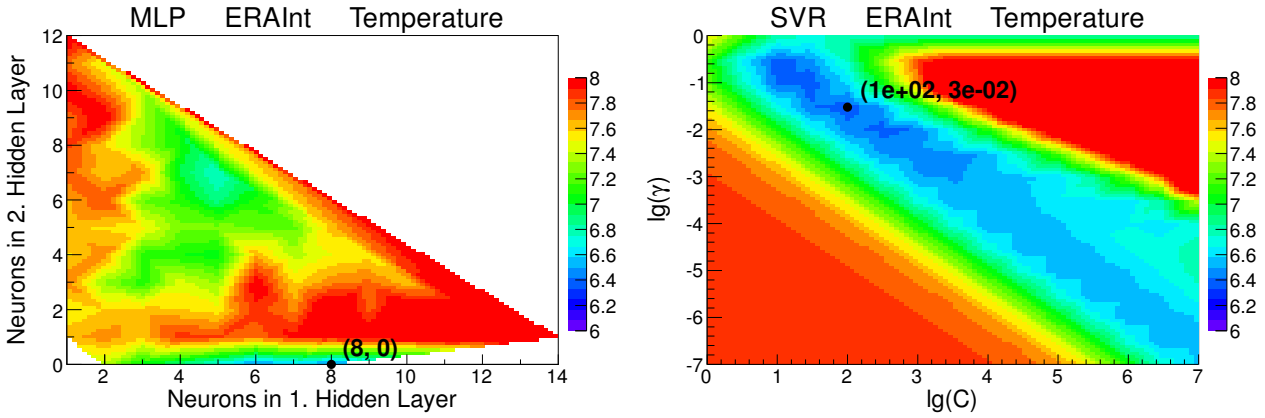

Figure 5.6: Cross-validation result for MLP (left) and SVR (right) received from the temperature in ERAInt, showing the mean prediction error in Kelvin (see Eq. 3.30). The optimal pairs of parameters are indicated.

Table 5.2: The optimal model architectures for FVX (K, C), MLP (L_1, L_2), and SVR (C, γ) for each data set. Results are shown separately for temperature (T) on the left and ozone (O_3) on the right. Not applicable where left blank.

	FVX	MLP	SVR	**T**	FVX	MLP	SVR	**O_3**
ERAInt	$2, 12$	$8, 0$	$10^2, 3 \cdot 10^{-2}$					
MERRA	$2, 9$	$5, 0$	$3 \cdot 10^3, 3 \cdot 10^{-3}$					
RANDEL					$2, 28$	$8, 0$	$10^5, 3 \cdot 10^{-2}$	
NIWA					$2, 6$	$6, 7$	$10^4, 10^{-1}$	
EMAC	$2, 21$	$5, 0$	$10^4, 3 \cdot 10^{-3}$		$2, 37$	$6, 6$	$10^5, 10^{-2}$	
WACCM	$2, 12$	$6, 0$	$10^3, 10^{-2}$		$2, 9$	$5, 7$	$10^4, 10^{-1}$	
MRI	$2, 21$	$5, 0$	$10^4, 3 \cdot 10^{-3}$		$2, 3$	$5, 0$	$3 \cdot 10^4, 3 \cdot 10^{-2}$	

There is actually a large space of parameters C and γ in the neighborhood of the optimal setting that lead to similar cross-validation results, making SVR a robust model for this application.

The model selection is now extended to all other data sets. Table 5.2 summarizes the optimal model architectures for FVX (K, C), MLP (L_1, L_2), and SVR (C, γ) for each data set. The results are shown separately for temperature (T) and ozone (O_3). The model settings for temperature are generally similar among the different data sets. Two clusters in FVX are favored with persistencies C between 9 and 21. One hidden layer in MLP is selected with number of neurons between 5 and 9. Also, selected settings for SVR are relatively similar and well within the range of low cross-validation errors as displayed on the right panel of Fig. 5.6. For ozone, the selected settings lead to slightly more free parameters to be trained with FVX and MLP and a larger variation among the different data sets. FVX still favores two clusters but now with persistencies C between 3 and 37. It is interesting to note that ozone variability in RANDEL and EMAC leads to a significantly higher C than variability in NIWA, WACCM, and MRI where the presistencies for temperature are not far apart. The larger the percistency C the more transitions between clusters are possible, implying

that the statistical model (FVX) becomes more complex (more parameters). However, It is difficult to conclude anything from the value of C about the variability found in the different data sets due to the many dimensions considered simultaneously. Also interesting to note is that ozone variability in NIWA, EMAC, and WACCM leads to an MLP with two hidden layers (more parameters) whereas variability in RANDEL and MRI leads to only one hidden layer (less parameters). The estimated optimal model architectures, as displayed in Table 5.2, are now used to perform the final training with each of the statistical models. The models are then evaluated and compared with respect to their regression performance.

5.3 Regression Performance

The regression performance of each statistical model is assessed with the help of the explained variance (EV) as defined in Eq. 3.36. A value of EV of around 30% or higher is an acceptable value as it corresponds to an approximate correlation coefficient between statistical model and truth of 0.5 (EV $\approx R^2$). This is shown exemplarily for one reanalysis, one observational data set, and one CCM. These representative data sets are ERAInt, RANDEL, and EMAC in Fig. 5.7 for temperature and in Fig. 5.8 for ozone for three different latitude ranges: The tropical (20° S-20° N), the midlatitude (20° N-60° N), and the polar region (60° N-90° N). Performance results for the southern hemisphere are similar and omitted for simplicity.

Concerning the temperature results, it is seen that the tropical regression performance is acceptable for most parts of the stratosphere, particularly for EMAC and the other CCMs (not shown). For ERAInt, a drop of performance in the upper stratosphere higher than 5 hPa is observed, also obtained for MERRA (not shown). A decrease of performance is also observable towards the lower stratosphere below 50 hPa which is strongest in midlatitudes. This is mainly an artifact of the dimension reduction, also visible in the ozone results in Fig. 5.8. Low values of explained variance are observed in Fig. 5.4 from 100 to 50 hPa without using a statistical model, particularly in the extratropics. There are interesting performance differences between the statistical models in Fig. 5.7. For ERAInt, MLP performs the worst and SVR the best. However, for the polar region there is a region between 20 and 3 hPa where FVX performs the best. There is no significant difference between the different statistical models in the polar stratosphere in EMAC and the other CCMs (not shown). For the tropics and midlatitudes, SVR performs best closely followed by MLP. FVX and LDA perform very similar for the temperature in EMAC. For the modeling of temperature, it is generally observed that LDA does not perform significantly worse than MLP or FVX. Only SVR is significantly better than the other methods throughout almost all regions and pressure levels.

The ozone modeling performance, as displayed in Fig. 5.8, is generally higher than that for temperature, especially for RANDEL across all considered pressure levels. This is also true for NIWA. For all statistical models and most part of the stratosphere, it is even above 60% EV corresponding to a correlation of approx. 0.8 between statistical model and truth. Again, SVR performs significantly better than any of the other methods. MLP and FVX perform very similar over all latitude ranges

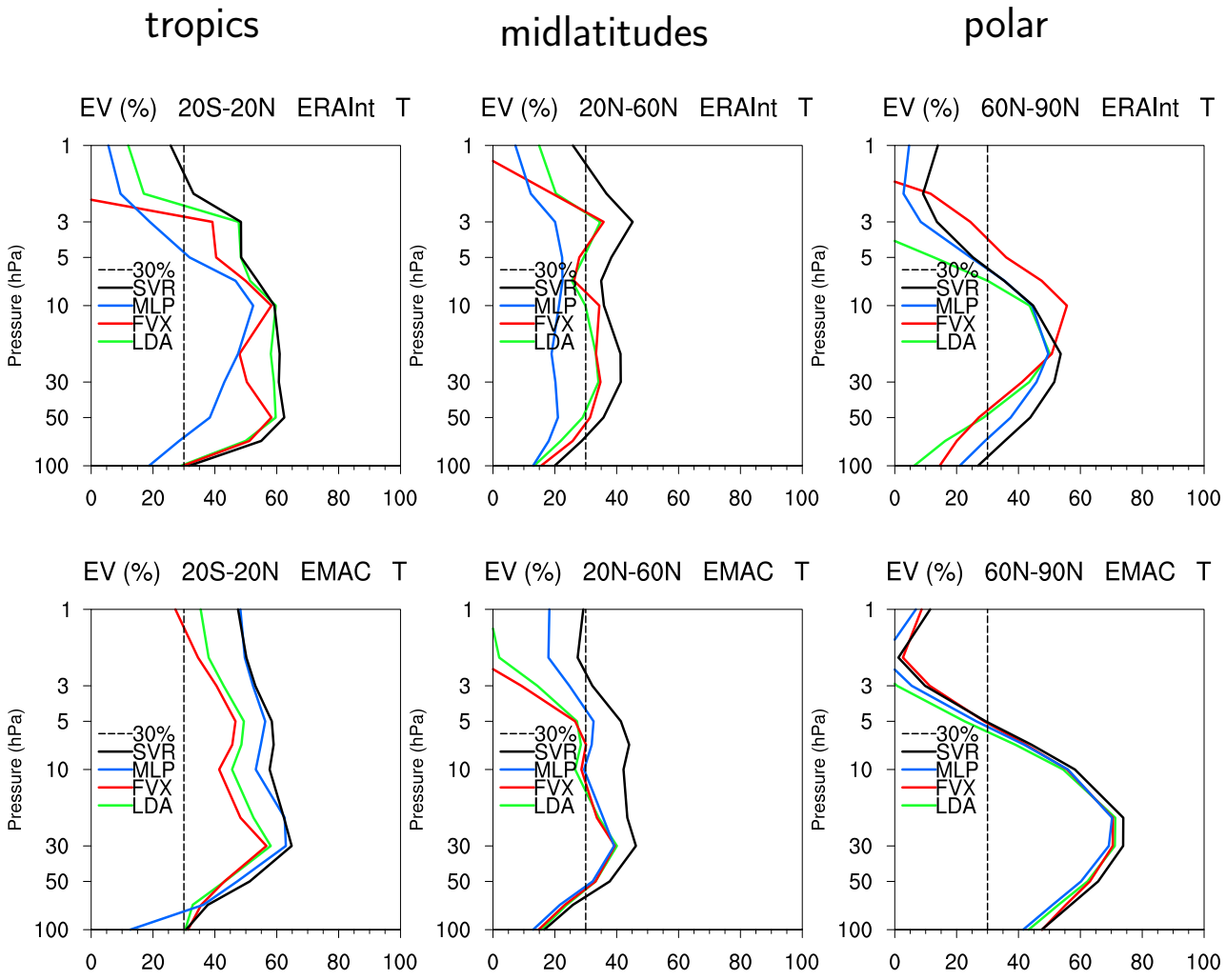

Figure 5.7: Explained variance (EV) for temperature for ERAInt (top) and EMAC (bottom) for different latitude regions. The considered latitude ranges from left to right are: The tropical (20° S-20° N), the midlatitude (20° N-60° N), and the polar region (60° N-90° N). The higher EV, the better the regression performance. Values of EV of at least 30% are acceptable as they correspond to an approximate correlation of 0.5 between statistical model and truth.

and pressure levels, except for the polar region in EMAC where MLP even drops below 30% EV. LDA clearly performs the worst for this application. Similar to the temperature results, EV in EMAC and the other CCMs (not shown) show a clear decrease in regression performance towards lower stratospheric levels (compare Fig. 5.7). The regression performance is generally lower at the upper stratosphere as observed for ozone and temperature. This is not an artifact of the dimension reduction but a result of insufficient information from external factors about the variability in this region. To achieve better modeling results at these pressure levels, other external factors governing the variability in this region, such as the semiannual oscillation, would have to be included.

The regression performance is summarized in Table 5.3 for temperature, ozone, all data sets, and the different statistical models. Displayed is the explained variance (EV) in percent averaged over the entire stratosphere. Results are shown separately for temperature (T) and ozone (O_3).

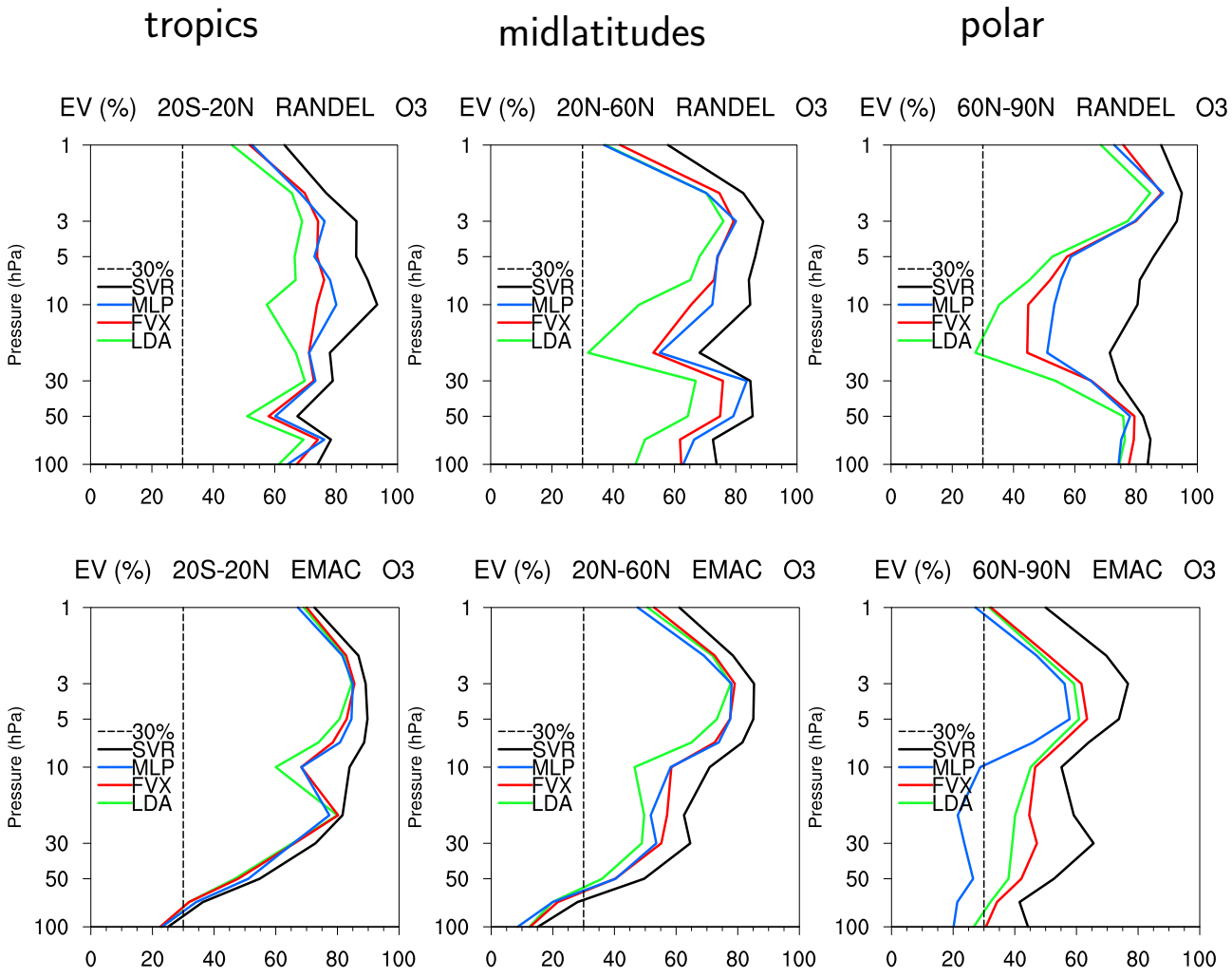

Figure 5.8: Explained variance (EV) for ozone for ERAInt (top) and EMAC (bottom) for different latitude regions. The considered latitude ranges from left to right are: The tropical (20° S-20° N), the midlatitude (20° N-60° N), and the polar region (60° N-90° N). The higher EV, the better the regression performance. Values of EV of at least 30% are acceptable as they correspond to an approximate correlation of 0.5 between statistical model and truth.

Concerning temperature, it is seen that LDA and FVX perform very similar for the two reanalyses ERAInt and MERRA, explaining about 27% of the variance. This is also true for the performance of MLP for MERRA. For ERAInt, MLP only explains about 23% variance. SVR clearly performs best for the reanalyses with a value of about 38%. For the CCMs, SVR also explains around 40% whereas the other methods perform significantly worse. In particular, LDA and FVX for MRI explain only 14% and 18%, respectively. Also, MLP is not much better with 21%. The reason for this poor performance might have to do with the internally generated QBO in MRI and more complex teleconnections with higher latitudes [Naoe and Shibata, 2010]. Concerning ozone, a very similar ranking among the statistical models is observed with SVR explaining most variance. As already seen in Fig. 5.8, EV is generally higher for ozone than it is for temperature. This is due to the large impact of stratospheric chlorine and bromine expressed by ESC compared to the smaller impact of

Table 5.3: Regression performance: The explained variance (EV) in percent averaged over the entire stratosphere for LDA, FVX, MLP, and SVR, presented for each data set. Results are shown separately for temperature (T) on the left and ozone (O_3) on the right. Not applicable where left blank. The mean across the data sets is given at the bottom along with its standard error ($\sigma/\sqrt{n}$).

	LDA	FVX	MLP	SVR	**T**	LDA	FVX	MLP	SVR	**O_3**
ERAInt	29	28	23	39						
MERRA	27	26	28	37						
RANDEL						55	64	65	79	
NIWA						53	63	68	81	
EMAC	30	31	35	42		50	53	48	63	
WACCM	29	26	23	44		49	54	46	78	
MRI	14	18	21	35		65	68	63	77	
Mean	26±3	26±2	26±2	39±1		54±3	60±3	58±4	76±3	

the linear trend term in the modeling of temperature. The next section will address this issue in more detail. It is interesting to note that the regression performance for ozone variability is the lowest in EMAC and WACCM across all statistical models. The reason for this might be an important, currently not included external factor such as the semiannual oscillation.

To obtain a final comparison between the different statistical models, the mean across the data sets is given at the bottom of Table 5.3 together with its standard error ($\sigma/\sqrt{n}$). For temperature, LDA, FVX, and MLP lead to equal performances with small error bars. SVR outperforms all other methods by about a factor of 1.5 with a mean EV of 39%. For ozone, SVR again performs much better than the other statistical models, outperforming them by a factor of about 1.3 with a mean EV of 76%. FVX and MLP perform very similar for ozone with approx. 59% EV. The mean EV value of 54% for LDA is smaller than that of FVX and MLP, but the uncertainty ranges overlap which makes it difficult to rank the performances differently. It can be concluded that SVR clearly explains significantly more variance when statistically modeling monthly temperature and ozone anomalies in the stratosphere. The performances of LDA, FVX, and MLP are not significantly different from one another for this particular application. It is investigated in the next section if the different regression results lead to different statistical impacts of the external factors.

5.4 Estimating the Statistical Importance

The statistical impact of the external factors on each of the statistical models and the variability in different data sets is investigated. This is done with the help of the relative impact as defined in Eq. 3.56. This impact depends on the variable (temperature or ozone) and the particular data set that the statistical model was trained with. Having the impact calculated, the natural climate variability inherent in the stratosphere is quantified from averaged statistical impacts across statistical models and the different data sets. This makes it possible to assign a robust uncertainty to the statistical impacts of the external factors on the different variables. In addition, it is shown where

Table 5.4: Weights for each statistical model used for the model averaging. They are derived from the individual explained variances for temperature and ozone. See text for explanation.

	LDA	FVX	MLP	SVR
T	0.22	0.22	0.22	0.33
O_3	0.22	0.24	0.23	0.31

the external factors have largest impacts in the stratosphere.

5.4.1 Individual Impacts

The individual impact of each external factor on a certain statistical model and data set is averaged across the stratosphere and displayed in Fig. 5.9 for temperature and in Fig. 5.10 for ozone. SST denotes the sum of the four SST factors, QBO the sum of the two factors QBO1 and QBO2, SFL the solar cycle, AOD the aerosol optical depth, AM the sum of NAM and SAM, SI the seasonal influence as the sum of sine and cosine term, TRE the linear trend term, and ESC the effective stratospheric chlorine (compare Chapter 2). The factors in sea surface temperature were added up due to their significant interrelationships [e.g., Deser et al., 2010]. The QBO factors were added up because they are two indices describing the same height-evolving phenomena in the equatorial stratosphere (see Section 2.2.1 for a description of the QBO). When averaging NAM and SAM impacts across the stratosphere, their impacts are almost identical in all here considered data sets. Hence, NAM and SAM contribute equally to the AM impact. Of course, all external factors influence each other to some extent which cannot be avoided within a statistical analysis. However, the grouping here was done according to the origin of the external factors, making it physically plausible.

The results are displayed for each statistical method with the corresponding standard error. The statistical impacts of each factor on the different methods are averaged using the model averaging procedure described in Section 3.7.3. The average (Mean), as displayed in Fig. 5.9 and Fig. 5.10, is weighted according to the individual explained variances (EV) presented in the previous subsection. The corresponding weights are summarized in Table 5.4, giving SVR the largest weight as it explains more variance than the other models. To obtain these weights, EV for a specific model is divided by the sum of all EVs. By doing so, it is assumed that the statistical model explaining more variance is able to more effectively weight the external factors according to their statistical importance.

Temperature

When examining the ERAint results in Fig. 5.9, it is noted that AM has the largest impact of about 25% where half of this value is attributed to NAM and SAM, respectively. Whereas AM impacts in ERAInt computed with SVR and MLP are close to the mean, the impact from LDA is 30% whereas that from FVX is the lowest with 20%. This deviation among the statistical models is reflected by a standard error of the mean of around 3%. The mean AM impacts for the other data sets are slightly

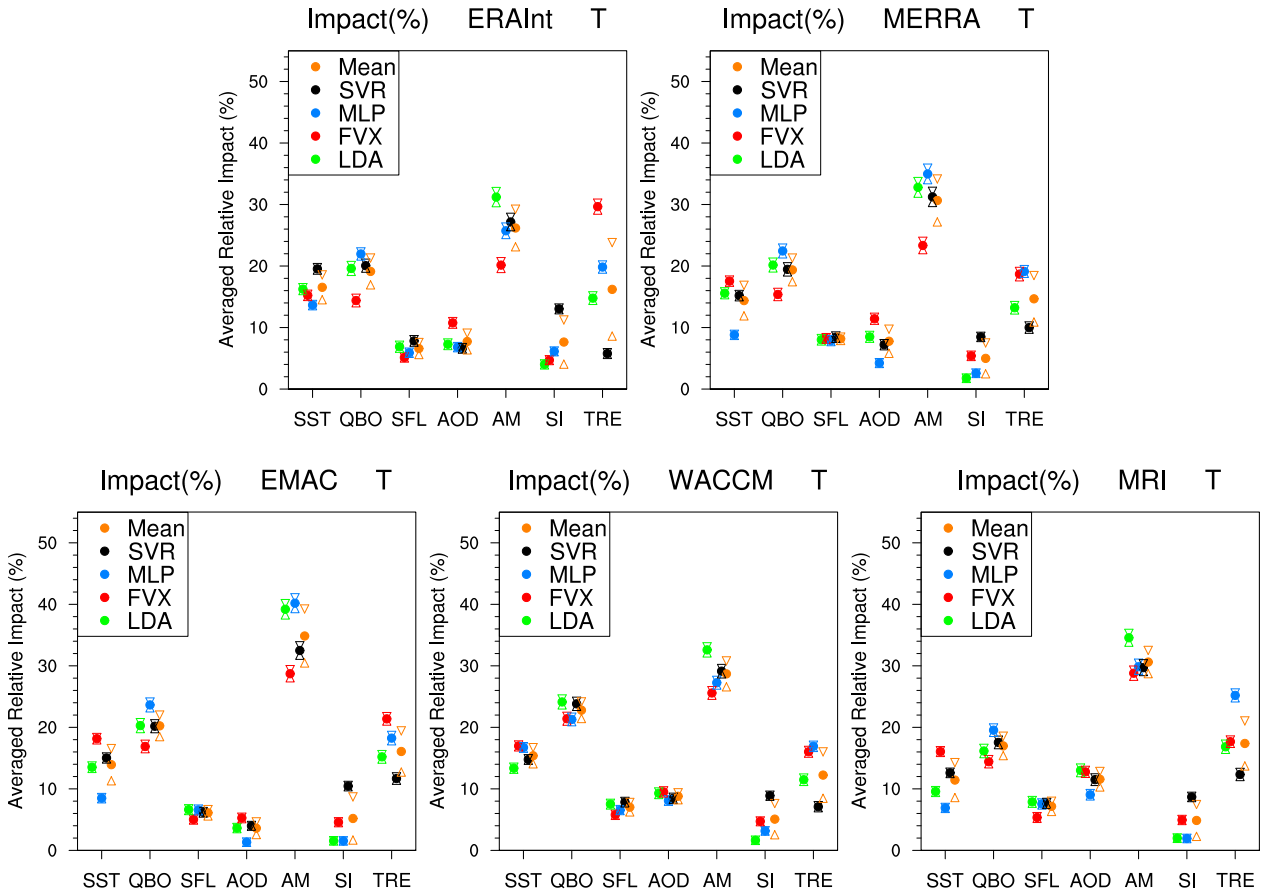

Figure 5.9: The impact of each external factor on **temperature** averaged across the stratosphere, displayed for the reanalyses (top) and the three CCMs (bottom). For each external factor, the impact is shown for each statistical model including its standard error ($\sigma/\sqrt{n}$). The average across the statistical models (Mean) is weighted according to the individual explained variances (see Table 5.4).

higher, varying from 30% to 34%. As for ERAInt, the AM impact from FVX in the other data sets is slightly lower than that from SVR, MLP, and LDA.

The second most important natural factor is the QBO, as observed for all data sets in Fig. 5.9 with only small differences among the statistical models. However, the QBO impact from FVX is once again slightly smaller than that from the other statistical models. The mean QBO impact is 20% for ERAInt, MERRA, and EMAC, 23% for WACCM and 17% for MRI. The third most important natural factor is SST which varies among the different data sets between a minimum of 12% for MRI and a maximum of 16% for ERAInt. For SST, FVX generally attributes the highest importance and MLP the lowest. Only for WACCM, the SST impacts among the statistical models agree well. The two least important factors are SFL and AOD which range between 3% (AOD for EMAC) and 12% (AOD for MRI) across the different data sets. The AOD impact variation between the statistical models varies on a range comparable to that of the QBO impact. In contrast, this variation for SFL is rather small, meaning the statistical models agree well about the importance of this factor in the different data sets. Also the variation of the mean SFL impact across the data sets

is very small and even negligible among the CCM simulations. A minimum of the mean SFL impact of 6% (CCMs) and a maximum of 8% (MERRA) is obtained. The mean SI impact is of the order of the SFL and AOD impact. It is 5% for all data sets but ERAInt (7%). The statistical models agree quite well throughout the different data sets except for SVR which appears to overestimate the SI impact by about 3%.

The impact of the linear trend term (TRE) for ERAInt varies substantially among the statistical models as seen in Fig. 5.9. This is reflected by the large standard error of 6% around the mean TRE impact of 16%. The TRE impact has its minimum at 5% (SVR) and its maximum at 29% (FVX). This points to a large uncertainty of the importance of this factor in ERAInt. However, this large deviation among the statistical models is not observed for the other data sets where the mean TRE impact error is not larger than 3%, being in the range of the mean AM impact error. The mean TRE impact does not vary much among the different data sets. The mean TRE impact has its minimum at 12% for WACCM and its maximum at 17% for MRI. Hence, there is relatively little variation observed for the mean TRE impact among the different data sets. The smallest attribution of the linear trend term is generally computed by SVR compared to the other statistical models. Only the TRE impact from LDA is always close to that of the mean TRE impact across the different data sets, making it a robust model for estimating trends. MLP and FVX seem to both slightly overestimate the TRE impact, depending on the considered data set.

Ozone

The corresponding impact of each external factor on ozone averaged across the stratosphere is displayed in Fig. 5.10. At first glance it is similar to that for temperature (see Fig. 5.10). However, there are significant differences among the different data sets, most notably when comparing NIWA with the remaining data sets. NIWA shows a large mean ESC impact of 38%, FVX even going up to 49% whereas the impact for SVR is around 30% which is more in agreement with mean ESC impacts found in RANDEL (28%), EMAC (24%), WACCM (22%), and MRI (34%). For NIWA, the second most important factor with a mean impact of approx. 15% is SST closely followed by QBO (13%), SFL (9%) and AOD (6%). An extremely small impact of 4% has AM in NIWA, not being in agreement with the other data sets which all point to significantly higher mean AM impacts ranging from 12% for EMAC up to 16% for MRI. NIWA is partly output of a regression model [Hassler et al., 2009] which is why natural variability might be underestimated. However, RANDEL is also partly output of a regression model but variability seems to be more realistically represented than in NIWA.

In contrast to NIWA, the QBO is more important for ozone variability in the other data sets. In fact, it is the most important natural factor with mean impacts of 22% for RANDEL, 23% for EMAC and WACCM, and 14% for MRI. The QBO impact in MRI might be under-represented because the QBO is internally generated in MRI and is subsequently significantly different in amplitude and period compared to the observed QBO [Shibata et al., 2005]. The same fact is observed for the QBO impact on temperature as seen in Fig. 5.9 where the QBO impact from MRI is significantly lower than that

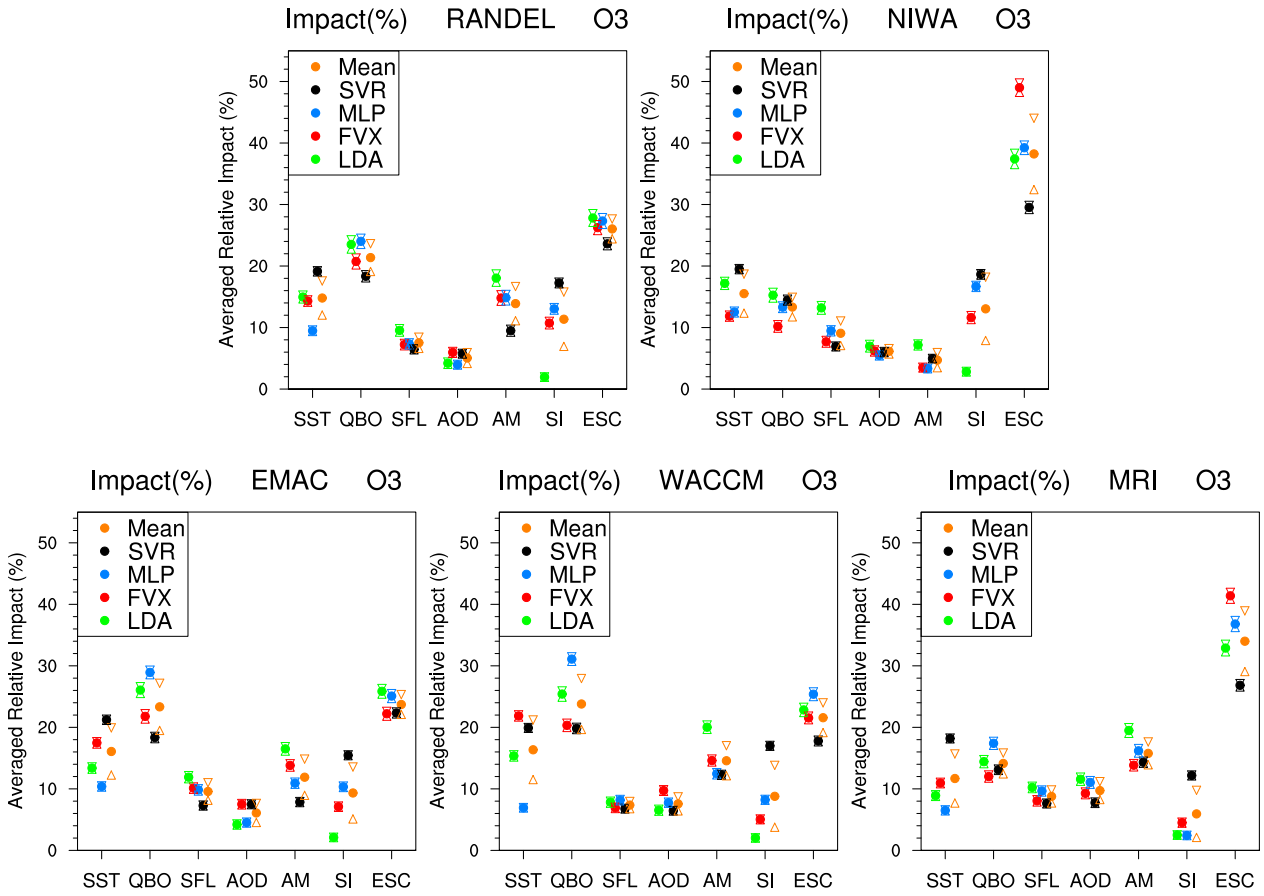

Figure 5.10: The impact of each external factor on **ozone** averaged across the stratosphere, displayed for the reanalyses (top) and the three CCMs (bottom). For each external factor, the impact is shown for each statistical model including its standard error ($\sigma/\sqrt{n}$). The average across the statistical models (Mean) is weighted according to the individual explained variances (see Table 5.4).

of the other data sets. The QBO is nudged in EMAC and WACCM (see Section 4.3) and therefore more realistically represented in terms of amplitude and period. Please note that the mean QBO impacts on ozone for RANDEL, EMAC, and WACCM are almost identical. Also, the QBO impacts on temperature are very similar between ERAint, MERRA, EMAC, and WACCM (compare Fig. 5.9). However, results for temperature are relatively similar among the different data sets. In contrast, the relative impacts on ozone are only similar between RANDEL, EMAC, and WACCM. This will also be reflected by the uncertainty ranges for the impact on ozone in the upcoming subsection.

MLP generally attributes more impact on ozone variability to the QBO than the other statistical models. This is especially pronounced for EMAC and WACCM. Another important factor is the SST impact which appears underestimated by MLP throughout all data sets by up to 9%. SVR on the other hand, appears to overestimate this impact by about 4%. The SST impacts from LDA and FVX are fairly similar through almost all data sets. The mean SST impact is the smallest for MRI (12%) and the largest for EMAC and WACCM (16%). The magnitude of the AM impact is comparable to that of the SST impact throughout all data sets (except NIWA). The mean AM impact on ozone

ranges from 12% for EMAC to 15% for MRI (excluding NIWA). Therefore, it is significantly smaller by at least a factor of two compared to the AM impact on temperature. Only LDA appears to overestimate the AM impact by 4% on average. The AM impacts from the remaining statistical models are close to eachother and to the mean AM impact. Similar to the temperature results, the mean impacts computed for SFL and AOD on ozone are the smallest, ranging for SFL from 8% (WACCM) to 10% (EMAC) and for AOD from 5% (RANDEL) to 10% (MRI). Similar to the temperature results, the statistical models do not vary significantly for the SFL and AOD impacts. They agree well for these two factors. The mean SI impact is of the order of the SFL and AOD impact or slightly larger. It varies quite substantially between the data sets from 6% (MRI) to 13% (NIWA). The mean SI impact shows the largest standard error among the mean impacts in all data sets aside from NIWA and MRI. SVR is most likely overestimating and LDA underestimating the SI impact on ozone. It cannot be answered at this point why the statistical models deviate to that extent for this factor.

There are factors that are weighted similarly and others that are weighted differently among the four statistical models. Strongly robust estimates are obtained for SFL and AOD thoughout all data sets where the impacts agree well. Other factors, including SST, QBO, and AM, lead to impacts where the different statistical models do vary resulting in a mean standard error of 4% on average. The largest differences are observed for the TRE impact in ERAInt and the ESC impact in NIWA and MRI where standard errors go up to 8%. The methods that deviate from the mean impacts are most commonly the advanced statistical models whereas LDA is almost always quite close (1-2%) to the mean impact. Only the AM impact is slightly overestimated by LDA and the SI impact underestimated in temperature and ozone. The application of advanced statistical models, such as MLP, SVR, and FVX, does not seem to be necessary for this application at first glance. However, the application of several independent statistical models makes it possible to quantify a robust estimate including uncertainty for the impact of factors influencing variability in temperature and ozone. For example, without the application of FVX, MLP, and SVR, reasonable estimates of the impact of the annular modes (AM) and the seasonal influence (SI) are not possible.

5.4.2 Averaged Impacts and Natural Variability

To summarize the statistical impacts, the individual mean impacts, as displayed in Fig. 5.9 and Fig. 5.10, are averaged across the different data sets as presented in Fig. 5.11. The different impacts and their errors are combined with the averaging procedure introduced in Section 3.7.3 while assigning equal weight to all data sets.

For temperature, it is observed that the annular modes (AM) have the largest impact of 30% (15% NAM, 15% SAM), followed by the QBO terms with an impact of 20%. SST and TRE are moderately important with 15% whereas SFL and AOD have an impact of 7 and 8%, respectively. The uncertainty ranges are of moderate size with 3-4% for SST, QBO, AOD, AM and only 1% for SFL, indicating robust results. The TRE impact uncertainty is quite large with 5% resulting

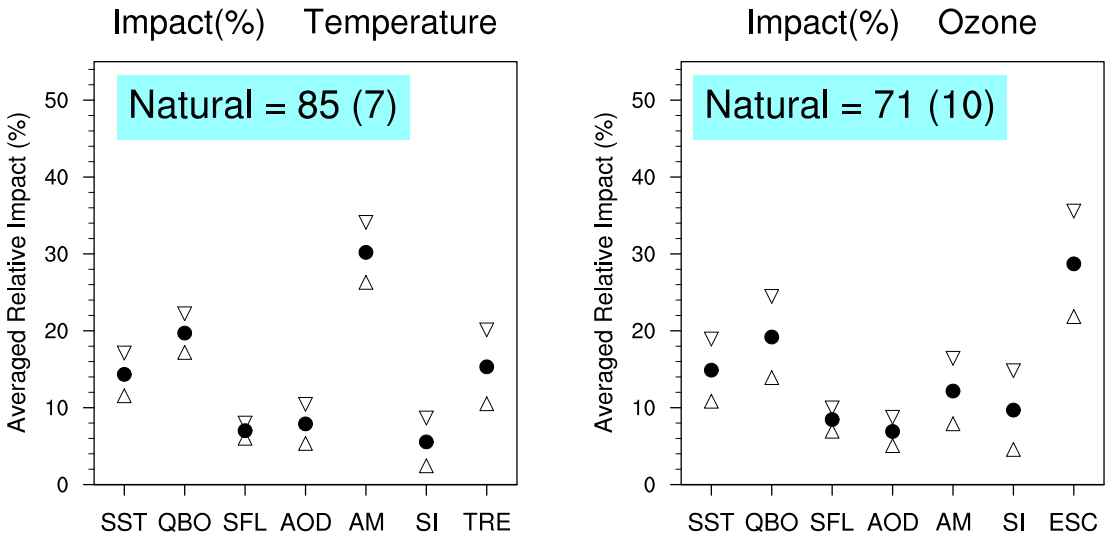

Figure 5.11: The impact of each external factor on stratospheric temperature (left) and ozone (right) averaged across the different data sets. The average was computed according to the model averaging procedure described in Section 3.7.3, assigning equal weights to all data sets. The sum of all natural factors (SST, QBO, SFL, AOD, AM, SI) is given on the light blue background including its standard error in parentheses, calculated from Gaussian error propagation.

from the variation in the data sets and the variation among the statistical models, especially for the TRE impact in ERAInt (compare Fig. 5.9). The SI impact is the smallest with $5\pm3\%$ which is not surprising as the temperature and ozone data were deseasoned prior to the statistical modeling.

The statistical impact of external factors on stratospheric temperature was also investigated in Chapter 8 of SPARC CCMVal [2010] (referred to hereafter as Ch8Val2). This was done for the tropical region (25° S-25° N) and for all CCMs that participated in CCMVal-2. The statistical method of choice for this analysis was multiple linear regression analysis (equivalent to LDA for regression problems). The external factors were ENSO (Nino3.4 index; compare Section 2.2.3), QBO (standard QBO plus one orthogonal term; compare Section 2.2.1), solar cycle (F10.7; compare Section 2.2.4), an artificial term for each of the three major volcanic eruptions (see Section 2.2.5), a trend term (linear for T, ESC for O_3; compare Section 2.3), and an offset term for seasonal influence. Annular modes were not included in the analysis of Ch8Val2 as it was assumend that they do not influence tropical processes significantly, at least not on monthly time scales. However, this is not true entirely as shown in the next section.

Fig. 5.11 presents a ranking of the impact of external factors on stratospheric temperature. A similar ranking was obtained in Ch8Val2, when averaging the absolute magnitudes of the regression coefficients across the different CCMs and the stratospheric levels from 100 to 1 hPa. The following approxmate values are absolute magnitudes: It is found that the QBO has the largest impact of 1 K, followed by the trend term with 0.8 K, and ENSO with 0.4 K which is in relative agreement to the ranking shown in Fig. 5.11. The approximate impacts of the solar cycle (0.4 K) and the volcanic eruptions (1 K) are about a factor of 2 higher than those found here. This is because SFL and AOD

have a much larger impact in the tropics which is the region that was mainly analyzed in Ch8Val2. In this study, the impacts are an average across the global stratosphere, not only the tropics, which decreases the impacts of SFL and AOD.

The sum of the impacts on temperature of all natural factors (SST, QBO, SFL, AOD, AM, SI) is given on the light blue background including its standard error in parentheses, calculated with Gaussian error propagation. It is observed that temperature variability in the stratosphere is determined to 85±7% by natural variability. Therefore, at least 100-(85+7) = 8% of the global monthly stratospheric variability in temperature cannot be explained by external factors representing *only* natural variability. The linear trend (TRE) impact is 15±4%, meaning that the 8% non-natural variability computed above are even smaller than the lower bound (15-4=11%) of the TRE impact. Therefore, variability in stratospheric temperature in the recent climate (1980-2005) was determined to at least 8% by forcings of non-natural, i.e. anthropogenic origin. As mentioned in Section 2.3, the combined effects of GHGs and ODS are responsible for the almost linear stratospheric temperature decrease in the recent and future climate. Randel et al. [2009] found a height-dependent cooling trend for 1979-2007 of about 0.5 K/decade in the middle and lower stratosphere and up to 1.5 K/decade in the upper stratosphere. This could also be reproduced by a range of chemistry-climate models [SPARC CCMVal, 2010].

Concerning the results for ozone, as displayed on the right-hand side of Fig. 5.11, the impacts are similar to those of temperature. However, AM is significantly less important for ozone with an impact of only 12% and SFL slightly more important with 8%. ESC is very important for changes in ozone with an impact of 29%. The QBO with 19% and SST with 15% are the most important natural factors for variability in stratospheric ozone whereas AOD is the least important with 7%. Therefore, impacts of SST, QBO, SFL, and AOD on ozone are not signficantly different from those on temperature. Major differences between the impacts on ozone and temperature are found for AM, SI, and the trend term (TRE or ESC) as shown in Fig. 5.11. The uncertainties are generally larger by 1-2% than those for temperature reflecting the large difference between impacts on ozone for NIWA and the other data sets (compare Fig. 5.10).

The impact of external factors on ozone was primiarly investigated in Ch8Val2 by using the same analysis as introduced above for temperature. A similar ranking was obtained in Ch8Val2, when averaging the absolute magnitudes of the regression coefficients across the different CCMs and the stratospheric levels from 100 to 1 hPa. The following approximate values are absolute magnitudes relative to an ozone climatology: The largest impact on ozone variability was found for the QBO with 5%, followed by ENSO with 3%. SFL and volcanic eruptions with 1% impact each are in relative agreement to the impacts found here. The factor with the smallest impact was found be ESC in the Ch8Val2 analysis with about 0.5%. This does not agree with the large ESC impact in this study because of the different latitude ranges. As mentioned in Section 2.3, ozone loss due to ESC is much larger in the extra-tropical and polar regions.

The sum of the impacts of all natural factors on ozone (SST, QBO, SFL, AOD, AM, SI) is given on the light blue background on the right of Fig. 5.11, including its standard error in parentheses.

Ozone variability in the stratosphere is determined to 71±10% by natural variability. Therefore, at least 100-(71+10) = 19% of the global monthly stratospheric variability in temperature cannot be explained by external factors representing *only* natural variability. The term representing trends in chlorine and bromine (ESC) impact is 29±7%, meaning that the 19% non-natural variability computed above are even smaller than the lower bound (29-7=22%) of the ESC impact. Therefore, variability in stratospheric ozone in the recent climate (1980-2005) was determined to at least 19% by forcings of non-natural, i.e. anthropogenic origin. The ozone loss in stratospheric levels is caused by the depeletion due to long-lived chlorine and bromine represented by ESC (see Section 2.3). The trend in ozone depends greatly on latitude and season but is greatest during late winter and spring over Antarctica with about 40-50% loss with respect to pre-1980 values. The ozone loss during spring over the Arctic is about a factor of two smaller and more variable. The loss in the tropics is much smaller with only a few percent [WMO, 2007]. In the following, it is investigated where the different external factors appear to be most important regionally in the stratosphere.

5.4.3 Regional Importance

This section examines where the external factors have their individual regions of highest statistical impact in the stratosphere averaged across the four statistical models and the different data sets. To do so, only the natural factors, other than AM, having the largest impact are displayed depending on the specific region in the stratosphere. This is shown on the upper panel of Fig. 5.12 for temperature and ozone. Blue denotes SST, green QBO, yellow SFL, and red AOD. A factor's impact is statistically significant on the 95% level if the corresponding region is *not* covered with a hashed pattern. For instance, the strongest natural factor aside from AM at the equator in 30 hPa is the QBO (note the green significant area). The impact of AM is superimposed using isolines with a distance of 20%. For temperature, it is observed how there are several centers of AM activity. Most pronounced over the poles past 60° N/S (compare Section 2.2.2) and strong in the lower tropical stratosphere for temperature. The NAM is responsible for the strongest AM activity (up to approx. 70% in the middle polar stratosphere) in the northern hemisphere and the SAM for AM activity in the southern hemisphere (up to approx. 60% in the middle polar stratosphere). For ozone, the overall AM impact is at least a factor of two smaller (compare Fig. 5.11) but the centers of activity remain in similar regions of the stratosphere. The highest activity, caused by the NAM on ozone, is found to be in the north-polar middle stratosphere (up to approx. 30%).

NAM and SAM do not only have an impact on the extratropical, but also on the tropical stratosphere, particularly notable for temperature in Fig. 5.12. This impact is especially large in the lower stratosphere (approx. 30%) between approx. 30° S-10° S and 10° N-30° N. This reflects the fact that anomolous temperatures in the polar region tend to lead to temperature anomolies in the lower tropical stratosphere which is an indirect effect of wave-mean flow interaction in the polar stratosphere [Kodera, 2006]. The impacts of NAM and SAM on ozone in the tropical lower stratosphere are more in midlatitudes and a factor of three smaller than those for temperature.

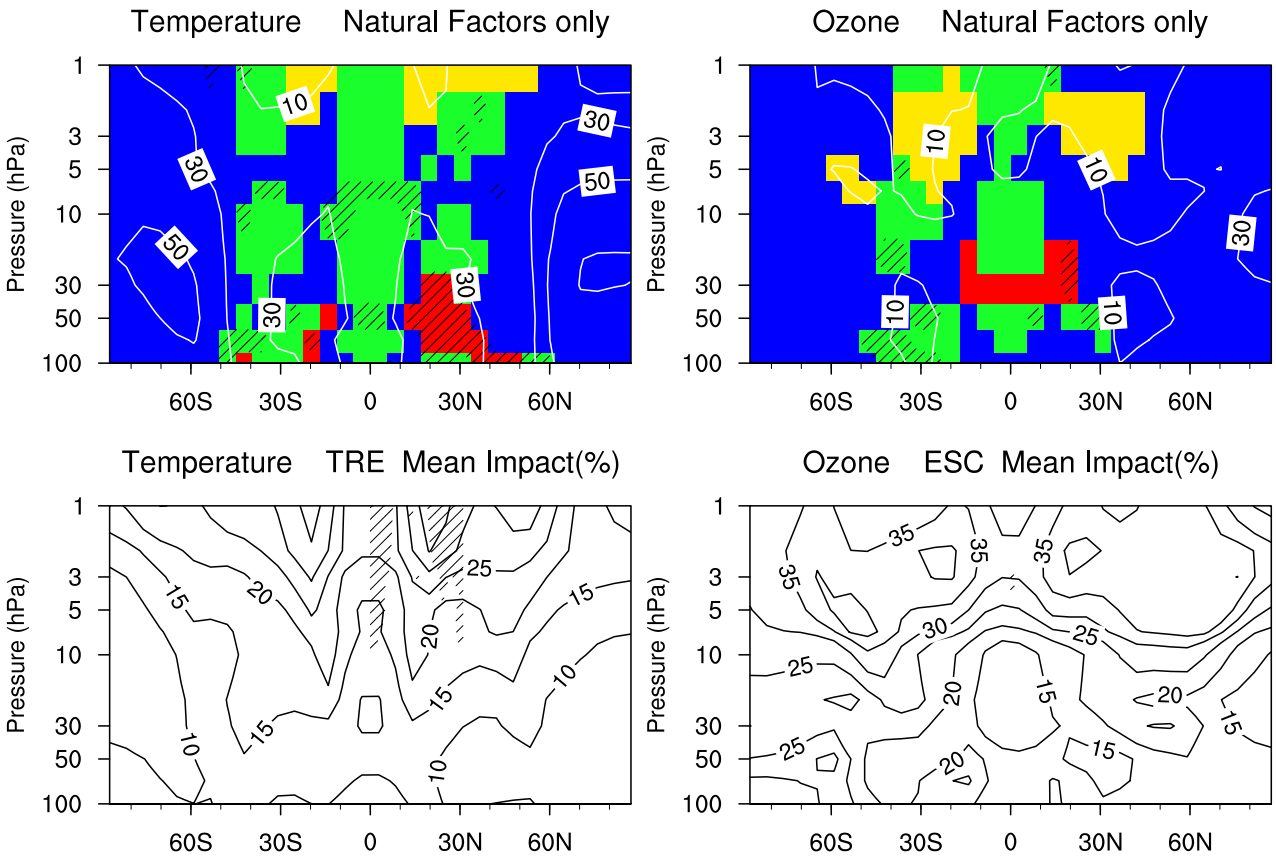

Figure 5.12: Individual regions of highest statistical impact of the external factors averaged across the four statistical models and the different data sets for temperature (left) and ozone (right). *Top:* The natural factors, other than AM, having the largest impact is displayed depending on the specific region in the stratosphere where blue denotes SST, green QBO, yellow SFL, and red AOD. The impact of AM is superimposed using isolines with a distance of 20%. *Bottom:* The statistical impact of TRE and ESC using isolines with a distance of 5%.

However, NAM and SAM impacts on ozone are much more pronounced in the middle and upper tropical stratosphere. They are only moderate (10-20%) but extend over a large range reaching from the equator to midlatitudes up to 60° S/N. It is not clear if this denotes an actual relationship or rather a statistical artifact.

The following focuses on the natural factors, other than AM, as displayed in the upper panel of Fig. 5.12. For both temperature and ozone, it is observed that the SST anomalies have the largest impact on the extratropical region past 50° N/S. This SST impact is only moderate with 10-15%, as observed in Fig. 5.11, but still larger than any of the other natural factors in the extratropics. Significant impacts of variabilities in SSTs, especially by ENSO on extratropical dynamics, have also been reported by, e.g., Camp and Tung [2007b], Calvo et al. [2009], and Mitchell et al. [2011]. However, there are also significant links between ENSO and temperature along with ozone in the tropical lower stratosphere [Randel et al., 2009; SPARC CCMVal, 2010], as observed in Fig. 5.12.

The tropical region is largely dominated by QBO variability for both temperature and ozone. The QBO impact on temperature is rather symmetric around the equator whereas the impact on

ozone is less pronounced on the northern hemisphere. The QBO is a phenomena of the equatorial stratosphere (compare Section 2.2.1). It was reported numerous times that the QBO affects tropical processes in the stratosphere to a large extent [e.g., Holton and Tan, 1980; Baldwin et al., 2001; SPARC CCMVal, 2010]. This is not only true for the tropics but also for the extratropics and the polar region in particular. Many studies have shown [e.g., Holton and Tan, 1982; Labitzke, 1987; Camp and Tung, 2007a; Blume et al., 2012] that the QBO governes dynamics in the polar stratosphere to a great deal, also shown in the following chapters.

The SFL impact on temperature is high and significant at the stratopause between 30° S-20° S and 20° N-30° N. The SFL impact centers are similar for ozone but are slightly larger and lower at around 3 hPa. In fact, the SFL impact for temperature and ozone is around 10% in the upper stratosphere from approx. 50° S to 50° N and around 5% in the lower equatorial stratosphere between 10° S and 10° N (not shown) in agreement with, e.g., Frame and Gray [2010]. The upper solar signal is a direct result of an increased UV absorbtion by ozone during solar maximum and, hence, increased heating in this region. The lower solar signal is a matter of ongoing research and is most likely a result of the dynamical interplay with the QBO [Matthes et al., 2010]. The CCM simulations during CCMVal-2 do not all show a lower solar signal, whereas the upper signal is quite consistent among the CCMs [SPARC CCMVal, 2010, Chapter 8]. The SFL impact on ozone is also significant but smaller for the polar stratosphere (more than 5%). Similar to SSTs and the QBO, the solar cycle has a significant influence on extratropical processes in the stratosphere (compare Section 2.2.4), in particular during northern hemisphere winter [e.g., Kodera and Kuroda, 2002; Labitzke et al., 2006; Camp and Tung, 2007a]. This is addressed in more detail in the following chapters.

The AOD impact on temperature is large in the lower equatorial stratosphere but not significant. A significant AOD signal is observed for ozone between approx. 20° S and 20° N at around 30 hPa. Both AOD signals range between 10% and 15% and are in agreement with previous work that showed that volcanic eruptions have their greatest influence in lower tropical levels of the stratosphere [e.g., Robock, 2000; SPARC CCMVal, 2010]. The influence of volcanic eruptions on the extratropical and polar stratosphere is still poorly understood [e.g. Frame and Gray, 2010; Mitchell et al., 2011]. This is because volcanic eruptions have to be powerful enough to inject sufficient aerosols into the stratosphere. During the observed stratospheric record, this has only happend three times (compare Section 2.2.5).

The impact of TRE on temperature is displayed on the lower panel of Fig. 5.12. The TRE impact is high in the upper stratosphere (up to 35%) and lowest over the poles in the middle stratosphere (less than 10%). A very similar pattern was found by [Randel et al., 2009] as mentioned earlier. Similar to Fig. 5.11, ESC has a larger and regionally wider impact on ozone than TRE on temperature. The ozone variability in the entire upper stratosphere from about 7 hPa to 1 hPa is governed by ESC to at least 30%. High ESC impacts are also observed over Antarctica (approx. 25%). ESC impacts are smallest in the equatorial middle and the north-polar middle stratosphere. This ESC impact pattern on ozone agrees well with previous work based on observations [Staehelin et al., 2001].

5.5 Long-Term Forecasting

In this section, statistical forecasts from 2006 to 2100 are made using each of the statistical models. These forecasts are then compared to REF-B2 model simulations made with EMAC, WACCM, and MRI (compare Chapter 4). This will answer two major questions: 1) Is it possible to adequately predict long-term stratospheric trends with statistical models? 2) Which of the statistical models is best suited for future projections? In order to make future projections, there has to be assumptions about the external factors. The SST variabilities along with the QBO factors from 1980 to 2005 have been simply repeated for the forecast period where it was ensured that connecting points are smooth. To project the 11-year solar cycle into the future, the representative solar cycle number 23 from 04/1996 to 07/2008 is repeated until 2100. Volcanic eruptions cannot be predicted which is why AOD is kept at its median computed from the training period (1980-2005). There can be no assumptions being made about the annular modes which is why they are held at their medians as well.

The resulting statistical forecasts on a yearly resolution, along with the corresponding REF-B2 model simulation (orange), are presented for the temperature on the left panel of Fig. 5.13. Displayed is the first principal component (PC) of the entire spatial domain calculated from 1980 to 2100 along with the variance it explains (EV). The slope m is calculated from a linear fit from 2005 to 2100 for the CCM and each statistical model. The orthonormal EOF, shown on the right panel of Fig. 5.13, corresponds to the first PC. It is now easy to estimate the trend in a certain region of the stratosphere. For instance, the first temperature PC of EMAC reveals a trend of -1.19 K/y. This value is now multiplied with 0.06, a value of the equatorial upper stratosphere of the corresponding first EOF. The resulting value of -0.07 K/y indicates the temperature trend for this region.

When examining the temperature results for all three CCMs, it is noted that EV is always close to 90% for the first principal component. This means that the first PC is highly representative for the long-term, interannual stratospheric behavior. Forecasting the first PC implies forecasting almost 90% of the overall stratospheric variance. It is observed for all three CCMs that the first PC can be forecasted well by the statistical models. The slopes between CCM forecast and statistical forecast are similar. This holds for temperature as shown in Fig. 5.13 but also for the corresponding ozone results as displayed in Fig. 5.14. It can be concluded that it is possible to forecast the long-term stratospheric behavior using statistical models. For temperature, the long-term behavior is governed by the linear trend (TRE). For ozone, the long-term behavior is governed by effective stratospheric chlorine (ESC). The interannual behavior is mainly governed by SFL but also by forcing in SST. The solar cycle signal is stronger for ozone (compare Fig. 5.11) and can be observed in the PCs. It is noted that the statistical ozone forecasts for EMAC get out of phase for the solar cycle. This is not an error in the statistical forecasts but simply caused by different solar cycle forcings. For the EMAC projection, the last four solar cycles form the recent climate were repeated whereas only one representative solar cycle was repeated for the statistical forecasts.

The forecast performance, in terms of slope, is summarized in Table 5.5 for temperature and

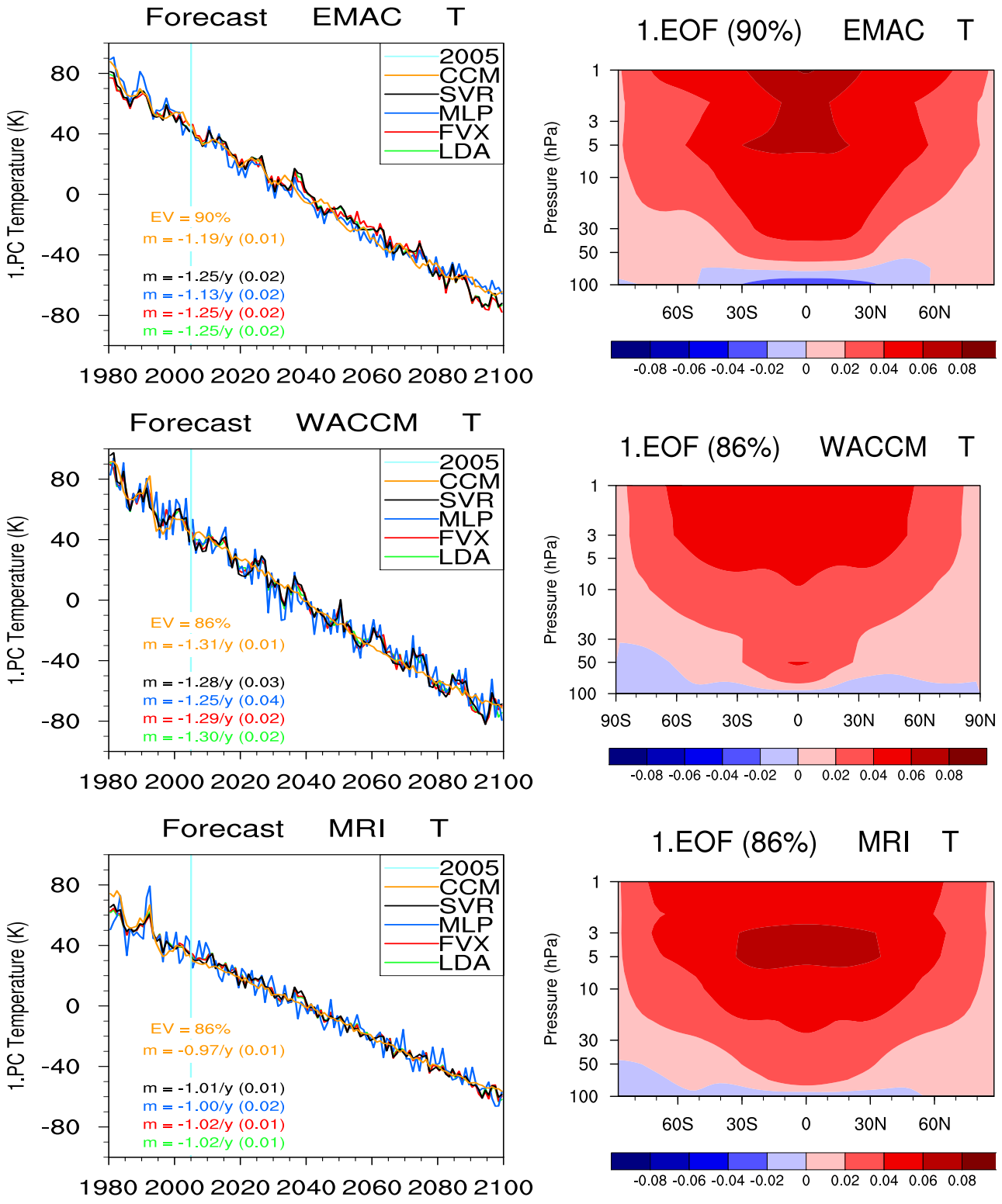

Figure 5.13: *Left:* Statistical temperature forecasts in a yearly resolution along with the REF-B2 model simulations for the first PC for EMAC (top), WACCM (middle), and MRI (bottom). The variance explained (EV) by the EOF and the different slopes for each time series calculated from 2005 to 2100 are also displayed. *Right:* The orthonormal EOF corresponding to the first PC along with the variance it explains in parentheses.

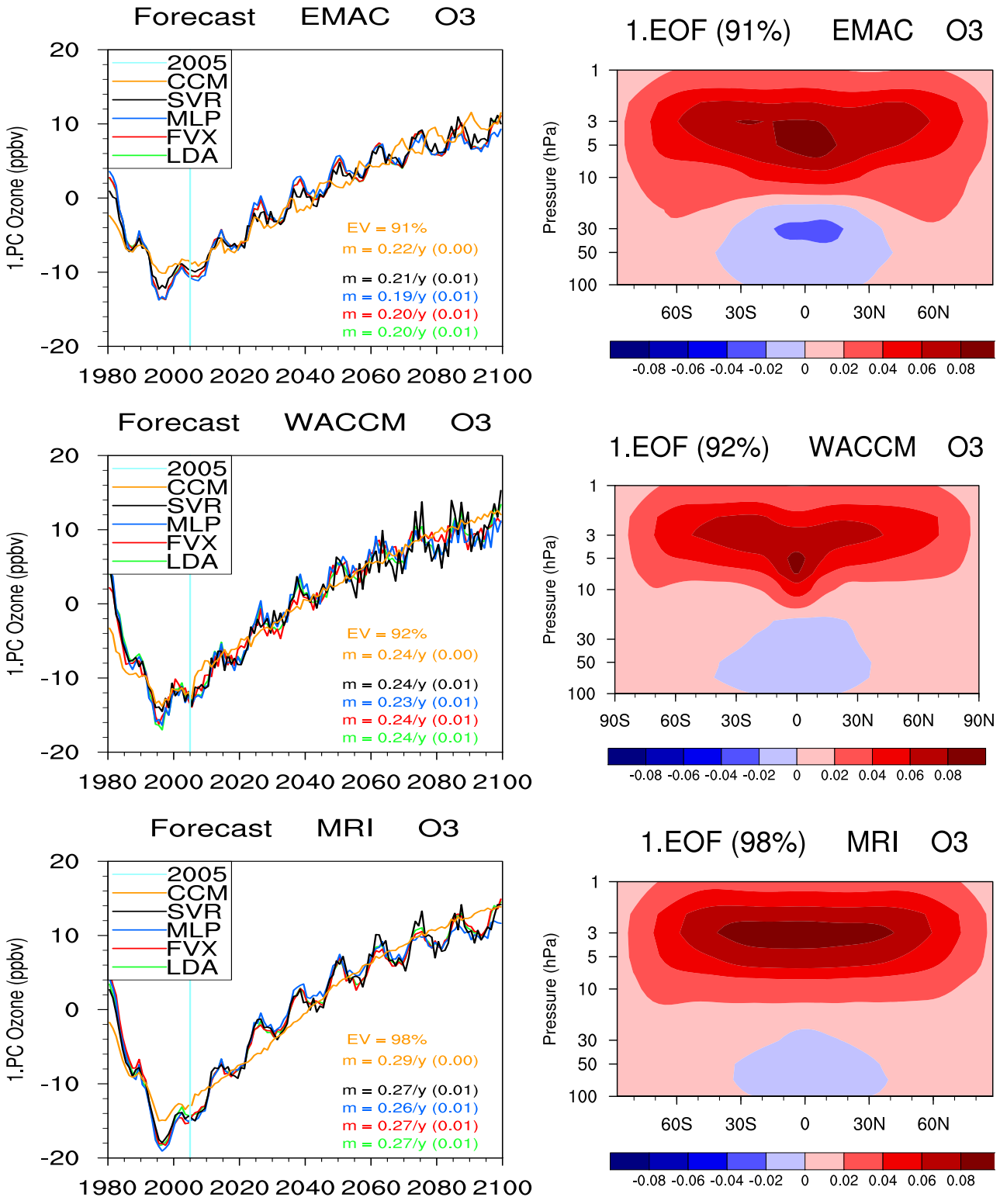

Figure 5.14: *Left:* Statistical ozone forecasts in a yearly resolution along with the REF-B2 model simulations for the first PC for EMAC (top), WACCM (middle), and MRI (bottom). The variance explained (EV) by the EOF and the different slopes for each time series calculated from 2005 to 2100 are also displayed. *Right:* The orthonormal EOF corresponding to the first PC along with the variance it explains in parentheses.

Table 5.5: Comparison of the different statistical slopes with respect to the simulated CCM slope for temperature and ozone. Shown is $1 - \frac{|m_{Mod} - m_{CCM}|}{|m_{CCM}|}$ averaged across the three CCMs so that values close to one denote a good forecasting performance.

	LDA	FVX	MLP	SVR
T	0.96	0.96	0.96	0.96
O_3	0.95	0.95	0.90	0.96

ozone with each statistical model averaged across the results for the CCMs. Shown is one minus the relative difference to the reference run, meaning that values close to one denote a good forecasting performance. It is observed that there is no difference between the different statistical models for temperature and only small differences for ozone. Only MLP reveals a significantly lower forecast performance for ozone than the other methods which may be caused by a possible overfitting during the MLP training. Overall, the long-range statistical forecasts for the stratosphere covering several decades are efficient. Furthermore, already a simple linear and stationary method, such as LDA, can forecast variability on these time scales. However, it was shown in Table 5.5 that most reliable forecasts can be made using SVR. However, the differences between the statistical model performances are not significant.

5.6 Concluding Remarks

In this study, the large-scale stratospheric variability in monthly zonal mean deseasoned temperature and ozone is modeled using purely statistical methods with the help of external factors known to influence stratospheric processes. The statistical methods, introduced in Chapter 3, are LDA, FEM-VARX, MLP, and SVR. These methods are partly linear, nonlinear, stationary and nonstationary. The wide range of different statistical approaches, makes it possible to assign robust uncertainties to the impacts of external factors. The learning approaches are trained with historical data representing the recent climate (1980-2005) from observations, reanalyses, and CCM simulations. Since the dimensionality of the data to be modeled is large, principal component analysis (see Section 3.1) was used to efficiently reduce dimensionality (see Table 5.1).

It was shown how optimal model architectures for FEM-VARX, MLP, and SVR can be determined with the help of information criteria and cross-validation (see Table 5.2). After being set up with the optimal setting, the statistical models are trained and evaluated to measure their regression performance. It was observed that there is no significant difference between the regression performance (see Table 5.2) of LDA, FEM-VARX, and MLP. SVR clearly outperformed the other statistical models by a factor of 1.5 for temperature and a factor of 1.3 for ozone in terms of averaged explained variance. For the following, it was assumed that a learning approach that can model more variability will also be able to more realistically weight external factors according to their statistical importance (see Table 5.4).

The statistical modeling in this study was performed with the help of a range of external

factors. The natural factors considered were: SST, QBO, SFL, AOD, NAM, SAM, and a seasonal influence (SI). In addition, it was aimed at including terms representing anthropogenic influence. For temperature, this was done by including a linear trend term representing the linear stratospheric temperature trend observed for the recent climate and simulated for the future (see Section 2.3). For ozone, ESC, representing chlorine and bromine as the main ozone depleting substances, was included (see Section 2.3). The individual impacts of all these factors on temperature (see Fig. 5.9) and ozone (see Fig. 5.10) were computed for the different statistical models and data sets. For temperature, it was found that the mean impacts agree quite well across the data sets. The statistcal models agree in the percent range for most factors but disagreed significantly for factors such as TRE in ERAInt or AM in EMAC by up to about 20%. For ozone, the variation among the statistical models is similar. The differences across the data sets are much larger. Especially the impacts found in NIWA, which is itself output of a regression model, are quite different from the other data sets. In NIWA, the ESC impact is overestimated whereas the QBO impact is underestimated. The QBO impact appears also underrepresented in MRI which may be caused by the internally generated QBO. The QBO in MRI is significantly different from the observed QBO in terms of amplitude and period.

Concerning the different statistical models, it was found that it were mostly the more advanced models (FVX, MLP, SVR) that deviated from the mean impact whereas the linear and stationary LDA was usually close to this mean value. However, including the different methods makes it possible to compute robust impacts including uncertainties. For AM and SI in particular, the computation of reasonable impacts would have been not feasible without including all four statistical methods. One challenge in the current study is that the number of training events, i.e. the number of steps in time, is quite small. The training period from 1980 to 2005 consists of a total of only 312 months which is why FEM-VARX and MLP do not perform significantly better than the linear regression technique. It will be shown in the following chapters that this drastically changes when more training events become available.

In order to obtain further insight, the individual impacts were averaged across all data sets and all statistical models. The results were robust estimates including uncertainties of the overall impact of external factors on stratospheric variability found in temperature and ozone (see Fig. 5.11). For temperature, it was found that the annular modes have the largest combined impact of 30%. However, half of that is attributed to NAM and SAM, respectively. Hence, the QBO with an impact of 20% is the most important factor for stratospheric temperature. Another important forcing is SST which showed an overall impact of 15%. The impact of the linear trend term is similar. In contrast, the impacts of SFL (7%), AOD (8%), and SI (5%) are rather small. A comparable factor ranking was observed for ozone. In fact, the impacts of SST, QBO, SFL, and AOD on ozone variability are not significantly different from those on temperature. The AM impact is found to be smaller (12%) and the ESC impact to be larger (29%) than the corresponding impacts for temperature. With these impacts and their uncertainties at hand, it was possible to quantify the overall natural variability inherent in the stratosphere. For temperature, the natural factors combined explain $85\pm7\%$ of the total temperature variability, implying that there is at least 8% variability that

is anthropogenically caused by the emissions of greenhouse gases and ozone depleting substances (see Section 2.3). For ozone, the natural factors are responsible for 71±10% variability indicating that at least 19% is of anthropogenic origin due to the emissions of ozone depleting substances (see Section 2.3). The importance ranking of the external factors found agree with previous work [e.g., SPARC CCMVal, 2010, Chapter 8]. Aside from the globally averaged impact, the regional importance of each external factor was computed, averaged across the data sets and the statistical models (see Fig. 5.12). Despite of the differences between data sets and models, large significant areas were observed. It was found that the AM and SST mainly govern variability in the extratropics, whereas the tropics are governed by the QBO, SFL (more upper), and AOD (more lower). The trend terms, TRE and ESC, are highest (up to 30-35%) in the upper stratosphere but take values of at least 10-15% in lower stratospheric levels (see Section 2.3).

Finally, long-term statistical forecasts until 2100 with each of the statistical methods were performed. This was done for temperature (see Fig. 5.13) and ozone (see Fig. 5.14) while making reasonable assumptions about the external factors. The natural factors have been either repeated (SST, QBO, SFL) or held at their medians (AM, AOD, SI). The trend term was linearly extended (TRE) for temperature. ESC was extended according to the A1 scenario for ozone (see Section 4.3). It was found that the statistical models are able to reliably forecast the long-term behaviour in the stratosphere. Aside from short-term interannual behaviour, the statistical forecasts agreed well with those made by the CCM simulations (REF-B2). It can be concluded that long-term stratospheric statistical forecasts are good and effective. Suprisingly, there is no significant performance difference between the statistical models (see Table 5.5). Only MLP performs slightly worse for the ozone forecast than the other models. From the results produced in this particular study, the usage of advanced statistical models, such as LDA, FVX, MLP, is justified when it is nessecary to compute reliable impacts of external factors including uncertainties. When only long-term forecasts are made, a linear and stationary model, such as LDA, is sufficient. This is because the trend term in the temperature and ozone forecasts dominates the overall behaviour.

One has to be careful when interpreting the results presented in this study. Statistical models incorporating external factors always make a range of assumptions. As observed, assumptions such as stationarity and linearity did not play a big role for this particular application. Other assumptions, such as the independence, might play a bigger role. It is known from the literature that all the forcings and modes of variability used here influence each other to some extent. As mentioned earlier, this cannot be avoided in a statistical analysis, but has to be thought of when analyzing results. Please note that almost zero correlation between two forcings, e.g., QBO and ENSO, does not indicate their physical independence. In addition, the relatively small number of training events may influence the regression performance and the statistical importance of external factors.

There are several improvements that could be made in future analysis of this kind. A range of temporal lags could be investigated to find out if lagging a certain factor improves the regression performance. This would dramatically increase computation time since the optimal model architectures have to be determined for every combinations of lags. The dimension reduction was done to

greatly decrease computation time. For one data set and one multivariate statistical method, such as FEM-VARX, it should be feasible to repeat the modeling on the full domain. However, it is not expected that results for impact or forcast will change signficantly but the regression performance may increase. It will also be interesting to include other factors, such as the semiannual oscillation, into the analysis to see if impacts and performance change.

Chapter 6

Classification of Sudden Stratospheric Warmings

The variability of the north-polar stratospheric vortex is a key dynamical feature of the middle atmosphere specifically, its breakdown during winter resulting in a sudden stratospheric warming (see Section 2.1). There have been several methods proposed in the past that can classify stratospheric warmings and measure the variability of the stratospheric vortex. Very common is the method based on the zonal mean zonal wind at 60° N and 10 hPa originally introduced by the Stratospheric Research Group Berlin [Labitzke and Naujokat, 2000] and incorporated by the World Meteorological Organization (WMO). It was used by Charlton and Polvani [2007] to compile climatologies of sudden stratospheric warmings derived from reanalyses data. It is a simple and effective method for measuring if and when a sudden stratospheric warming takes place leading to a vortex breakdown. Another method is based on the Northern Annular Mode (NAM) [Baldwin and Dunkerton, 2001] computed from geopotential anomalies in the middle stratosphere. The NAM measures the deviation from the climatological mean state of the polar middle atmosphere. It measures the amount of disturbance but cannot alone be used to detect the occurrence of a vortex breakdown. It is widely used to detect downward propagation into the troposphere. The method based on 2D-moments [e.g., Mitchell et al., 2011] is a different way of measuring vortex variability. In contrast to the zonal wind measure and the NAM, it directly examines the geometrical structure of the vortex, such as position and size. In addition, it is used to measure the vortex strength.

In this chapter[1], a method is proposed that extends and combines the zonal wind measure and the NAM approach but does not examine the vortex structure. It incorporates significant atmospheric forcings, called external factors, that play an important role in the winter time evolution of the polar stratosphere. These externals factors are the Quasi-Biennial Oscillation (QBO) [e.g., Holton and Tan, 1980, 1982], the El Niño-Southern Oscillation (ENSO) [e.g., Manzini et al., 2006], and the 11-year solar cycle (SFL) [e.g., Gray et al., 2010]. These forcings interact and create a complex and nonlinear dynamical response [e.g., Calvo et al., 2009; Richter et al., 2011]. There are previous

[1] This chapter, except for Section 6.9, was published with a few minor changes in Blume et al. [2012].

efforts, such as e.g., Labitzke and Kunze [2009a], Camp and Tung [2007a,b], or Mitchell et al. [2011], that statistically investigated the impact of these forcings on the evolution of the polar vortex. Their analysis methods are linear, incorporating only a few factors at the same time. Here, a nonlinear method is used with three external factors simultaneously to classify not only sudden stratospheric warmings but also minor and major final warmings as well as undisturbed vortex states at the same time. The classification procedure is a continuous analysis of stratospheric warming events for 52 consecutive winters on the period from 1958 through 2010. In addition, the derived framework is used to classify stratospheric warmings in data from chemistry-climate model simulations.

In contrast to previous methods, the proposed classification method does not lead to a yes/no criterion but a continuous probability measure, which has the advantage of detecting the amount of deviation from the climatological mean state of the Arctic stratosphere. This disturbance of the polar vortex can then end up in one of the aforementioned stratospheric warming events. Dealing in terms of probabilities has the advantage of obtaining a temporal evolution of the likelihood of occurrence of a stratospheric warming state (e.g. major warming state), given the remaining states. In addition, the temporal onset, duration, and intensity of stratospheric warming events is calculated independently of a particular pressure level.

In this chapter, a wide class of supervised learning methods is considered and a classification method for stratospheric warmings based on a nonlinear statistical model, a neural network, is proposed. A *supervised* statistical method needs fixed pairs of input and output objects presented to it during training, meaning that the true outcome is known a priori (compare Chapter 3). It is shown that a nonlinear model is suited better to recognize the complex, nonlinear response between atmospheric forcings and polar vortex variability. Moreover, it is demonstrated that the approach based on a neural network can classify not only major mid-winter stratospheric warmings (referred to hereafter as *major warmings*), but also minor stratospheric warmings (referred to hereafter as *minor warmings*) as well as major final stratospheric warmings (referred to hereafter as *final warmings*). So-called Canadian warmings will be grouped into the class of minor warming events.

Major and final warmings are characterized by a strong anomalous temperature increase at most pressure levels of the Arctic middle stratosphere, accompanied by a breakdown of the polar vortex and a reversal of the zonal stratospheric flow in mid-latitudes from westerlies to easterlies. Major warmings are often preceded by blocking situations in the troposphere over the Atlantic and/or Pacific sector [Martius et al., 2009]. Major warmings happen on average every other year during mid-winter; hence there is enough time for the polar vortex to recover after a major warming. The polar vortex does not recover after a final warming as they take place at the transition between winter and summer circulation. Please note that final warmings naturally happen every year whereas final warmings [Labitzke and Naujokat, 2000] in this work have to be accompanied by an anomalous temperature increase with respect to a climatology (*major final warming*). Minor warmings are characterized by an anomalous temperature increase, and do not lead to a reversal of the zonal stratospheric flow in mid-latitudes, but to a disturbed polar vortex. Minor warmings often take place more than once in a given winter and are typically more upper stratospheric events. Canadian

warmings are minor warmings with the difference that anomalous temperatures are observed mainly in lower levels of the polar stratosphere. They are characterized by an enhancement of the Aleutian high [Labitzke and van Loon, 1999].

6.1 Data and External Factors

Three reanalysis data sets are utilized in this study (compare Chapter 4): ERA40, ERAInt, and NCEP. Time series from ERA40 and ERAInt have been combined to one data set that is referred to hereafter as ERA. This combination is justified by a small approximately Gaussian residual with zero mean calculated from the overlapping period (1989 to 2002) between the time series used in this work, separately calculated from ERA40 and ERAInt. In this combined set, ERA40 data have been used until 3/1/1989 and ERAInt data thereafter. This date has been selected because stratospheric temperatures and winds are very similar at and around this date, leading to a smooth transition between the two data sets. Both data sets, ERA and NCEP, are utilized for the time from 10/1/1958 through 5/1/2010, which covers 52 winters. ERA is utilized to train the statistical model. NCEP is utilized for validation, because it is quite different from ERA in the polar region due to sparseness of observations, especially on the daily scale and for the pre-satellite era [Labitzke and Kunze, 2005; Charlton and Polvani, 2007]. Also, it only reaches up to 10 hPa leading to potentially different variability compared to ERA. ERA and NCEP have many input factors in common but especially during the pre-satellite era, forcings in sea surface temperature (as seen in e.g. ENSO) along with equatorial stratospheric winds (as seen in e.g. QBO) are significantly different. Except for the zonal wind, all time series are normalized to ensure similar magnitudes according to Eq 3.52.

This analysis makes use of three *external factors* which describe large-scale phenomena important for the stratosphere. It has been shown in previous studies [e.g., Labitzke and van Loon, 1988; Camp and Tung, 2007a,b; Mitchell et al., 2011] that there exists a complex link between the external factors, namely the Quasi-Biennial Oscillation (QBO), the El-Niño Southern Oscillation (ENSO), the solar cycle (SFL), and the vortex variability. These studies showed e.g., that the least perturbed vortex state is solar minimum and QBO west. It was also shown that the polar vortex is more likely to break down during El Niño-like conditions. Other studies have shown that this link is nonlinear [Calvo et al., 2009; Richter et al., 2011]. The idea behind this study is to incorporate these external factors to classify stratospheric warmings on the one hand, and on the other to obtain insight into their statistical importance and interrelationships. A brief description of the corresponding indices is given in the following.

The QBO index is the 50 hPa zonal mean zonal wind anomaly averaged between 5° S and 5° N (compare Section 2.2.1). For a representation of ENSO (compare Section 2.2.3), the Nino3.4 index [Trenberth, 1997] is used, which is the area-weighted average in sea surface temperature anomalies in the box from 170° E to 120° E and from 5° S to 5° N. As a proxy for the solar irradiance, the daily radio flux at a wavelength of 10.7 cm (F10.7; `ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/`

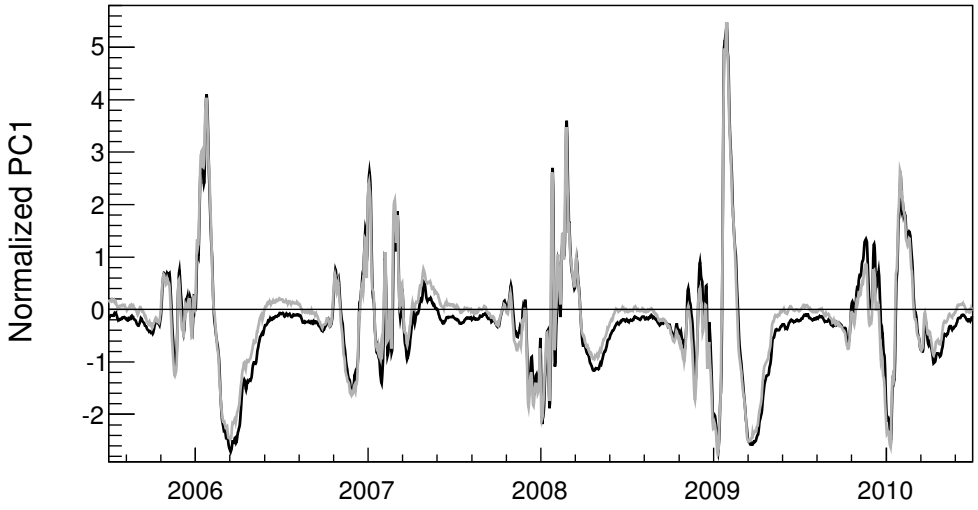

Figure 6.1: Normalized first principal component for ERA (black) and NCEP (gray) for a sample period from summer 2005 to summer 2010 covering 5 winters. Labeled is the first of January of the particular year.

`SOLAR_RADIO/FLUX/Penticton_Observed`) is utilized (compare Section 2.2.4). There are a few missing values in the daily F10.7 which were filled by a linear interpolation, computed from the neighboring measurements. In order to reduce short-term fluctuations, the daily external factors have been sent through a low-pass filter calculating the 10-day running mean. This value was chosen to be more than a few days but a lot less than a month. Hence, daily short-term extremes are avoided in ENSO, QBO, and SFL; but an approximately weekly resolution is retained.

6.2 Temperature Representation

The classification procedure uses stratospheric temperatures because *warming* events have to be detectable naturally in Arctic temperatures. Temperature time series are considered at 10, 20, and 30 hPa levels where stratospheric warmings are always observed. They are also observable in upper and lower parts of the polar stratosphere depending whether the event is a vortex split or displacement event. Vortex splitting events tend to be observable near-instantaneously throughout most parts of the polar stratosphere (approx. 20-40km) whereas vortex displacement off the pole increases with altitude [Matthewman et al., 2009].

The temperature time series are taken as a area-weighted average over the north-polar cap between 60° N and 90° N and are anomalies relative to their individual long-term mean. This treatment makes the time series equivalent to the northern annual mode in temperature (not geopotential) at the respective levels. The resulting temperature time series are highly correlated. But, the inputs to a classification approach should be decorrelated. A principal component analysis (see Section 3.1) of the three time series reveals that the first principal component (PC1) explains more than 90%

of the overall variance in both ERA and NCEP (not shown). Therefore, PC1 is solely used as a robust representation of the temperature anomalies in the Arctic middle stratosphere, not favoring a particular pressure level. The normalized PC1 is displayed in Fig. 6.1 for ERA and NCEP, for a sample period of 5 winters from autumn 2005 to spring 2009. Because of the high degree of explained variance, PC1 is not used for classification only but also as a measure for the intensity of a stratospheric warming event (see Section 6.8). Intensity is therefore not only a measure of strength but also of vertical expansion. It is taken as the maximum PC1 value during a warming event.

6.3 Training Sample

The main property of a *supervised* statistical model is that fixed sets of input and output objects are presented to it during training. The output is often called the *truth*, which has to be obtained externally. In this study, it is made use of, among other, the zonal mean zonal wind at 60° N and 10 hPa to receive time series of the four vortex states. The *training sample* is called the set of data that is presented to the statistical model during training. The statistical model will learn from the training sample. It can then be evaluated without using anything but the polar-cap temperature and the external factors. Here, the model learns the patterns between the given inputs and the vortex variability, making it possible to classify warming events in frequency, intensity, and duration, but also to learn about impacts and relationships of the input factors.

Four classification time series have been produced, representing four different states of the Arctic stratosphere. The first three are major (W^{major}), minor (W^{minor}), and final (W^{final}) stratospheric warmings. The last is the undisturbed state (W^{undis}) in which no stratospheric warmings take place. Please note that W^{undis} does not denote that the polar vortex is not perturbed at all. It simply denotes the absent of stratospheric warming events. Three time series are used to calculate the training sample: PC1, the zonal mean zonal wind at 60° N and 10 hPa, $U^{10,60\mathsf{N}}$; and the long-term mean of the 30 hPa polar-cap temperature, $\overline{T^{30}}$.

The training sample is computed from ERA as follows. Firstly, the disturbed state W^{dis} at time t is defined as

$$W_t^{\mathsf{dis}} := \begin{cases} 1 & \mathrm{PC1}_t > 1 \\ 0 & \mathrm{otherwise} \end{cases} , \tag{6.1}$$

which means that PC1 needs to exceed one standard deviation. This only happens during winter time. The value of one sigma is relatively robust towards the number of derived major warmings, and leads to just the right amount of average minor and final warmings per winter compared to observations. The *undisturbed state* W^{undis} is now given by

$$W_t^{\mathsf{undis}} := 1 - W_t^{\mathsf{dis}} , \tag{6.2}$$

which denotes the state that is least disturbed. Please note that the polar stratosphere is constantly perturbed by the dissipation of planetary waves [Labitzke and van Loon, 1999]. The next task is to

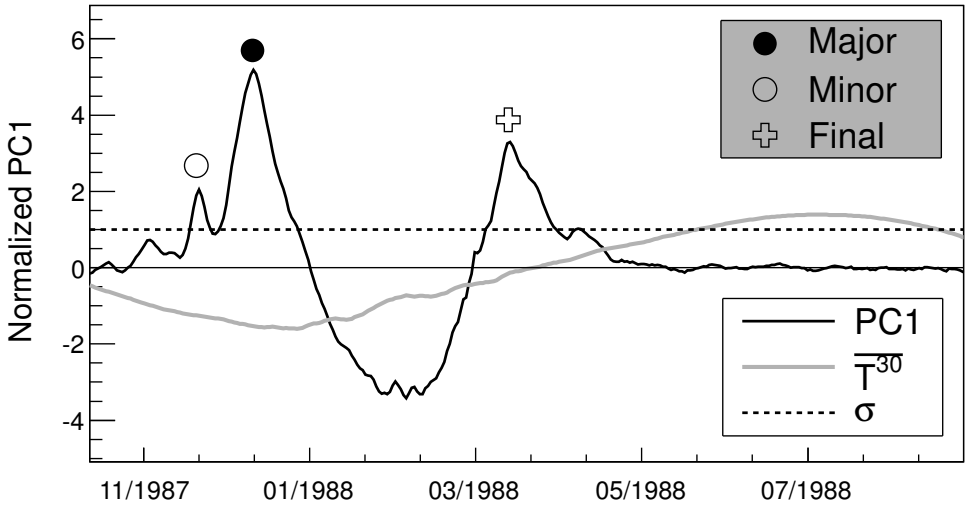

Figure 6.2: Normalized PC1 for the winter 87/88. The long-term mean of the 30 hPa temperature $\overline{T^{30}}$, the standard deviation $\sigma = 1$, and the estimated stratospheric warming events are displayed. Labeled is the first day of the particular month in a given year.

extract major, minor, and final warmings from W^{dis}. It is started with final warmings. $\overline{T_t^{30}}$ denotes a temporal measure so that $\overline{T_t^{30}} < 0$ represents the winter and $\overline{T_t^{30}} > 0$ the summer period ($\overline{T_t^{30}}$ is normalized). Therefore, values close to zero represent the transition between winter and summer or vice versa.

To classify major final warmings (referred to simply as *final warmings*), the following definition is appropriate

$$W_t^{\text{final}} := \begin{cases} 1 & W_t^{\text{dis}} = 1 \quad \wedge \quad \overline{T_t^{30}} \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{6.3}$$

which implies that disturbed states that happen in the transition phase from winter to summer are counted as final warming events. There are no disturbed states at the transition from summer to winter.

In order to determine major warmings, $U^{10,60\text{N}}$ needs to be incorporated. According to, e.g., Charlton and Polvani [2007], a major warming event takes place if $U^{10,60\text{N}} < 0$ (easterlies) during the winter time. Therefore, the *major warming state* is defined as

$$W_t^{\text{major}} := \begin{cases} 1 & W_t^{\text{dis}} = 1 \ \wedge \ W_t^{\text{final}} = 0 \ \wedge \ U_t^{10,60\text{N}} < 0 \\ 0 & \text{otherwise} \end{cases}. \tag{6.4}$$

A neighborhood of 5 days was added in which the zonal mean zonal wind can become easterly. The peak in temperature in the middle stratosphere during a major warming is usually a few days earlier

than the wind reversal in 60° N. The *minor warming state* is now simply given by

$$W_t^{\text{minor}} := \begin{cases} 1 & W_t^{\text{dis}} = 1 \ \wedge \ W_t^{\text{final}} = 0 \ \wedge \ W_t^{\text{major}} = 0 \\ 0 & \text{otherwise} \end{cases} . \tag{6.5}$$

During the procedure of computing an appropriate training sample, it was ensured that contiguous events in W^{dis} where assigned to only one type of warming state. The warming states fulfill the condition

$$W_t^{\text{major}} + W_t^{\text{minor}} + W_t^{\text{final}} + W_t^{\text{undis}} = 1 \quad \forall t \, . \tag{6.6}$$

For instance, the results for the winter 1987/88 are displayed in Fig. 6.2. The minor, major, and final warming events are observed clearly. The time axis labels indicate the first day of the particular month in a given year.

6.4 Memory in the System

Since there might be certain memory in the system, an estimate of the temporal lags of the external factors (QBO, ENSO, SFL) that minimize the classification error has to be determined. For reasons of simplicity and to reduce computational efforts, this calculation is restricted to a linear classification procedure using linear discriminant analysis (see Section 3.2) and only one target. This target has been chosen to be W^{major} as major warmings are of greatest interest.

A temporal lag larger than zero for SFL does not seem to reduce the classification error at all. Therefore, the SFL lag has been fixed to zero, and only the lags for QBO and ENSO have been varied between 0 and 180 days. An analysis with a step size of 1 day has been performed to find the optimal lags of 93 days for the QBO and 140 days for ENSO. These lags minimize the classification error and are used in all further analysis steps.

After estimating a set of lags for the external factors, it is interesting to calculate linear correlations between all input time series. It is generally favorable to use uncorrelated input variables when facing classification problems. The correlation matrix (not shown) reveals that there is no correlation apparent between any of the input variables. This also holds when keeping all time series at zero lag.

6.5 Statistical Methods for Classification

In this section, three different supervised statistical methods previously introduced in Chapter 3 are compared with respect to their ability to classify sudden stratospheric warming events. These methods are linear discriminant analysis (LDA), the support vector machine with linear kernel (LSVM), and the multi-layer perceptron (MLP). LDA and LSVM represent the group of linear classifiers, whereas MLP is a nonlinear classifier.

Table 6.1: Performance measures (see Section 3.6.1) for LDA, LSVM, and MLP. The largest value (best performance) is underlined for the particular class and the particular performance measure.

	S			ε_S			I_{ROC}		
Class	LDA	LSVM	MLP	LDA	LSVM	MLP	LDA	LSVM	MLP
Major	0.857	0.814	<u>0.864</u>	0.490	0.542	<u>0.945</u>	0.984	0.981	<u>0.987</u>
Minor	0.831	0.733	<u>0.851</u>	0.080	0.107	<u>0.935</u>	0.963	0.941	<u>0.983</u>
Final	0.822	0.527	<u>0.862</u>	0.044	0.095	<u>0.950</u>	0.953	0.801	<u>0.997</u>
Undisturbed	0.898	<u>0.909</u>	0.882	0.869	<u>0.985</u>	0.981	0.995	0.998	<u>0.999</u>

The tuning parameters for LSVM and MLP have been chosen somewhat intuitively for this comparison (LDA does not have tuning parameters). For LSVM, the cost parameter C was varied between 0.1 and 10 and the value with the best performance ($C = 1$) was selected for further analysis. For the MLP, 10 neurons in the first and 5 neurons in the second hidden layer are chosen. These values are of the same order as the number of inputs to avoid overfitting. The MLP was trained 10 times with different, randomly chosen initial parameters and the realization with the best performance was kept. The training for each method was performed in such a way that events where assigned alternating to train and test data sets.

The classification results are presented in Table 6.1 for LDA, LSVM, and MLP with respect to the performance measures introduced in Section 3.6.1. The largest value (best performance) is underlined for the particular class and performance measure. Firstly, MLP clearly outperforms the linear models in all performance measures when classifying stratospheric warmings. Out of the linear models, LSVM performs better than LDA for the warming classes with respect to ε_S but worse with respect to S and I_{ROC}. If the goal is to only discriminate between undisturbed and disturbed states, LSVM is even slightly better than MLP. This is not unexpected since the only difference between a disturbed and undisturbed Arctic stratosphere is a simple linear cut on PC1 (see Eq. 6.1). The goal of this study is the correct classification of stratospheric warmings. Hence, MLP clearly wins this method comparison with respect to the given performance measures. MLP seems to be able to classify stratospheric warmings rather well as all performance measures are close to one. Hence, MLP is the method of choice for the following analysis. In the next section, the MLP analysis is explained in greater detail and a pathway towards an optimal MLP model architecture is presented.

6.6 Optimal Model Architecture

Neural networks are widely used methods for efficient pattern recognition [Ripley, 1996]. Here, an artificial neural network recognizes patterns in temperature anomalies and external factors to classify stratospheric warming events as major, minor, and final warmings. More specifically, the neural network here is a multi-layer perceptron (MLP) in which all neurons of a certain layer are connected via synapses to all neurons in the neighboring layers (see Section 3.4). The training is

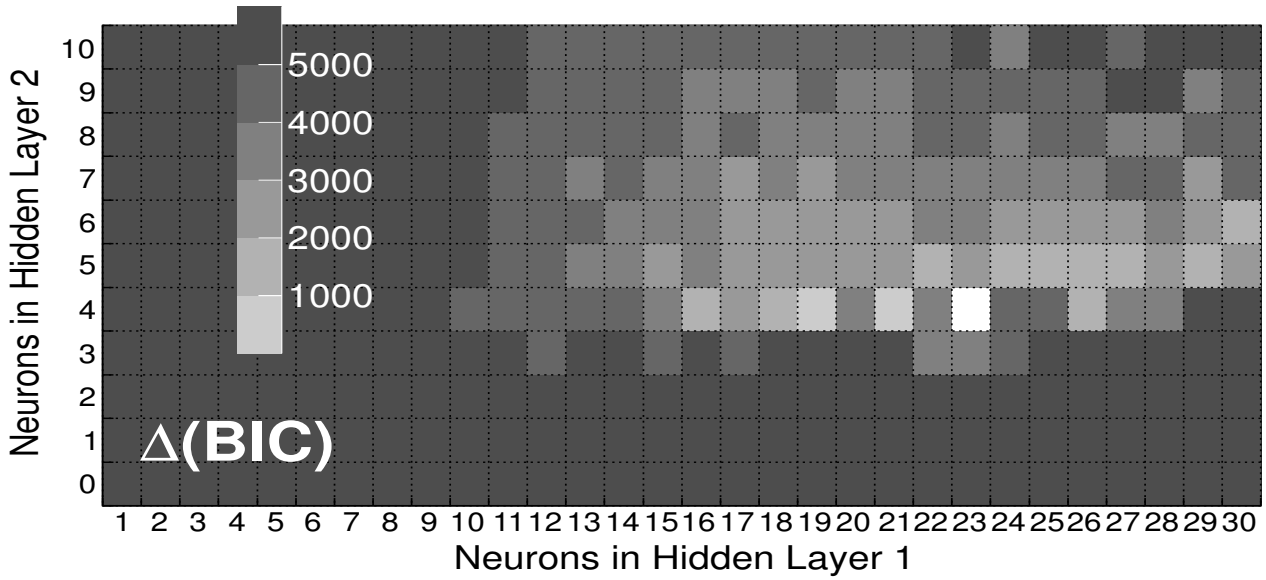

Figure 6.3: The BIC differences according to Eq. 6.7 for varying number of hidden neurons. The white square ($\Delta = 0$) denotes the optimal model architecture with 23 neurons in the first and 4 neurons in the second hidden layer. Please note the valley of small BIC values around the optimum.

performed so that events are assigned alternating to train and test data set. The input layer consists of five input neurons, which are $\overline{T^{30}}$, PC1, QBO, ENSO, and SFL. The output layer consists of four neurons representing four different states of the polar stratosphere. The first three are major (W^{major}), minor (W^{minor}), and final (W^{final}) stratospheric warmings. The last is the undisturbed state (W^{undis}), in which no stratospheric warmings take place.

The *optimal model architecture* of the MLP is estimated. The number of hidden layers as well as the number of hidden neurons within these layers need to be determined. The dimensions of input and output layers have been specified in Section 6.1. Each MLP setting is considered to be a separate statistical model. Methods from information theory (see Section 3.7.1) were shown to have remarkable ability to discriminate between statistical models. In comparison with cross-validation (see Section 3.7.2), this approach is computationally a lot less expensive and leads to the model setting with the best descriptive power whereas cross-validation focuses more on forecasting (see Section 3.7). As mentioned above, events are assigned alternating to train and test data set, therefore, incorporating a simple cross-validation with neighboring events that helps to avoid overfitting.

The information criterion leading to the optimal model architecture in this study is the *Bayesian Information Criterion* BIC as given by Eq. 3.40. This version of the BIC is applicable under the assumption that the errors are independent and identically distributed according to a Gaussian distribution [Priestley, 1983]. This assumption holds for this problem (not shown). The BIC can be understood as an estimator for the balance between explained variance and the number of free model parameters. The model with the smallest information criterion of all tested models is the

preferred model. Hence, the BIC differences can be defined as

$$\Delta_i = \mathrm{BIC}_i - \mathrm{BIC}_{\mathrm{min}} \ , \tag{6.7}$$

where $\mathrm{BIC}_{\mathrm{min}}$ denotes the minimal BIC value within the sample of tested models and i one model out of this sample ($\Delta = 0$ for the best model).

In order to determine the optimal model architecture, the MLP needs to be trained many times with different model configurations. The MLP training has been repeated 10 times with different, randomly chosen initial parameters for each model configuration. In order to reduce the effect of local minima, the resulting σ_e^2 used to calculate the BIC is taken as the mean of those 10 optimizations.

The number of hidden neurons is varied in the hidden layers. The results of Eq. 6.7 are displayed in Fig. 6.3 where the white square indicates $\Delta = 0$. This procedure was repeated using the Akaike Information Criterion as given by Eq. 3.42, which led to a more complicated model architecture with significantly more free parameters and was therefore rejected. The resulting optimal model setting has two hidden layers with 23 neurons in the first and 4 neurons in the second layer. The MLP has now been trained 100 times with this specific architecture. The run with the smallest error is chosen. The classification results of this run are presented in the following.

6.7 Probabilities of Stratospheric Warmings

In this section first classification results based on conditional probabilities for each of the classes are presented. Additionally, the statistical method is validated. To ensure that the MLP response values can be interpreted as conditional probabilities, the value y_i of output neuron i needs to be transferred via the softmax function as given by Eq. 3.23. Having computed conditional probabilities, a threshold value is determined for each class above which a certain probability is significantly different from the background. This is called the cut_i at class i. To do so the area-normalized background probability distribution $P_{B,i}$ is integrated for each class i so that

$$\alpha = \int_0^{\mathrm{cut}_i} P_{B,i}(x)\mathrm{d}x \ , \tag{6.8}$$

where $\alpha = 0.999$, so that a probability p_i greater than cut_i is significantly different from the background at a confidence level of 99.9%. The resulting values are 0.32, 0.34, and 0.25 for major, minor, and final warming class, respectively. The cuts are relatively small, which indicates a good classification performance.

6.7.1 Three Sample Winters

Further insights into the MLP response are now envisaged. Three adjacent sample winters are selected that include all three types of warming events. Fig. 6.4 shows the evolution of the prob-

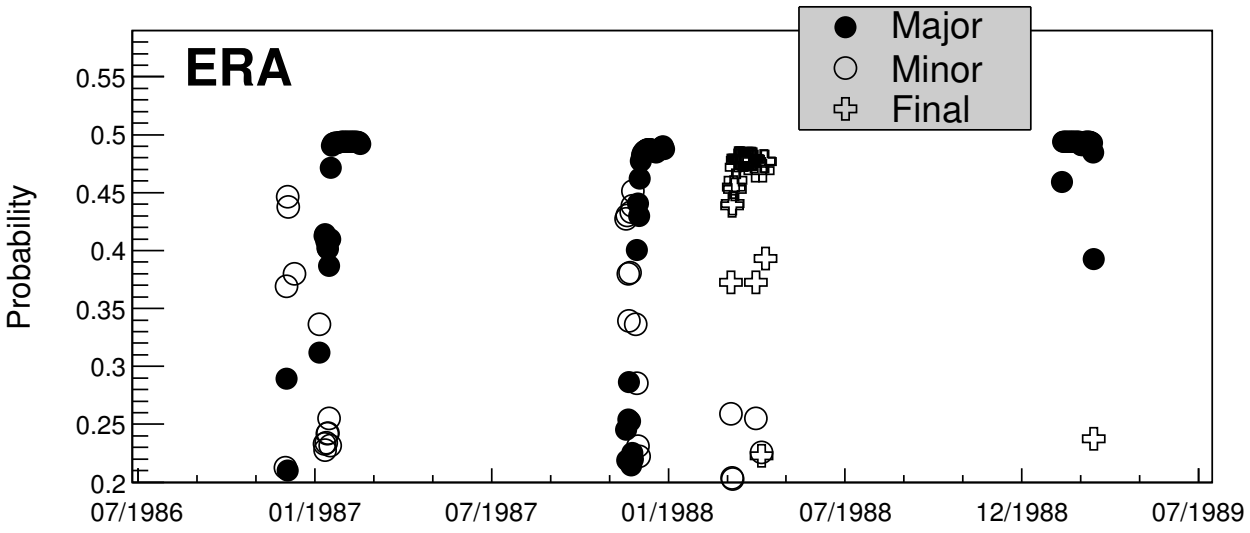

Figure 6.4: Evolution of the probabilities in ERA for major, minor, and final warmings for the three winters in the period from summer 1986 to summer 1989. Labeled is the first day of the particular month in a given year.

abilities for major, minor, and final warmings for the period from summer 1986 to summer 1989. Results for ERA (training data set) and NCEP (validation data set, not shown) are very similar. The winter 1987/88 appears to be the most variable on this period. A minor warming in November 1987, lasting about 5 days, is observed. A major warming [Baldwin and Dunkerton, 1989; Naujokat et al., 1988] takes place in the beginning of December 1987, lasting about 20 days. A short minor warming appears as a precursor to this major warming. Ultimately, a final warming lasting about 15 days takes place in March [Labitzke and Naujokat, 2000]. The probabilities shown in Fig. 6.4 give a good representation of what was observed (compare Fig. 6.2). The classification performance is now assessed in greater detail.

6.7.2 Classification Performance

In addition to the performance measures used in Section 6.5, the classification performance is assessed with respect to the mean prediction error (MPE) as defined in Eq. 3.30. For a perfect classification it is expected that MPE equals zero. Table 6.2 presents the performance measures for each class as calculated from the ERA classification results. A very high classification performance is obtained. The separation S, the signal efficiency ε_S, and the integral of the ROC curve I_{ROC} are very close to one for all classes. This represents a very good ability of discriminating signal from background events.

MPE is very close to zero for all classes, which implies that only in rare cases the MLP response is not close to the data that the MLP has been trained with (the training sample, see section 6.1). Therefore, the MLP is able to reliably detect major, minor, and final warming states, and of course, the undisturbed state.

Table 6.2: The performance measures for the optimal MLP setting for each class.

Class	S	$\varepsilon_S(\varepsilon_B = 0.01)$	I_{ROC}	MPE
Major	0.992	0.969	0.996	0.0023
Minor	0.985	0.929	0.990	0.0026
Final	0.962	0.968	0.987	0.0015
Undisturbed	0.990	0.991	0.995	0.0010

Table 6.3: Relative impact in percent on the MLP response according to Eq. 6.9 for each input neuron and output class.

Input	Major	Minor	Final	Undisturbed
$\overline{T^{30}}$	22.2	22.4	54.0	0.8
PC1	26.9	23.7	31.2	97.5
QBO	19.3	19.9	4.6	0.7
ENSO	17.4	17.9	4.9	0.6
SFL	14.3	15.9	5.3	0.4

6.7.3 Impact of the Input Neurons

It is of great interest to estimate the individual *impact* of the five input neurons on the MLP response. This gives an insight into the statistical importance of each of the input factors. The impact $I_{i,k}$ of input factor i on output class k is simply chosen to be the variance of MLP response differences given by

$$I_{i,k} = \mathrm{Var}(y_k^{(i)} - y_k) \ , \tag{6.9}$$

where y_k denotes the MLP response at output neuron k and $y_k^{(i)}$ the corresponding MLP response where input factor i was set to zero. If an input neuron had no impact on the MLP Eq. 6.9 would give zero. Table 6.3 presents the relative impact in percent on the MLP response according to Eq. 6.9 for each input neuron and output class.

It is observed that the impacts are quite different for different output classes. For the undisturbed case only PC1 plays an important role. This is expected as the undisturbed state is simply defined by a linear cut on PC1 (see Section 6.3). The final warming state is mostly governed by PC1 and $\overline{T^{30}}$ since the definition of the final warming state was only based on these two factors (see Section 6.3).

When looking at major and minor warming states, the external factors become more important and necessary to discriminate major from minor warmings. The QBO shows the largest impact, followed by ENSO and the solar cycle in agreement with previous studies [e.g., Labitzke and Kunze, 2009b; Camp and Tung, 2007a,b; Mitchell et al., 2011] that also investigated the impact of these forcings and found a similar ranking. Hence, the neural network combines QBO, ENSO, and SFL in a nonlinear fashion to distinguish between major and minor stratospheric warmings. Therefore, the external factors, namely QBO, ENSO, and SFL, should be incorporated in order to classify

stratospheric warmings successfully. It was mentioned earlier that there is practically no linear correlation between any of the input time series. However, as Table 6.3 shows, there exist nonlinear combinations of input factors that lead to different stratospheric warming states.

6.8 Stratospheric Warming Climatologies

This section presents stratospheric warming climatologies extracted from resulting probabilities for 52 winters from 1958 through 2010. In order to identify stratospheric warmings, a threshold is defined above which a signal in one of the output neurons is counted as an event signal. An event signal has to be significant; hence, it needs to exceed the cut values (see Section 6.7). In order to get an estimate for the training data set ERA, the first derivative $\mathrm{d}Q/\mathrm{d}p$ of the cumulative density function of the response distribution is calculated of each warming class. As an increasing derivative denotes a regime change, the thresholds is defined where $\mathrm{d}Q/\mathrm{d}p$ starts rising from its constant level with increasing quantiles. The resulting thresholds for ERA are 0.41, 0.41, and 0.45 for major, minor, and final warming events, respectively. It is found that the resulting ERA warming event numbers and distributions are not sensitive with respect to slightly different thresholds.

As the validation set NCEP is a priori unknown, and to avoid counting events caused by a possible overfitting, a reasonable NCEP threshold needs to be found that is larger than any of the ERA thresholds but smaller than the theoretical limit given by Eq. 3.23. An NCEP threshold of 0.47 for all warming classes was selected leading to reasonable distributions and event numbers as presented in the following. The resulting NCEP events are more sensitive with respect to this threshold than the ERA events but can still be changed in the percentage range and the event numbers and monthly distributions would not change significantly.

6.8.1 Warming Events

To obtain stratospheric warming events, contiguous warming days need to be grouped together. To do so, minimal temporal distances between adjacent warming events are defined. Is this distance exceeded without output neuron i above the given probability threshold, then warming event i is finished and a new warming event may eventually take place. For these distances 30 days are chosen for major warmings, 5 days for minor warmings, and 100 days for final warmings. The number for final warmings is rather arbitrary as they may only take place once a year during the transition from winter to summer circulation. 30 days are selected for major warmings because it is known from observations [e.g., Labitzke and van Loon, 1999] that major warmings may last 20 days but that neighboring major warmings in the same winter are at least one month apart. The relatively short period of 5 days for minor warmings was chosen since minor warmings are usually not proceeded by a great cooling in the Arctic stratosphere, as major warmings are [Labitzke and van Loon, 1999; Charney and Drazin, 1961]. Therefore, adjacent minor warming events can be closer than major warming events.

Table 6.4: Total number of stratospheric warming events (top) and relative number of events per year (bottom) for the different warming classes and the two data sets. The uncertainties are given in parentheses (standard error of mean).

Data	Major	Minor	Final	Total
ERA	31	74	27	132
NCEP	26	76	28	130
ERA	0.6 (0.1)	1.4 (0.2)	0.5 (0.1)	2.5 (0.3)
NCEP	0.5 (0.1)	1.5 (0.2)	0.5 (0.1)	2.5 (0.3)

First results of this procedure are shown in Table 6.4 for ERA and the validation data set NCEP. Absolute number (upper part) and relative number (lower part) of warming events are presented. It is observed that values for ERA and NCEP are very similar for all warming classes. This indicates a successful validation of the classification procedure using NCEP. Only the major warming case shows slightly less events in NCEP than in ERA. This discrepancy for major warmings has also been reported by Charlton and Polvani [2007].

To summarize, there is a major warming event approximately every other year whereas minor warmings happen at least once a year on average. (Major) Final warmings take place every second year, too. These results are in good agreement with Charlton and Polvani [2007] who find approximately 0.6 SSWs per winter. Labitzke and Naujokat [2000] find approximately 0.5 major mid-winter warmings, approximately 1 minor warmings (half of which are Canadian warmings), and 0.25 major final warmings per winter. The differences between ERA and NCEP found in this study are due to differences in the two data sets, particularly in PC1 and ENSO during the pre-satellite era before 1979.

6.8.2 Change in Circulation

The question remains whether the detected major and final warming events lead to a vortex breakdown and therefore a change in circulation (easterly zonal winds) in the stratosphere in mid-latitudes. Minor warmings should only slow down the circulation but not reverse it. In order to tackle this question, the zonal mean zonal wind at 60° N and 10 hPa is incorporated. If the zonal wind is negative (easterlies) then a change in circulation took place and the polar vortex broke down. An interval of 20 days around the central warming date of major and final warmings was considered to find the minimum zonal wind.

The result of this analysis is shown in Fig. 6.5 for all warming classes and both data sets. Values for minor warmings temporally very close to major or final warmings are not shown as they lead to ambiguous wind results. The numbers represent the winter in which the warming took place; e.g. 98 means the winter 1998/99. The zonal wind reversed for almost all major and final warmings in ERA and NCEP, which confirms the classification procedure.

There are only a few clear falsely detected major warmings for which the vortex was disturbed,

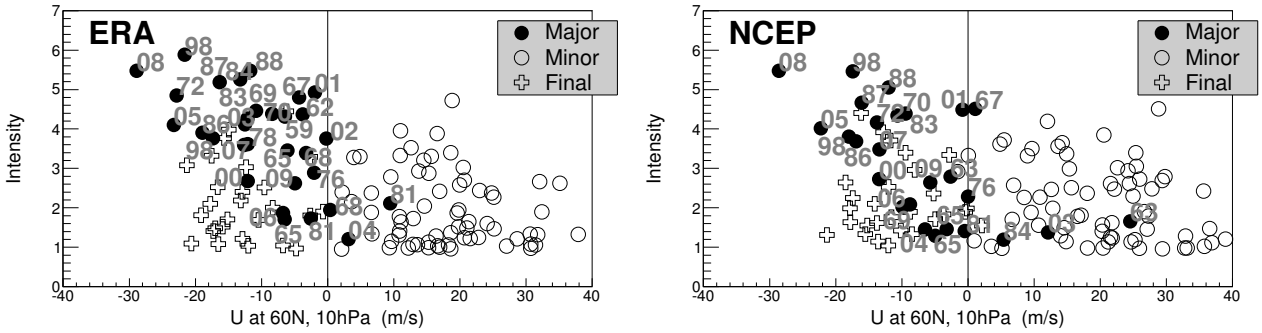

Figure 6.5: Scatter diagram of stratospheric warming intensity against the zonal mean zonal wind at 60° N and 10 hPa. Numbers represent winters; (e.g. 98 means the winter 98/99) in which major warming(s) took place. The results are shown for ERA (left) and NCEP (right). Values for minor warmings temporally very close to major or final warmings are not shown as they lead to ambiguous wind results.

and the circulation slowed down but did not reverse. In ERA these are the winters 81/82 and 04/05. In NCEP these are the winters 63/64, 84/85, and 03/04. It is the nature of a statistical method that it is never 100% effective. However, all final warming events were classified correctly. Despite of a few differences, the classified stratospheric warmings are in good agreement with previous studies [Charlton and Polvani, 2007; Labitzke and Naujokat, 2000]. None of the detected minor warming events led to a change in circulation.

6.8.3 Stratospheric Warming Frequencies

The classification results are now analyzed and presented in more detail with respect to their occurrences and intensities. Monthly climatologies of major, minor, and final warmings are shown in Fig. 6.6. The uncertainties are displayed as error bars. Firstly, the distributions for ERA and NCEP are similar. Most major warmings take place in January. Minor warmings happen all throughout the winter but most take place in February for ERA and January for NCEP, whereas final warmings clearly peak in March and April. There are no major warmings taking place in November, which is in agreement with observations [Labitzke and Naujokat, 2000]. Major warmings show highest intensities with large variability followed by minor and final warmings. As expected, the minor warming intensities peak in January and decrease towards beginning and end of the winter. The final warming intensities are also very variable and peak in March.

Charlton and Polvani [2007] show monthly distributions for major warmings retrieved from a classification method based on the zonal mean zonal wind at 60° N and 10 hPa. These results are similar to the distribution for major warmings shown in Fig. 6.6. There have been a few SSWs found by Charlton and Polvani [2007] in November which were most likely Canadian warmings. They found more SSWs in March simply because some of those are counted as final warmings in the present analysis.

It is of great interest to investigate the temporal evolution of the three warming classes over the

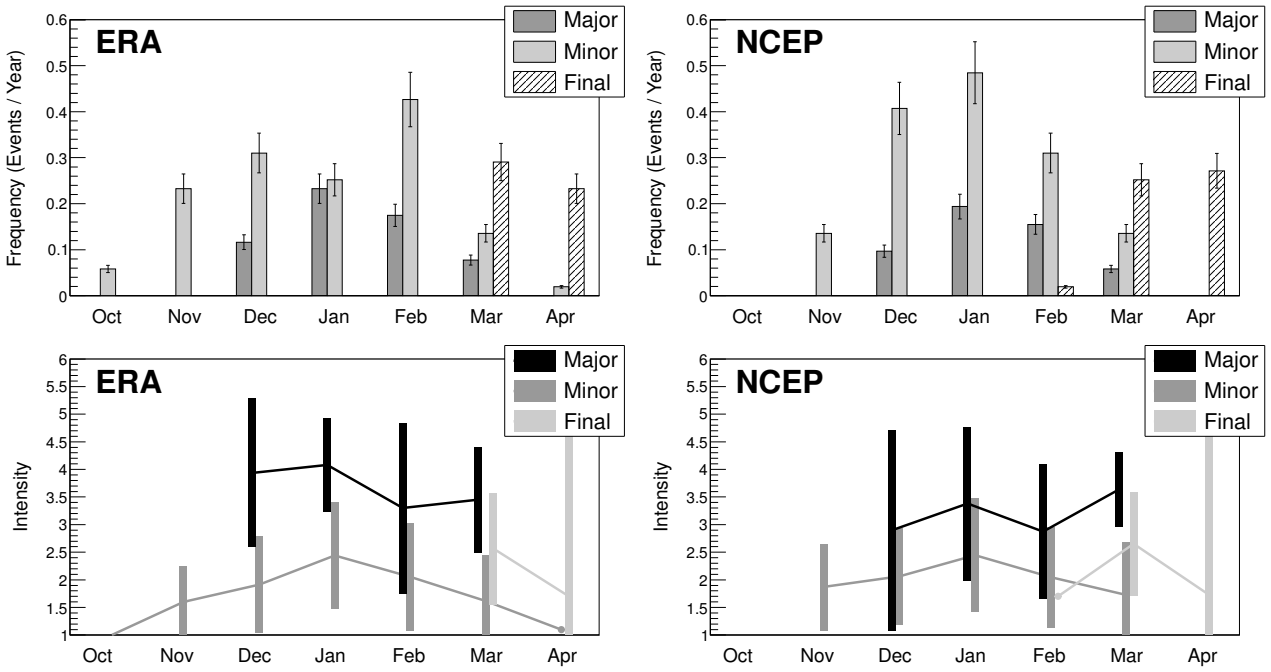

Figure 6.6: Monthly distributions in events/year (top) and intensity (bottom) of the three warming classes for ERA (left) and NCEP (right). The error bars represent the standard error of mean (top) and the standard deviation (bottom), respectively. Please note the different color scheme for frequency (histograms) and intensity (graphs).

52 year period. Their frequency of occurrence and intensity in bins of 4 years is presented in Fig. 6.7. The frequency distributions resemble observations rather well [Labitzke and Naujokat, 2000]. For instance, the clear minimum of major warming activity observed in the 1990s is obtained. There are also periods of higher major warming activity in the 1970s. Minor warmings were especially frequent during the 1980s and 1990s. Final warmings do not show significant occurrence variabilities. The results for ERA and NCEP in Fig. 6.7 are again qualitatively similar. Differences appear mostly during the pre-satellite era before 1979. In comparison to Charlton and Polvani [2007], differences for the major warming case are mainly due to different methodologies and classification strategies.

The intensities presented in Fig. 6.7 are also similar in ERA and NCEP. A great decrease in major warming intensity is observed during the 1970s and large intensities in the past 30 years. The mean minor warming intensities seem to be rather constant through out the whole period whereas the final warming intensities show a peak in the 1980s to then decrease to minor warming levels.

Mean intensity and the corresponding standard deviation of the three warming events and their duration in days for ERA and NCEP is presented in Table 6.5. The results for ERA and NCEP agree rather well. On average, major and final warmings last about 20 days and minor warmings only 8 days. There is a large variability in duration as the standard deviation takes values of about 10 days for each warming class. On average, Major warmings are twice as intense as minor and final warmings with medium variability.

Table 6.5 also shows the linear correlation between intensity and duration for each warming

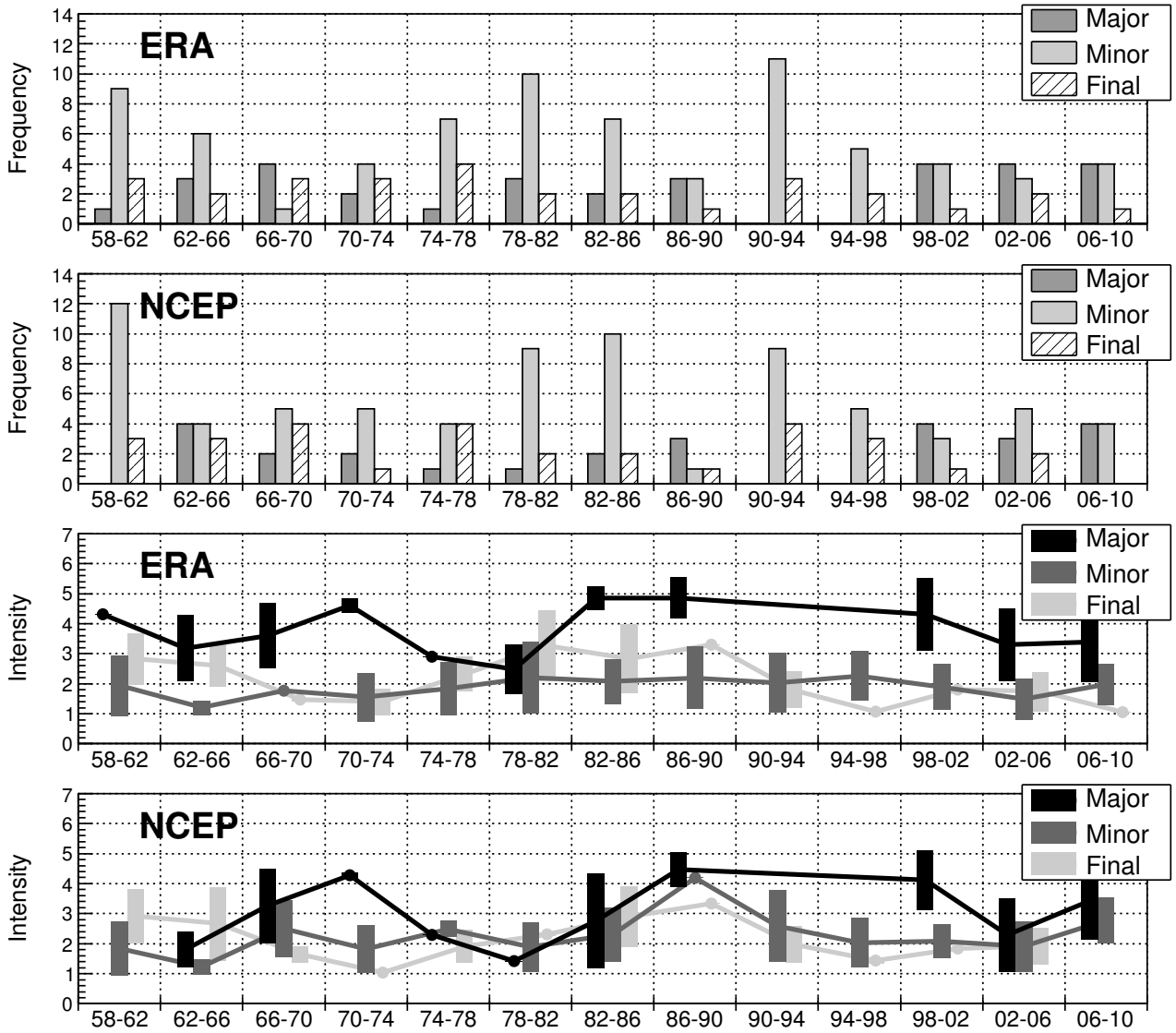

Figure 6.7: Distributions in bins of 4 years of occurrence (top) and intensity (bottom) of the three warming classes for ERA and NCEP, respectively. All bins start at September 1 and stop at August 31 of the respective years. Error bars represent one standard deviation around the mean. Please note the different color scheme for frequency (histograms) and intensity (graphs).

class. All correlation factors are significant (t-test) at the 95% confidence level. For ERA, all correlation factors are greater than 0.6 which leads to the expected conclusion that warmings with larger intensities generally last longer, and vice versa. For NCEP, the correlation factors are slightly smaller.

6.8.4 Marginalized Probability Distributions

The neural network can be considered as a function (classifier) mapping from a 5-dimensional input space to a 4-dimensional probability space. In order to retain an understanding of the relationships between the input factors despite the high dimensionality, the resulting probability distributions are

Table 6.5: Mean intensity in standard deviations of stratospheric warming events and their mean duration in days for the different stratospheric warming events in ERA (top) and NCEP (bottom). The corresponding standard deviation is given in parentheses. The correlation between duration and intensity is also given. All correlation factors are significant (t-test) at the 95% confidence level.

Data	Class	Intensity	Duration	Correlation
ERA	Major	3.7 (1.2)	23.0 (10.7)	0.61
	Minor	1.9 (0.9)	8.4 (8.3)	0.75
	Final	2.2 (1.0)	20.1 (10.9)	0.67
NCEP	Major	3.2 (1.4)	16.7 (12.3)	0.53
	Minor	2.1 (1.0)	8.9 (10.5)	0.52
	Final	2.2 (0.9)	20.2 (10.5)	0.41

marginalized. Motivated by previous studies, the relationships between QBO, ENSO, and SFL are particularly interesting. Therefore, these factors have been varied and the resulting MLP response investigated.

PC1 has been fixed and the responses have been averaged for the mid-winter between December and February (DJF). Additionally, the results have been split for solar maximum and solar minimum conditions where a value of 120 solar flux units (sfu) of the f10.7 cm solar radio flux was used to separate the two regimes. The resulting marginalized probability distributions are shown in Fig. 6.8 for the major warming state. The shading denotes the probability of the occurrence of a major warming and the black thick line an approximately significant probability of 0.3. The numbers in Fig. 6.8 represent the winter in which the major warming took place; e.g. 87 means the winter 1987/88.

A moderately high disturbance of PC1 $= 3$ (compare Fig. 6.5) was selected, implying that the condition for the disturbed state is fulfilled and the MLP discriminates between two states; major warmings and minor warmings. Due to the aforementioned averaging, the patterns in Fig. 6.8 represent climatological mean states for DJF. These patterns are highly nonlinear, which emphasizes the usefulness of a nonlinear statistical method. Previous studies have shown, that the considered forcings interact and create a complex and nonlinear dynamical link [e.g., Calvo et al., 2009; Richter et al., 2011].

The superimposed boxes in Fig. 6.8 represent a schematic frequency distribution of QBO and ENSO; the larger the box the greater the population density of a particular bin, i.e. a large box stands for a high frequency of a particular combination of QBO-ENSO values and vice versa. Hence, these population densities are naturally different for solar maximum and minimum. Highly populated regions are observed, but also combinations of QBO and ENSO that have not been seen in the data at all. The larger the population, the more the MLP response can be trusted. In regions with zero population (no boxes), the MLP predicts probabilities. Considering the good validation results for NCEP, which is an unseen data set, the MLP predictions can be considered trustworthy. Nevertheless, they need to be confirmed by data from chemistry-climate model simulations.

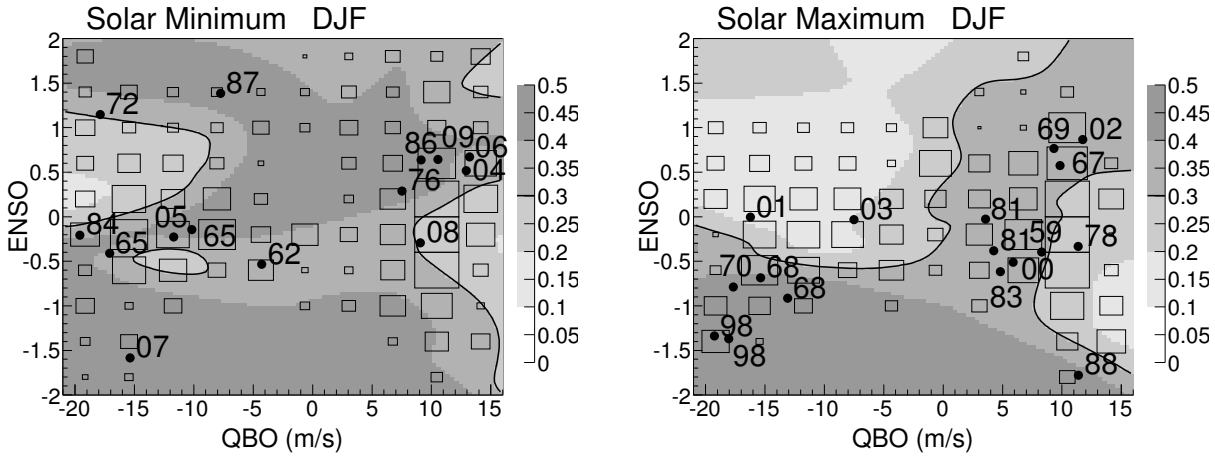

Figure 6.8: Marginalized probability distributions (shading; black line denotes $p = 0.3$) for the **major warming state** depending on ENSO and the QBO for solar minimum (left) and solar maximum (right) for PC1 $= 3$, denoting a moderate vortex disturbance. The darker the shading the higher the probability for a major warming. The numbers denote the winter of a major warming; e.g. 87 denotes the winter 87/88. The superimposed boxes represent a schematic frequency distribution of QBO and ENSO; the larger the box the greater the population density for a particular bin, i.e. a large box stands for a high frequency of a particular combination of QBO-ENSO values.

There are two main regions that are not populated. The first is the region of large negative ENSO values (La Niña) and small absolute QBO values (around zero) for both solar maximum and minimum. La Niña events are rather rare and the transition between QBO west and QBO east and vice versa is very fast (often within a month), whereas a QBO phase (east or west) can last about a year. The other underpopulated region is that of large positive ENSO values (El Niño) during solar maximum for almost all values of the QBO. Hence, El Niño events are only rarely found during solar maximum conditions.

Fig. 6.8 presents various probability features for major warmings. Despite the averaging, almost all major warming events fall into the significant area of $p \geq 0.3$, indicating a robust classification. Please note the aforementioned averaging over SFL regimes and the mid-winter implying that probabilities for a specific event may be different from what is shown in Fig. 6.8. There are regions of high probabilities for QBO west and solar maximum conditions, as also found by Camp and Tung [2007a]; Labitzke and Kunze [2009b]. However, there is a region for strong QBO west in both solar maximum and minimum, in which moderate and La Niña-like ENSO events show only small probabilities. The high population density in this region makes the probabilities particularly trustworthy. This indicates strong nonlinear relationships between QBO and ENSO as also found by e.g., Calvo et al. [2009]; Richter et al. [2011]. Linear interrelationships as emphasized by e.g., Camp and Tung [2007a,b] and Labitzke and Kunze [2009b] are not sufficient to explain this pattern.

The very intense major warming of the winter 08/09 (solar minimum, QBO west, and slightly negative ENSO values) is very close to the significant region in Fig. 6.8. Hence, this major warming along with the major warming in 06/07 are part of the nonlinear rules determined by the MLP,

whereas these events have been previously treated as exceptions from linear rules [e.g., Labitzke and Kunze, 2009a].

Despite the high probabilities, only a few major warmings are found to happen during the transition from QBO west to east or vice versa (see Fig. 6.8). This is because of the aforementioned fast transition between QBO phases (west↔east). Moreover, it is known from observations [Baldwin et al., 2001] that the QBO phase transition takes place mostly during the northern hemisphere summer. By definition, sudden stratospheric warmings take place only during the winter time.

During QBO east and solar maximum conditions, only negative ENSO values show significant probabilities. During QBO west, moderate and El Niño-like conditions lead to significant major warming probabilities. For solar minimum and QBO east, strong positive ENSO events lead to large probabilities, too. A probability minimum is observed for ENSO values close to zero. This minimum appears also for QBO west but for slightly negative ENSO values and more dependent on the strength of the QBO. In general, the probability for a disturbance to become a major warming leading to a vortex breakdown is greater during solar minimum conditions (note the large significant area) than during solar maximum. As also found by Butler and Polvani [2011], El Niño-like and La Niña-like conditions make the occurrence of major stratospheric warmings more likely as opposed to neutral ENSO conditions. The only exception is the small major warming probability for El Niño-like conditions during solar maximum and QBO east.

6.9 Classification in CCM Simulations

This section aims at classifying stratospheric warmings in REF-B1 CCM simulations for the period from July 1960 to June 2005, considering only full winters. The CCMs are EMAC-FUB, MRI, and WACCM (see Section 4.3). The same framework as above is applied including the computation of a training sample (see Section 6.3) for each CCM and a corresponding training of the neural network incorporating the five input factors ($\overline{T^{30}}$, PC1, QBO, ENSO, SFL) while using a lag of 93 days for the QBO and 140 days for ENSO (compare Section 6.4). $\overline{T^{30}}$, PC1, and QBO are computed from the output of each CCM, whereas ENSO (Nino3.4 index) is computed from HadISST1 (see Section 4.1) and SFL is the F10.7 radio flux as in Section 6.1. Lagging the external factors in the same way as for the reanalyses above has three advantages: 1) It keeps classifications results comparable between reanlyses and CCMs, 2) The optimal neural network architecture for the CCMs is the same as for the reanalyses with 23 neurons in the first and 4 neurons in the second hidden layer (see Fig 6.3), and 3) The same event thresholds are used as for ERA which are 0.41 for major warmings, 0.41 for minor warmings, and 0.45 for final warmings. More than 95% of the warming events recognized by the neural network in each of the different CCM simulations were classified correctly. The classification was tested using the zonal mean zonal wind at 10 hPa and 60° N (compare Fig. 6.5).

Table 6.6: Relative number of stratospheric warming events per year (bottom) for the different warming classes and the three CCMs over the 45-yr period. The uncertainties are given in parentheses (standard error of mean).

Data	Major	Minor	Final	Total
EMAC	0.8 (0.1)	1.8 (0.3)	0.5 (0.1)	3.0 (0.4)
MRI	0.4 (0.1)	2.0 (0.3)	0.5 (0.1)	2.9 (0.4)
WACCM	0.6 (0.1)	1.6 (0.2)	0.3 (0.1)	2.6 (0.4)

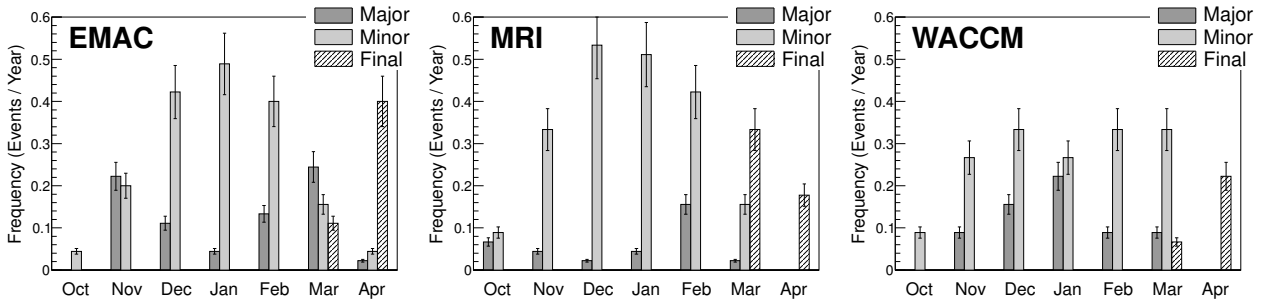

Figure 6.9: Monthly distributions in events/year of the three warming classes for EMAC (left), MRI (middle), and WACCM (right). The error bars represent the standard error of mean.

6.9.1 REF-B1 Simulations: EMAC, MRI, WACCM

Stratospheric warmings are classified in the REF-B1 simulations made with EMAC, MRI, and WACCM using the neural network approach. Relative event numbers per year are shown in Table 6.6 for major, minor, and final warmings and the different CCM simulations. The number of major warmings per year are significantly different across the CCMs. Most major warming events are found in EMAC (0.8), followed by WACCM (0.6), and MRI (0.4). These results agree well with the number of major SSWs found in chapter 4 of SPARC CCMVal [2010] where major SSWs were classified using the zonal wind criterion as in Charlton and Polvani [2007]. The number of minor warmings are statistically indistinguishable between the different CCMs averaging to approx. 1.8/yr. The number of final warmings is 0.5/yr for EMAC and MRI and significantly smaller with 0.3 for WACCM. The total number of events is again indistinguishable across the CCMs averaging to 2.8/yr.

When comparing the CCM with reanalyses results, as shown in Table 6.4, it is found that major and minor warming frequencies are best represented in WACCM whereas final warming frequencies agree well with EMAC and MRI. Also, the total number of warming events in the considered reanalyses agrees best with WACCM results. However, the sampling errors are quite high due to the short time series (only 45 years), making it difficult to distinguish between event numbers from the different CCMs. That being said, WACCM appears to have the most realistic representation in terms of average relative event numbers.

The corresponding monthly distributions of major, minor, and final warming events for the

Table 6.7: Relative impact in percent of the input factors on the MLP response for major warmings according to Eq. 6.9 and the different CCMs.

Input	EMAC	MRI	WACCM
$\overline{T^{30}}$	28.8	25.2	24.1
PC1	25.3	23.6	28.4
QBO	18.7	14.0	19.4
ENSO	15.4	21.1	15.3
SFL	11.8	15.9	12.8

different CCM simulations are shown in Fig. 6.9. First off, the major warming distributions agree well with those found in chapter 4 of SPARC CCMVal [2010]. Slight differences are due to different methodologies as already pointed out in Section 6.8. EMAC simulates most major warmings in early (Nov.) and late winter (Feb. and Mar.) similar to MRI which even simulates a few major warmings in October but only little in March. Major warmings in EMAC and MRI are underrepresented during mid-winter. The major warming distribution as found in WACCM agrees best among the CCMs with those found in the considered reanalyses (see Fig. 6.6) except for slightly too many major warmings in November and too few in February. Other than that the major warming distribution in WACCM peaks in January with approx. 0.2/yr just like the distributions in ERA and NCEP.

Despite of the overall, slightly higher number of minor warmings per year in the CCMs, the monthly minor warming distributions are quite similar to those in reanalyses. Interestingly, the minor distributions of EMAC and MRI are very similar to that of NCEP and appear to be normally distributed around mid-winter. WACCM, on the other hand, has a minor warming distribution very similar to that of ERA with an almost flat distribution throughout the winter. Only the high (0.3/yr) minor warming frequency in March, as found in WACCM, does not agree with the reanalyses. Final warmings take place in March and April where events in April are most likely to take place in EMAC and WACCM. March is the month with highest final warming frequency in MRI. In contrast, final warmings populate March and April almost equally in ERA and NCEP (see Fig. 6.6).

It is interesting to compute the statistical importance of each of the input factors. To do so, the impact on the MLP response for major warmings according to Eq. 6.9 and the different CCMs is shown in Table 6.7. The other warming states are omitted here for simplicity because impacts on minor warmings are very similar to those on major warmings, more over, final warmings are only governed by $\overline{T^{30}}$ and PC1. From Table 6.7 one can see that $\overline{T^{30}}$ and PC1 have the largest impact all throughout averaging to about 25%. For EMAC and WACCM this is followed by the QBO, as an external forcing, with approx. 18%. Moderatly important for EMAC and WACCM are ENSO with approx. 15% and SFL with approx. 12%. It is interesting to note that the impacts of the external forcings (QBO, ENSO, SFL) are very similar in EMAC and WACCM. As noted earlier, ENSO and SFL are equal for all simulations and the QBO is almost identical in EMAC and WACCM as the QBO is nudged in both CCMs (see Section 4.3). Equal forcings does not imply similar teleconnections to the polar middle stratosphere. However, this appears to be the case here as shown in Table 6.7.

It should be noted that the ranking of the input factors for major warmings in ERA, as shown in Table 6.3, is best represented in the impacts for WACCM.

MRI shows a ranking of the input factors that is different from EMAC and WACCM, only the impacts of $\overline{T^{30}}$ and PC1 are similar. A significant difference is observed for the external forcings. The QBO with 14% is less important in MRI whereas ENSO with 21% and SFL with 16% are more important. The QBO in MRI is internally generated which could be the reason for this difference across the CCMs.

In order to obtain insight into the interrelationships between the input factors during mid-winter, marginalized probability distributions for the major warming state for DJF are shown in Fig. 6.10. They are computed in the same way as the distributions for ERA presented earlier in Fig. 6.8. The top panel shows results for solar maximum and the bottom panel for solar minimum. The boxes represent a frequency distribution of QBO-ENSO pairs. As ENSO (Nino3.4 index) and SFL are equal across all CCMs, it is noted that the QBO in EMAC and WACCM is rather similar whereas the QBO in MRI is significantly different. The QBO in MRI appears to be weaker in magnitude.

The probability patterns for the major warming state in Fig. 6.10 are significantly different across the CCMs and do not have much in common with the ERA pattern in Fig. 6.8. The only pattern that agrees in most part with ERA is the WACCM pattern for solar minimum. This does not hold for solar maximum conditions. However, the CCM patterns are still highly nonlinear confirming previous studies [e.g. Garfinkel and Hartmann, 2007; Calvo et al., 2009]. They represent relationships between the external forcings and SSWs as shown in the literature (compare Section 6.8.4). At least, MRI and WACCM show an increased probability for major warmings during QBO east which could indicate that the Holton-Tan mechanism [Holton and Tan, 1980, 1982] is present in these simulations. EMAC, on the other hand, shows an increased probability for QBO west for both solar conditions. All CCM patterns point to a high probability during El Niño events (large ENSO values) in agreement with the literature on this subject [e.g., Camp and Tung, 2007b; Mitchell et al., 2011]. Work made by Butler and Polvani [2011] reported that La Niña events also lead to more major warmings. This is found for EMAC and WACCM during solar minimum and for ERA (see Fig. 6.8). It is concluded that some of the established connections between external forcings and SSWs are recovered. However, the nonlinear CCM patterns are different to eachother and do not agree with the ERA patterns. Robust teleconnection patterns could not be estimated. The non-linear relationships between the external forcings and the major warming probability are significantly different across the CCMs.

6.9.2 A Study of Sensitivity with WACCM

So far, stratospheric warmings in reanalyses and realistically forced CCM simulations (REF-B1) were classified. Here, stratospheric warmings in two artificially altered WACCM simulation runs are recognized by the neural network and results are compared to the REF-B1 WACCM run. The two altered runs are equally forced as the REF-B1 run except that 1) The QBO is not nudged resulting in a constant easterly phase which is referred to hereafter as WACCM-QBO and 2) Climatological

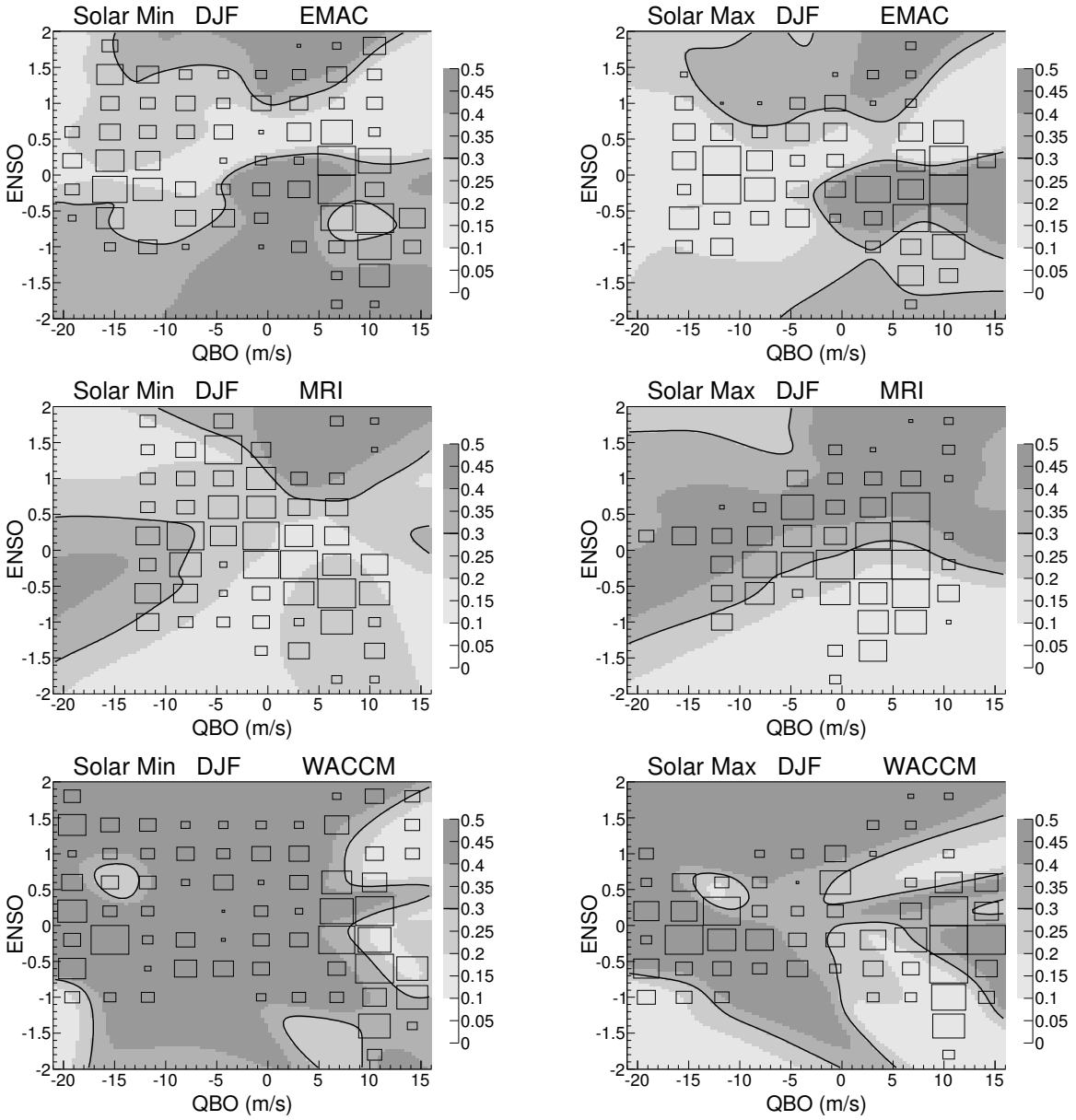

Figure 6.10: Marginalized probability distributions (shading; black line denotes $p = 0.3$) for the **major warming state** depending on ENSO and the QBO for solar minimum (left) and solar maximum (right) for PC1 $= 3$, denoting a moderate vortex disturbance. The darker the shading the higher the probability for a major warming. The superimposed boxes represent a schematic frequency distribution of QBO and ENSO; the larger the box the greater the population density for a particular bin, i.e. a large box stands for a high frequency of a particular combination of QBO-ENSO values. Distributions are shown for EMAC (top), MRI (middle), and WACCM (bottom).

SSTs are prescribed resulting in neutral ENSO phase conditions referred to hereafter as WACCM-SST. The three runs, WACCM, WACCM-QBO, and WACCM-SST, are also analyzed for 45 years from June 1960 to July 2005. Richter et al. [2011] has already partly analyzed these runs concerning sudden stratospheric warmings but only for a period of 30 years. They analyzed an additional fourth run with constant easterly phase instead of a QBO *and* climatological SSTs. Richter et al. [2011]

Table 6.8: Relative number of stratospheric warming events per year (bottom) for the different warming classes and the different WACCM sensitivity runs over the 45-yr period. The uncertainties are given in parentheses (standard error of mean).

Data	Major	Minor	Final	Total
WACCM	0.6 (0.1)	1.6 (0.2)	0.3 (0.1)	2.6 (0.4)
WACCM-QBO	0.4 (0.1)	2.0 (0.3)	0.5 (0.1)	2.9 (0.4)
WACCM-SST	0.6 (0.1)	1.8 (0.3)	0.4 (0.1)	2.8 (0.4)

found that the latter run is far off from realistic polar stratospheric variability. They conclude that at least one of the forcings (QBO, SSTs) has to be incorporated in the WACCM model simulation to obtain observed frequencies of sudden stratospheric warmings.

In order to classify stratospheric warmings with the neural network framework, the external forcings (QBO, ENSO, SFL) are solely taken from the REF-B1 WACCM simulation. The internal variability factors ($\overline{T^{30}}$, PC1) are computed from each individual simulation. This will indicate, e.g., how important the QBO time series still is for vortex variability even though a QBO is not simulated (as in WACCM-QBO). Similar to the REF-B1 classification results, the neural network is able to classify more than 95% of all stratospheric warming events correctly as measured with the zonal wind criterion (compare Fig. 6.6).

The relative event numbers per year are shown in Table 6.8 for major, minor, and final warmings and the different CCM simulations. WACCM and WACCM-SST show a realistic frequency of major warmings with 0.6/yr. The frequency of WACCM-QBO is significantly smaller with 0.4/yr. Richter et al. [2011] find the opposite and observe that WACCM-QBO is more realistic than WACCM-SST. They have analyzed a 30 year period. The differences are caused by sampling on different periods. The minor warming frequencies of the artificial runs (approx. 1.9/yr) are slightly higher than that of WACCM. They average to approx. 1.8 minor warmings per year. The frequencies of final warmings of the artificial runs are also slightly higher than that of WACCM and are actually more in agreement with the final warming frequency in ERA and NCEP (0.5/yr). The total number of warmings is statistically indistinguishable and averages to approx. 2.8 warmings per year.

The corresponding monthly distributions of major, minor, and final warming events for the different simulation runs are shown in Fig. 6.11. The major warming distributions are quite similar across the different simulations. Only the Februar and the November frequencies in WACCM-QBO are significantly lower than those of WACCM and WACCM-SST. The frequencies of final warmings are slightly higher in the artifical runs, as already observed in Table 6.11. However, they maximize in April in agreement with WACCM. Larger differences are observed for minor warmings. WACCM produces realistic minor warming distributions whereas WACCM-QBO simulates much more minor warmings in mid-winter and WACCM-SST simulates more minor warmings in early winter than WACCM. From the overall frequencies (see Table 6.8) and the monthly distributions (see Fig. 6.11) it can be concluded that including the QBO is more important than including SST variability to

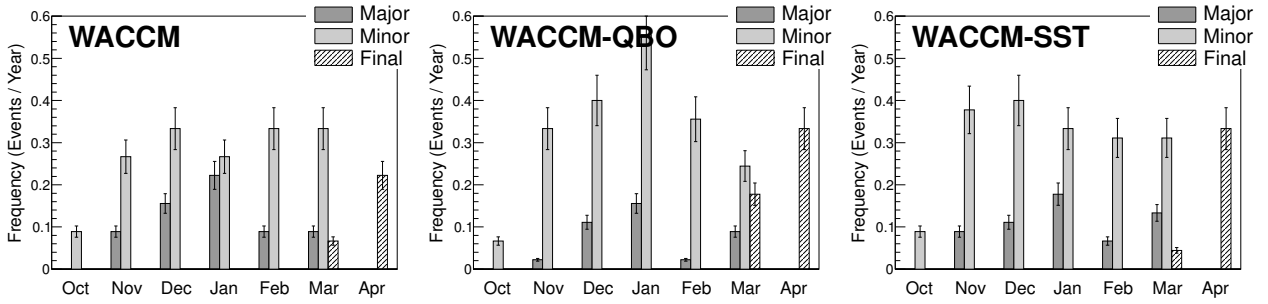

Figure 6.11: Monthly distributions in events/year of the three warming classes for WACCM (left), WACCM-QBO (middle), and WACCM-SST (right). The error bars represent the standard error of mean.

Table 6.9: Relative impact in percent of the input factors on the MLP response for major warmings according to Eq. 6.9 and the different WACCM sensitivity runs.

Input	WACCM	WACCM-QBO	WACCM-SST
$\overline{T^{30}}$	24.1	24.2	21.2
PC1	28.4	27.2	29.2
QBO	19.4	10.1	24.4
ENSO	15.3	22.4	13.7
SFL	12.8	16.0	11.4

generate realistic SSW frequencies.

In order to quantify the statistical importance of these factors, the impact according to Eq. 6.9 on the neural network response for major warmings is shown in Table 6.9, omitting minor and final warmings for simplicity. As expected, the QBO impact (10.1%) for WACCM-QBO is now much lower than that for WACCM, making ENSO and SFL relatively more important. The opposite is observed for WACCM-SST, where the ENSO impact is now smaller than that for WACCM, making the QBO impact higher than that for WACCM. SFL impact for WACCM-SST is approx. the same as for WACCM. The QBO impact on WACCM-QBO and the ENSO impact on WACCM-SST are small but far from negligible. This means that the neural network is still assigning statistical importance to these factors for discriminating between major and minor warmings, even though they are not part of the corresponding simulation runs. This is surprising and could be caused by the short data record and the relatively low number of major stratospheric warmings on this record. Also, a possible overfitting during the training of the neural network can not be ruled out.

6.10 Concluding Remarks

This study classifies stratospheric warmings by considering Arctic stratospheric temperature anomalies together with atmospheric forcings (or external factors) which influence the polar vortex, namely the QBO, ENSO, and the solar cycle (SFL). The classification procedure is applied to data from the

ERA-40/ERA-Interim (ERA) reanalyses and the NCEP/NCAR (NCEP) reanalysis for 52 consecutive winters from 1958 to 2010. Optimal lags of the external factors are determined using linear discriminant analysis.

Three supervised learning approaches (LDA, LSVM, MLP) are introduced and compared with respect to their ability to classify stratospheric warmings. It is shown that the nonlinear MLP outperforms the linear methods (Table 6.1). This is in agreement with previous work showing that the external factors nonlinearly influence the polar vortex evolution [e.g., Garfinkel and Hartmann, 2007; Calvo et al., 2009; Richter et al., 2011]. The MLP is therefore used as the method of choice to classify stratospheric warmings in major, minor, and major final warming events. This approach extends and combines the zonal wind measure and the NAM approach applied in previous studies. It incorporates the polar cap temperature and significant external factors simultaneously leading to a continuous probability measure, indicating the amount of deviation from the climatological mean state.

It is shown how an appropriate training sample (Fig. 6.2) can be calculated. Using this training sample, the optimal MLP architecture is determined using methods from information theory (Fig. 6.3). Using various performance measures, the classification procedure is successfully validated (Table 6.2). It is shown how resulting stratospheric warming probabilities (Fig. 6.4) are post-processed.

The statistical impact of the input factors on the individual output classes is computed (Table 6.3). It is shown that the atmospheric variability factors are an essential part of the classification procedure as they discriminate between minor and major stratospheric warmings. They are less important for final warmings and show only a small impact on the undisturbed state (Table 6.3). Inspite of the absence of any linear correlations between the external factors, there are nonlinear combinations that help distinguish between warming classes. The QBO was found to have the largest impact, followed by ENSO and the solar cycle. This ranking was also found by previous work [e.g., Labitzke and Kunze, 2009b; Camp and Tung, 2007a,b; Mitchell et al., 2011] that investigated the influence of these forcings on the polar vortex.

It is shown that detected major and final warming events lead to a vortex breakdown and a reversal of the zonal flow at 60° N (Fig. 6.5) except for a few cases (two in ERA, three in NCEP). Reasonable distributions of stratospheric warming events by month and year of occurrence and intensity are presented (Fig 6.6 and Fig 6.7), which are in agreement with previous work made by Charlton and Polvani [2007] and Labitzke and Naujokat [2000] who also compiled climatologies of stratospheric warming events. On average, major warmings show intensities that are twice as large as those of minor or final warmings. Final warmings last as long as major warmings but twice as long as minor warmings. There are largely positive, significant correlations greater than 0.6 between intensity and duration of the warming events (Table 6.5).

Marginalized probability distributions depending on QBO and ENSO, for both solar maximum and solar minimum conditions, are presented (Fig. 6.8). The results contain the linear QBO-SFL relationships presented by Camp and Tung [2007a] and Labitzke and Kunze [2009b]. However, it is

shown that the interrelationships between the external factors are nonlinear as previously suggested. QBO-SFL relationships are nonlinearly modulated by ENSO [Calvo et al., 2009]. It appears that El Niño-like conditions [Camp and Tung, 2007b] during QBO west favor the occurrence of major warmings and vice versa during QBO east. This pattern is more prominent for solar maximum than for solar minimum. For solar minimum, El Niño-like conditions and QBO east also point to large major warming probabilities. It is found that major warmings are generally more likely during solar minimum conditions. For solar minimum, there are only two regions that do not favor major warmings, which are small but positive ENSO values during QBO east and small but negative ENSO values during QBO west. This pattern also depends on the strength of the particular QBO phase. As discovered by Butler and Polvani [2011], major warmings are more likely during El Niño-like and La Niña-like conditions as opposed to neutral ENSO conditions. An exception to this is only observed for El Niño-like conditions during solar maximum and QBO east. In addition, it is shown that the extraordinary major warming of the winter 08/09 lies close to the significant climatological area that indicates a possible vortex breakdown. Therefore, this event is part of the nonlinear rules learned by the MLP. A three dimensional animation through the winter of the probabilities indicated in Fig. 6.8 can be found under `http://nathan.gfz-potsdam.de/doc/sswanim.gif`[2].

Using the current neural network framework, stratospheric warming events in REF-B1 chemistry climate model simulations could be succesfully classified. It was shown how major warming frequencies (Table 6.6) and corresponding monthly distributions (Fig. 6.9) in EMAC, MRI, and WACCM agree well with results presented in chapter 4 of SPARC CCMVal [2010]. A major warming frequency of 0.8/yr was found in EMAC, 0.4/yr in MRI, and the more realistic frequency of 0.6/yr in WACCM. It was found that the monthly WACCM distributions for major, minor, and final warmings agree best with warming distributions computed from the reanalyses. Monthly major warming distributions in EMAC and MRI simulations are too low during mid-winter. However, their minor warming distributions appear to be realistic. Concerning the importance of the external factors, a similar ranking as for the reanalyses was found for EMAC and WACCM assining most impact to the QBO, followed by ENSO, and the solar cycle (Table 6.7). For MRI, the ENSO impact is largest, followed by the solar cycle, and the QBO. Using these external factors, probability distributions for the major warming state were calculated (Fig. 6.10). Even though some important relationships between the external factors and vortex variability could be recovered, the estimation of a robust proabability pattern was not successful. This indicates significantly different nonlinear interrationships between the external factors across the CCMs. However, it is promising that the WACCM pattern for solar minimum is similar to the corresponding ERA pattern for the major warming state (see Fig. 6.8).

In order to measure the sensitivity of the external factors to the frequencies of stratospheric warmings and to the neural network classification, two additional artificially forced WACCM simulations were analyzed. These simulations were equally forced as the WACCM REF-B1 run except that 1) The QBO is not nudged resulting in a constant easterly phase (WACCM-QBO), and 2) Climatological SSTs are prescribed resulting in neutral ENSO phase conditions (WACCM-SST). The

[2]Alternatively: `http://wekuw.met.fu-berlin.de/~ChristianBlume/doc/sswanim.gif`

warming frequencies found in these simulations are quite similar to those estimated from WACCM (Table 6.8). Largest differences are found for WACCM-QBO which led to less major and more final warmings on the 45 year period. From the frequencies and the monthly distributions (Fig. 6.11), it was seen that including the QBO is more important than including SST variability to generate realistic SSW frequencies. The impact of the input factors on the neural network response for major warmings was computed (Table 6.9). Even though the QBO impact for WACCM-QBO and the ENSO impact for WACCM-SST were reduced compared to WACCM, they are not negligible which could be caused by a the short data record and a possible overfitting of the neural network. To rule out the latter point, cross validation should be applied in a future analysis and compared to results obtained with information criteria.

Several further improvements of the current statistical framework are possible. The introduction of the geopotential height into the MLP input layer would further enhance the classification results, as it provides direct information about the polar vortex strength. Introducing a memory of one or two days would also improve the classification but, at the same time, would exponentially increase computation time. Incorporating volcanic influences may also improve the classification procedure. It is shown that a statistical model with the current set of input factors needs to recognize nonlinear patterns to reliably classify stratospheric warmings. However, there are not only neural networks that can cope with this challenge. One may also think of applying e.g., support vector machines with nonlinear kernels or nonlinear functional discriminant analysis.

Chapter 7

Forecasting Polar Stratospheric Variability

It was shown in the previous chapter that stratospheric warming events, as the main manifestation of polar stratospheric variability, can be reliably classified if a nonlinear statistical method is used for pattern recognition while incorporating significant external factors. The study in the present chapter[1] investigates statistical models with respect to their ability to model and forecast polar stratospheric variability. Four partly nonstationary, nonlinear models are assessed, previously introduced in Chapter 3: linear discriminant analysis (LDA); a cluster method based on finite elements (FEM-VARX); a neural network, namely a multi-layer perceptron (MLP); and support vector regression (SVR). These methods model time series by incorporating all significant external factors simultaneously, including ENSO, QBO, the solar cycle, volcanoes, to then quantify their statistical importance.

When making polar stratospheric forecasts, general circulation model runs consisting of multiple observation constrained ensemble members are performed. These forecasts are reliable on a daily scale but on a seasonal scale they quickly become computationally very expensive and loose their forecast skill [Gerber et al., 2009; Kuroda, 2010; Dörnbrack et al., 2012]. This study investigates statistical models that are mathematically much simpler and demand significantly less computer power, and even though they do not simulate physical processes explicitly, one can learn about underlying relationships. A wide class of partly nonstationary and nonlinear statistical models are considered with respect to their ability to model and seasonally forecast geopotential and temperature anomalies representing variability in the polar middle stratosphere from 10 hPa to 30 hPa. In contrast, common statistical methods analyzing polar stratospheric variability are linear [e.g., Camp and Tung, 2007b; Crooks and Gray, 2005] and do not consider more than a few atmospheric forcing factors at the same time.

The statistical models are trained for the time period from 07/01/1980 through 06/31/2005 where training denotes the process of minimizing a method specific cost function that quantifies the deviation from the truth. Once trained, the statistical methods are used for hindcasting the period from 07/1/2005 to 04/30/2011. The sensitivity of the statistical models is tested with respect to small changes in the external factors. While making reasonable assumptions about the external

[1] This chapter, except for Section 7.4, was published with a few minor changes in Blume and Matthes [2012].

factors, the statistical model with the best generalization performance is used to forecast the winter 2011/12. The forecast performance is validated with actual observations from that winter.

7.1 Data and External Factors

This work makes use of two reanalysis data sets resolving the stratosphere, available up to 0.1 hPa from 1979 to present. These data sets are the ERA-Interim (hereafter ERA) reanalysis [Simmons et al., 2006] and the MERRA reanalysis [Rienecker et al., 2011], both considered from 7/1/1980 through 6/31/2011. ERA will be used to train the statistical models on the period from 7/1/1980 through 6/31/2005 and to test the models in a hindcast experiment from 7/1/2005 through 6/31/2011. MERRA is used to validate the final results but not further utilized. Please refer to Section 4.2 for a more detailed description of these data sets.

Two daily target geopotential and temperature time series are computed so as to represent the variability in the polar middle stratosphere. Anomalies of the area-weighted average on the polar cap (60° N–90° N) are computed at 10, 20, and 30 hPa. A subsequent principal component analysis [Jolliffe, 2002] of the three time series reveals that the first principal component (P1) explains more than 90 % of the overall variance in both ERA and MERRA. Therefore, only P1 was retained for both geopotential (P1Z) and temperature (P1T). P1Z and P1T are both positive for weak and warm vortex events and negative for strong and cold vortex conditions. P1T was recently used in Blume et al. [2012] to classify sudden stratospheric warmings events while incorporating important external forcings.

A polar cap average of geopotential anomalies is equivalent to the Northern Annual Mode (NAM) [Baldwin and Dunkerton, 2001], only reversed in sign. The NAM is a popular scalar index to measure polar stratospheric variability [e.g., Thompson, 2003; Baldwin and Thompson, 2009]. The NAM is the leading principal component of geopotential anomalies north of 20° N. Tthe polar cap method is, however, simpler and the resulting time series for geopotential and temperature are positively correlated ($R = 0.8$) [Baldwin and Thompson, 2009], pointing in the same direction during extreme vortex events. P1Z and P1T are physically closely correlated. A lead-lag correlation analysis between P1Z and P1T reveals that there is a correlation of approx. 0.7 when lagging P1T with 10 days whereas lagging P1Z leads to only 0.3. For instance during a sudden stratospheric warming, the temperature anomaly usually appears first and the actual vortex breakdown a few days later. In addition, P1T reflects the strong stratospheric cooling (overturning) proceeding most major warmings. For simplicity, P1Z is referred to hereafter as *geopotential* and P1T as *temperature*.

This analysis makes use of nine physical *external factors* which describe large-scale phenomena important for the polar stratosphere. Their purpose is to improve model variability and to obtain insight into relationships and impacts of the various forcings. The factors representing variability in sea surface temperatures (SSTs) are the El Niño- Southern Oscillation (ENSO) [Trenberth, 1997], the Pacific Decadal Oscillation (PDO) [MacDonald and Case, 2005], and the Atlantic Multidecadal Oscillation (AMO) [Delworth and Mann, 2000]. Deser et al. [2010] reviews variabilities in sea surface

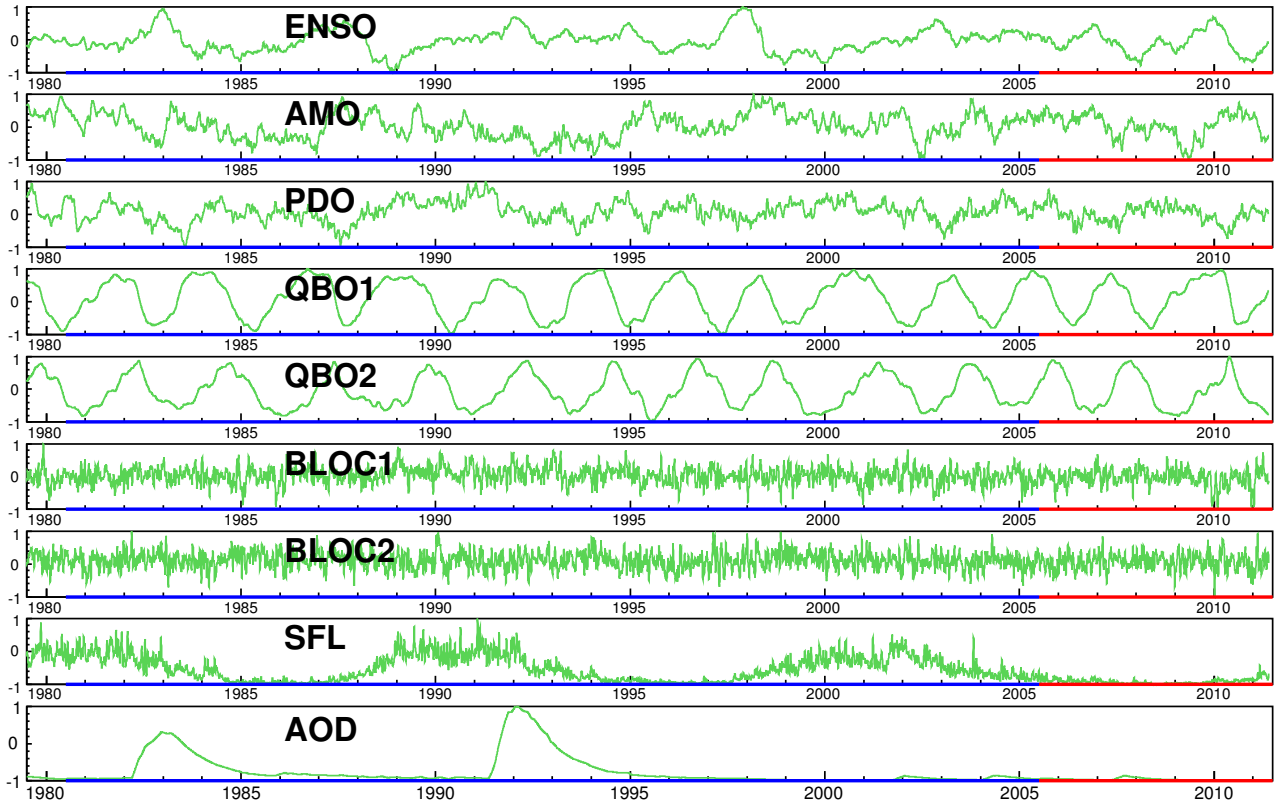

Figure 7.1: The different external factors in green as used in this study, omitting sine, cosine, and trend terms for simplicity. Blue denotes the training and red the hindcast period.

temperature and describes how corresponding indices can be computed. In this study, they have been calculated with an EOF analysis of detrended SST anomalies from 60° S to 60° N with ENSO being the leading EOF (see Section 2.2.3).

Furthermore, the first two principal components of equatorial stratospheric zonal wind anomalies (QBO1 and QBO2) [Wallace et al., 1993] are included. Factors representing tropospheric high-latitude blockings [Martius et al., 2009; Woollings and Hoskins, 2008] are the first two principal components of geopotential anomalies between 35° N and 85° N at 500 hPa (BLOC1 and BLOC2). BLOC1, as the leading principal component, is equivalent to the NAM in 500 hPa and represents blockings in both the Atlantic and Pacific sectors simultaneously. Moreover, the F10.7 cm radio flux representing solar variability (SFL; `ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA`) and the aerosol optical depth (AOD; `http://data.giss.nasa.gov/modelforce/strataer`) representing volcanic eruptions are included. The external factors as used in this study are described in Section 2.2.

Additionally, three baseline factors representing the seasonal cycle (one sine and one cosine with a period of one year) and a linear trend term are included. Since the different external factors have different magnitudes and physical units, they are normalized on the full period from 1980 to 2011, such that the minimum is at -1 and the maximum at $+1$. To reduce short-term fluctuations and extreme values, the daily external factors are low-pass filtered using a 5-day running mean.

In order to obtain an idea on how the different factors vary with time, their time series are visualized in Fig. 7.1. The time series of the sine, the cosine, and the trend term were omitted for

simplicity. It is noted that the factors vary on very different timescales from days (e.g. BLOC1) to decades (e.g. AMO). Others vary on both scales such as SFL and factors such as AOD represent only singular events.

In order to improve regression results, optimal time lags of each of the nine physical external factors were calculated using a lead-lag correlation analysis. Every external factor was correlated with geopotential (temperature) for different lags from 0 to 365 days. The largest statistically significant correlation on this period indicates the optimal lag. Zero lags are obtained for PDO, BLOC1, BLOC2, and AOD for both geopotential and temperature. For ENSO, 96 (82) days; for AMO 8 (185) days; for QBO1 173 (137) days; for QBO2 0 (264) days; and for SFL 0 (50) days are computed for geopotential and temperature, respectively.

7.2 Optimal Model Architecture

All models but LDA depend on a set of free tuning parameters that needs to be determined which is called the model architecture. The optimal model architecture (combination of tuning parameters) aims at meeting the principle of Occam's Razor [Ariew, 1976] stating that the simplest model is the preferred if it contains just as much information as any of the more complicated models. There are two major branches found in the literature of information theory [Burnham and Anderson, 2002] aiming at selecting the optimal model which are information criteria and cross-validation. The approach to be used depends on the statistical method and the specific application. Please refer to Section 3.7 for an overview of model selection.

For FEM-VARX, the optimal architecture was determined with the use of the Akaike information criterion [Akaike, 1974; Horenko, 2011] where the parameter setting leading to the smallest criterion is preferred. The Bayesian information criterion [Schwarz, 1978] applied to FEM-VARX led to a far too simple model ($K = 2$) with only little explanatory power on the training period and was therefore rejected. For MLP and SVR, a 5-fold cross-validation [Kohavi, 1995] was conducted in which the training data were partitioned into 5 equally-sized contiguous subsets (folds). The model architecture with the largest correlation calculated from the tested subsets has been selected. It is only possible to use cross-validation for SVR model selection. Cross-validation was chosen for MLP because this study aims at forecasting variability.

In the following, the optimal values are given in parentheses, for geopotential and temperature, respectively. For FEM-VARX, K (5, 5) denotes the number of clusters and C (146, 112) the persistency threshold. For MLP, L_1 (8, 3) and L_2 (5, 0) denote the number of neurons in the first and second hidden layer, respectively. For SVR, γ (1, 0.2) denotes the radial scaling parameter and C (0.3, 0.3) the trade-off parameter. Please refer to Chapter 3 for a description of the statistical methods and their parameters.

LDA and SVR only lead to global solutions, whereas FEM-VARX and MLP might run into local minima during training. In order to reduce this effect, a total of 30 models were trained for FEM-VARX and MLP. For FEM-VARX, K is fixed and values of C were chosen slightly different from the

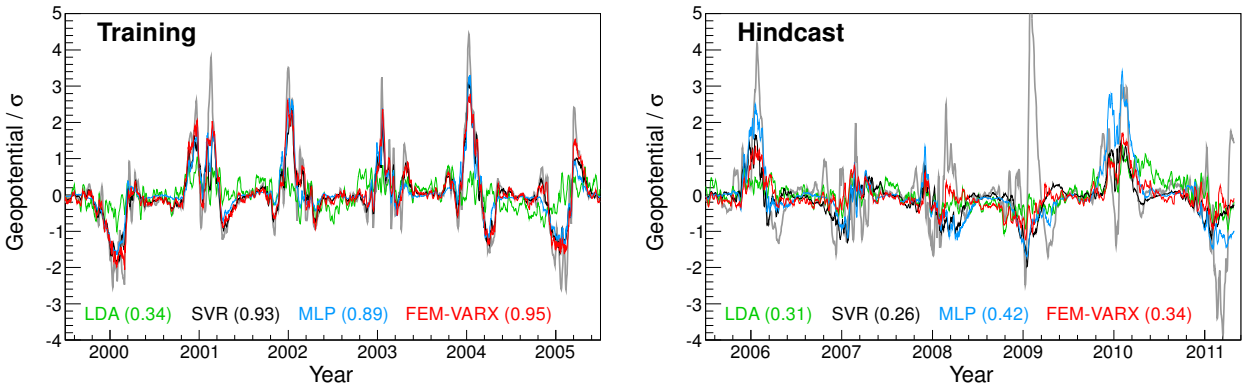

Figure 7.2: Geopotential results for the training period (left) and the hindcast period (right) for each of the statistical models together with the correlation coefficient R (in parentheses) calculated between the particular model and truth (gray). The hindcast is shown for the full hindcast period whereas the training results are only shown for a representative period (last six years). Labeled is the first of January of the particular year. The 95 % confidence interval of the correlation factors are ± 0.02 for the training and ± 0.04 for the hindcast period, computed with bootstrap based on case resampling (see Section 3.8).

optimal value. For MLP, those pairs of L_1 and L_2 were chosen that were ranked highest according to the cross-validation. The regression and forecast results are the average across these realizations.

7.3 Training and Hindcast Period

The statistical models are trained with data from the training period (1980–2005) while being set up with optimal model architectures as described above. After being trained, the models are used to hindcast the period from 2005 through 2011, meaning that the models are evaluated with the available external factors from this period. The result of this procedure is presented for geopotential in Fig. 7.2, where the truth is shown in gray. The correlation coefficient between each model and truth is given in parentheses. The training period is modeled well ($R \approx 0.9$) by all models except LDA ($R \approx 0.3$). FEM-VARX possesses the highest explanatory power over the training period. Please note that all external factors are used in a resolution of a few days. For the hindcast period a large drop in correlation is observed for all methods, largest for SVR. Only FEM-VARX ($R = 0.34$) and MLP ($R = 0.42$) lead to satisfactory results on the hindcast period. They are able to approximate most anomalies and hindcast the general behavior in 5 out of 6 winters. Looking more closely, significant differences between truth and hindcast become evident. The most obvious is the sudden stratospheric warming in January 2009. This is an extraordinary strong warming [Labitzke and Kunze, 2009a] during solar minimum and QBO west which was not expected according to the Solar-QBO relationships by Labitzke and Kunze [2009b] and Camp and Tung [2007a]. This is an example of variability that cannot be explained using the current statistical models reflecting the chaotic nature of the system. However, the present set of external factors might not be optimal and needs further investigation.

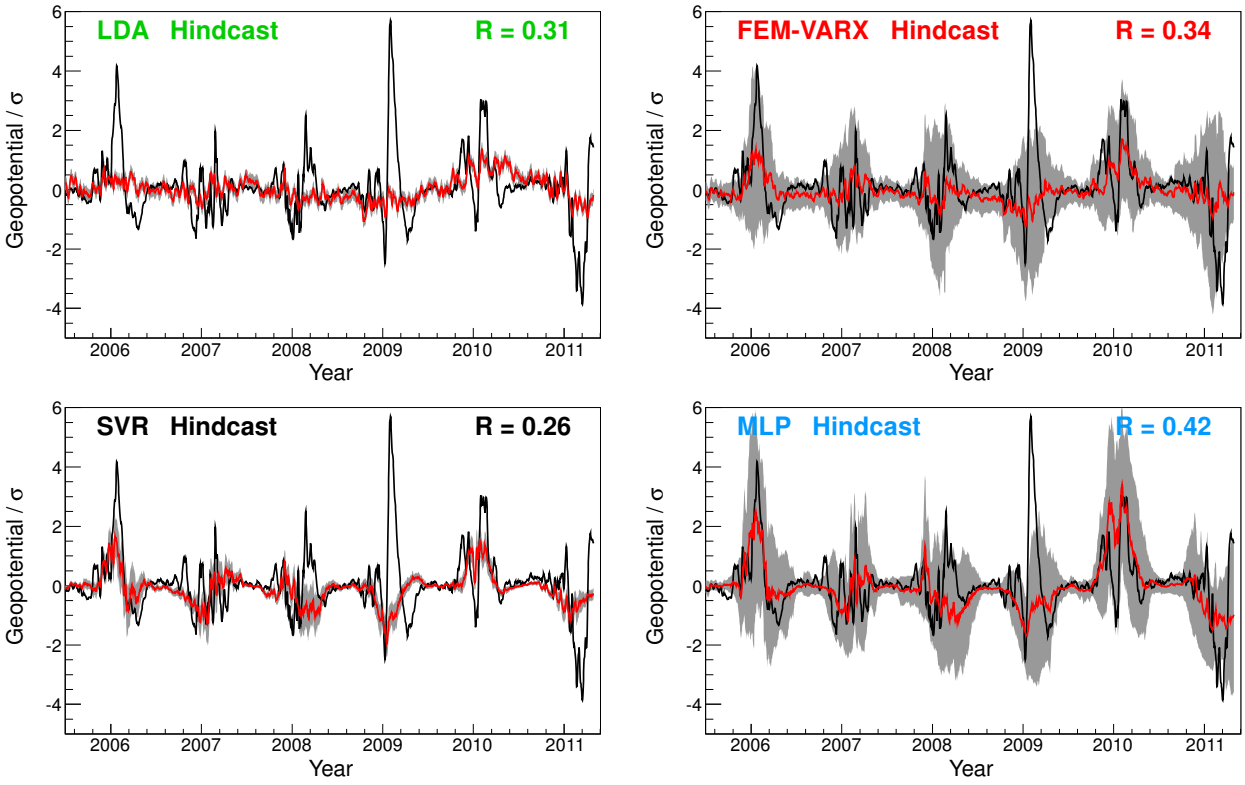

Figure 7.3: Individual hindcast results of the different statistical models (red) along with the truth (black). Also shown is the correlation coefficient R between model and truth. The shaded area (gray) denotes a 95% confidence interval computed with Gaussian process sampling as described in Section 3.8. For FEM-VARX and MLP, the bootstrap sampling was applied to each of the 30 realizations which were subsequently combined using the model averaging procedure presented in Section 3.7.3.

The temperature results (not shown) are similar to the geopotential results with slightly smaller correlations. For the training period, correlations of 0.25 (LDA), 0.53 (SVR), 0.86 (MLP), and 0.92 (FEM-VARX) are obtained. For the hindcast period, correlations of 0.22 (LDA), 0.16 (SVR), 0.25 (MLP), and 0.24 (FEM-VARX) are computed for temperature. These correlations, except for SVR, are statistically indistinguishable. However, the same ranking of the different models as for geopotential is still obtained.

It can be seen that it is possible to statistically model and hindcast polar stratospheric variability. The MERRA reanalysis is utilized to validate the regression results by evaluating each statistical model with MERRA data on both, the training and the hindcast period. This leads to very similar results and correlation coefficients compared to ERA (not shown). The ERA results can be considered as robust and trustworthy.

7.4 Model Sensitivity

This section aims at estimating the sensitivity of each model response with respect to small changes in the external factors. This is done on the hindcast period but is also representative for the training

period. The sensitivity is quantified using the bootstrap method based on Gaussian process sampling as described in Section 3.8. To summarize, in Gaussian process sampling every external factor is randomly altered according to a Gaussian distribution, independently for each time step. After having calculated 1000 randomly generated sets of external factors (bootstraps), the statistical model can be evaluated 1000 times. Finally, a confidence interval is calculated for the mean response from the different model responses. This can be somewhat compared to an operational weather forecast where not only a single run is computed but rather an ensemble of runs which makes it possible to assign an uncertainty to a forecast for each time step.

Since there is only a single model available for LDA and SVR, the boostrapping sampling was only applied once to these models. For FEM-VARX and MLP, however, the bootstrap sampling was computed for each of the 30 realizations which were subsequently combined using the model averaging procedure presented in Section 3.7.3. Therefore, the resulting uncertainty range for FEM-VARX and MLP is a combination of the differences between the 30 model realizations and the bootstrap sampling applied to each of these realizations.

The individual hindcast results of the different statistical models are presented in Fig. 7.3. The shading in gray denotes the 95% confidence interval computed with the bootstrap sampling. It is observed that LDA and SVR are robust against changes in the external factors. Whereas the LDA uncertainty is constant all throughout (because LDA is a linear method), the SVR uncertainties are slightly higher during winter than during summer, reflecting the larger winter variability in the polar stratosphere. This is also true for the uncertainties found in the FEM-VARX and MLP response. While their uncertainties are rather small during summer they become quite large during winter. The confidence intervals of FEM-VARX and MLP are by a factor of 7 larger than those of LDA and SVR. This is mainly caused by the 30 model realizations that were incorporated for FEM-VARX and MLP. In fact, the single model bootstrap sampling errors are of the same order as those found for SVR. It is interesting to note that the truth is mostly within the 95% confidence of the 30 realizations. Interestingly, only the sudden warming in January 2009 would even leave a $4\,\sigma$ confidence range during its maximum for both FEM-VARX and MLP, highlighting once more this extraordinary event. The use of independent model realizations is a powerful and straightforward way of estimating model uncertainties. It can be generally recommended when using statistical models that may run into local minima, such as FEM-VARX and MLP.

7.5 Impact of External Factors

The statistical importance (or *impact*) of each of the external factors on the statistical models is calculated (compare Section 3.9). The impact I_k is the standard deviation of the difference between model responses so that $I_k = \sigma(Y - Y^{(k)})$, where Y is the original model response and $Y^{(k)}$ is the model response for external factor k held constant at its median. I_k represents the averaged response deviation from the equilibrium response given by Y. The relative impact is then simply I_k divided by the sum of all impacts for one statistical model. This is shown in Fig. 7.4 for geopotential

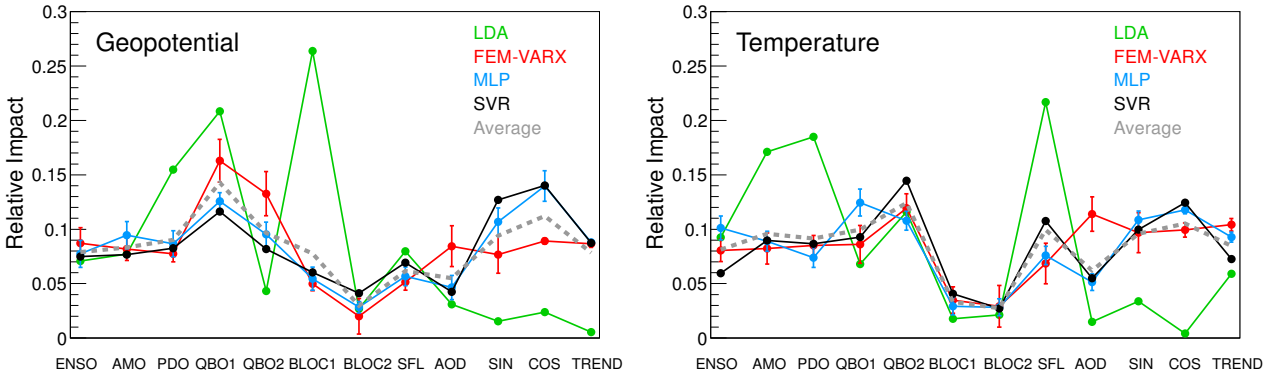

Figure 7.4: Relative impact of the external factors on each of the statistical models for geopotential (left) and temperature (right). The average impact (gray) is calculated as a weighted mean over the different models where the weights are calculated from the correlation coefficients (see left panel of Fig. 7.2) on the training period. The error bars of the FEM-VARX and MLP impacts denote the 95% confidence interval calculated from the 30 model realizations.

and temperature along with a weighted average over all four models. The weights were determined from the correlation coefficients of the training period, meaning that FEM-VARX is given the largest weight and LDA the smallest. It is observed that the impacts of FEM-VARX, MLP, and SVR are very similar, whereas LDA misinterprets the importance of factors, such as the impact of high-latitude blockings (BLOC1) on the geopotential or the solar cycle (SFL) on the temperature. LDA assumes linear and stationary relationships which is not valid for the polar stratosphere [e.g., Calvo et al., 2009; Richter et al., 2011].

Apart from LDA, the impacts in geopotential and temperature are very similar across the different models. A large impact of the QBO terms and a medium impact of SFL are observed in agreement with e.g., Holton and Tan [1982], Labitzke and Kunze [2009b] and Camp and Tung [2007a]. QBO1 is more important than QBO2 for geopotential and vice versa for temperature. The ENSO impact on vortex variability is moderate, as also found by Camp and Tung [2007b] and Mitchell et al. [2011]. The AMO and PDO impacts are of similar magnitude. There are only two sufficiently powerful volcanic eruptions (El Chichón in 1982 and Mt. Pinatubo in 1991) [Robock, 2000]. Therefore, the impact of the aerosol optical depth (AOD) index is very small for this period. It is worth noting that the AOD impacts vary significantly across the four models reflecting a large uncertainty for this forcing, possibly caused by the small number of eruptions important for the stratosphere. Surprisingly, the two factors representing tropospheric high-latitude blockings (BLOC1/2) show a relatively small importance. Especially for the modeling of temperature, they can be omitted. However, the BLOC1 impact on geopotential is of the same order as the SFL impact and needs to be accounted for. However, as stressed by Woollings and Hoskins [2008], BLOC1 represents high-latitude blockings in both the Atlantic and Pacific sectors simultaneously and cannot be used as a proxy for all blocking situations. Another challenge with blockings is that they are shown to precede SSWs but they also appear without an SSW following [Martius et al., 2009], making it difficult to use them for statistical modeling and therefore resulting in small statistical impacts. The sine and

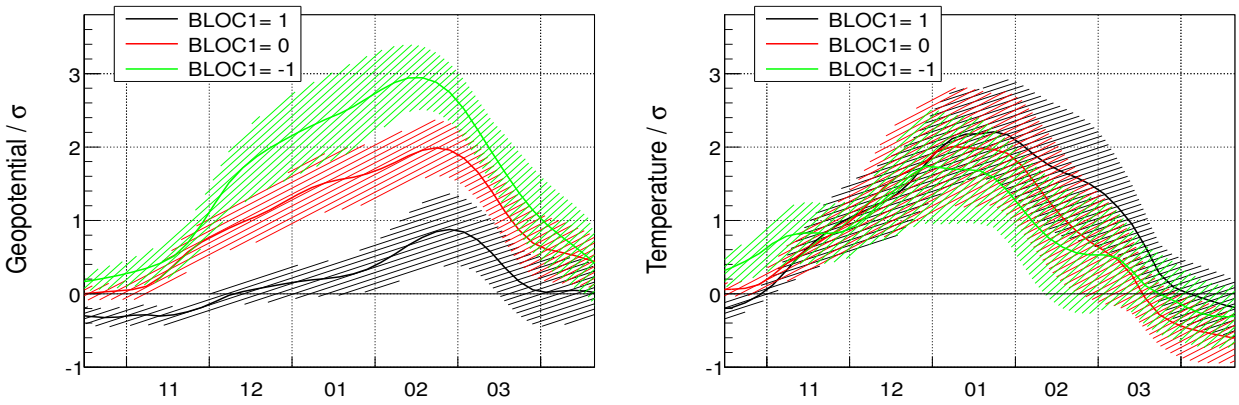

Figure 7.5: MLP forecast for the winter 2011/12, holding the external factors constant and varying only the sine and cosine terms. The assumptions about the external factors are partly received from predictions made by the NOAA and partly from scientific reasoning (see text). The forecast is shown for geopotential (left) and temperature (right) for three different conditions of BLOC1. The hatched area denotes the 95% confidence interval calculated from the 30 model realizations. Please note the additional uncertainty imposed by the only moderate hindcasting performance (see right panel of Fig. 7.2).

cosine terms largely influence the model response, which reflects the strong seasonal dependence of the dynamics in the polar stratosphere. The linear trend term was also found to be relatively strong ($\approx 10\%$).

7.6 The Winter 2011/12 - Forecast

MLP performs best over the hindcast period (see Fig. 7.2) and is therefore used to predict the winter 2011/12. As this winter lies in the future, there have to be made assumptions about the external factors while taking into account the optimal lags from Section 7.1. For SST variability along with SFL, predictions from the NOAA climate prediction center (`http://www.cpc.ncep.noaa.gov`) are used. This leads to -0.5 for ENSO, 0 for AMO, -0.3 for PDO, and -0.5 for SFL. Please note that the external factors are normalized between -1 and $+1$ for the period from 1980 to 2011. BLOC2 has a very small impact (see Fig. 7.4) which is why it is set to zero. AOD is held at -1, as future volcanic eruptions that might affect the stratosphere are unknown. The trend term is held at one (its value in 2011), as an approximate value for the extension of only one winter. By extending the QBO oscillation with a period of 28 months, 0.8 is obtained for QBO1 and 0 for QBO2.

Figure 7.5 shows the resulting MLP forecast for the winter 2011/12 by only varying the sine and cosine terms, for geopotential and temperature and for three different conditions of BLOC1. A value of -1 represents extremely pronounced high-latitude blocking situations [Woollings and Hoskins, 2008], whereas $+1$ represents no high-latitude blockings at all. For moderate values of BLOC1, the synoptic situation remains unclear and regional blocking situations may still occur. It is shown in Fig. 7.5 that the geopotential forecast changes significantly with varying BLOC1. However,

for minimum and average BLOC1 conditions, the geopotential forecast is well above one standard deviation. This also holds for the temperature forecast, which is almost unaffected by BLOC1 changes, indicating the small statistical importance of BLOC1 on the temperature response (see Fig. 7.4). To summarize, both forecasts tend to be positive and well above one sigma, indicating extreme variability and a warm stratospheric winter with a weak stratospheric vortex. Since the anomalies in Fig. 7.5 are quite large denoting extreme conditions, a sudden stratospheric warming is likely to take place in late January, early February 2012. The temperature anomaly leads and is proceeded by the geopotential anomaly.

The winter 2011/12 will most probably coincide with a westerly QBO in 50 hPa and weak solar activity (NOAA). Hence, the finding contrasts the Solar-QBO relationship found by Labitzke and Kunze [2009b], which predicts a cold and undisturbed polar stratosphere under these conditions. Correspondingly, Camp and Tung [2007a] found that solar minimum conditions and a westerly QBO point to the least disturbed vortex state. Moreover, work made by e.g., Camp and Tung [2007b] and Mitchell et al. [2011] indicates that a warm and disturbed polar stratosphere is more likely to take place during warm ENSO phases (El Niño) than during cold ENSO phases (La Niña). This is also in contrast to the present forecast, since the ENSO index is most likely to be moderately negative for the winter 2011/12 according to the NOAA predictions. However, since the impacts of the individual external factors do not add up linearly, a nonlinear statistical method is certainly more appropriate. This analysis, in addition to being nonlinear, incorporates all the significant external factors simultaneously.

7.7 The Winter 2011/12 - Observations

After the statistical forecast was made, the winter 2011/12 has passed, making it now possible to directly compare the prediction of the polar vortex conditions with actual observations. The NCEP/NCAR reanalysis [Kalnay et al., 1996] is used as a reference as it provides a gridded, freely available data set until almost present day. This is shown in Fig. 7.6 for geopotential and temperature which represent polar cap anomalies and were computed as described in Section 7.1. Hence, they are comparable to the statistical forecast presented in Fig. 7.5. In order to measure directly if the vortex broke down, the zonal mean zonal wind at 60° N and 10 hPa is shown on the right panel of Fig. 7.6. During a vortex breakdown, i.e. a sudden stratospheric warming, the zonal wind is smaller than zero (easterlies).

It is observed in Fig. 7.6 that the polar vortex in early winter was stronger and colder than usual, indicated by the negative temperature anomalies in November and December 2011. Also the zonal wind is slightly stronger than the climatology during early winter with a maximum of 48 m/s indicating a strong polar vortex. Then, suddenly, the temperature rises by almost $3\,\sigma$ at the end of December within a few days. The geopotential follows approx. two weeks after and both reach maximum values of $2.5\,\sigma$ (geopotential) and $3.5\,\sigma$ (temperature) in mid January to then decrease to climatological values within approx. four weeks (geopotential) and two weeks (temperature). The

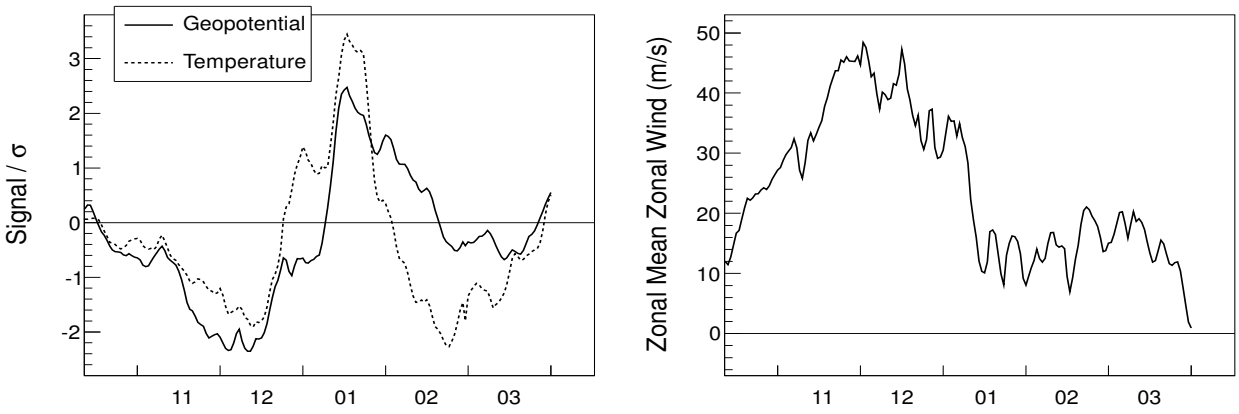

Figure 7.6: Observed polar stratospheric variability of the winter 2011/12 until March 31st for geopotential and temperature (left) and the zonal mean zonal wind at 60° N and 10 hPa (right).

zonal wind follows somewhat the progression of the geopotential and reaches a first minimum of 8 m/s in mid January. The zonal wind, however, does not drop below values of 6 m/s for the rest of the actual winter so that no vortex breakdown, i.e. no major stratospheric warming took place. At the end of March, the transition toward the stratospheric summer circulation is observed.

There was no vortex breakdown, but the vortex conditions of the winter 2011/12 were extreme. The vortex was anomalously weak from mid January to mid February and anomalously warm from end of December to end of January. This is a classical example of a minor warming during mid-winter [Labitzke and Naujokat, 2000]. Therefore, a winter 2011/12 with warm and weak stratospheric vortex conditions was predicted correctly in Section 7.6. Even the temperature maximum for mid January was predicted correctly as shown in Fig. 7.5. The forecast for the geopotential is already large in mid January (for BLOC1=-1,0) but does not peak until February. It is notable that the magnitudes of the forecast (except for BLOC1=1) are similar to those found in observations. In addition to correctly forecasting weak and warm vortex conditions in Section 7.6, a vortex breakdown for late January, early February was also predicted. This did not take place as mentioned earlier and was not forecasted correctly. However, the vortex remained the weakest from mid January to mid February, as observed in Fig. 7.6 for the geopotential and the zonal wind.

7.8 Concluding Remarks

A novel statistical treatment of variabilities in the polar middle stratosphere was presented, making use of four independent and different statistical models. For the first time, partly nonstationary and nonlinear statistical methods were trained with polar stratospheric geopotential and temperature anomalies incorporating all significant external factors simultaneously (Fig. 7.1). It was shown that, with the help of external factors, FEM-VARX and MLP are able to model and satisfactorily hindcast the variabilities (Fig. 7.2). With the help of several model realizations along with bootstrap sampling, reasonable confidence intervals could be calculated, enclosing most variability (Fig. 7.3). However,

a degree of internal chaotic variability remains, seen in the sudden stratospheric warming of January 2009, that cannot be forecasted using the current statistical models and the current set of external factors.

The statistical impact of each of the external factors on the statistical models was computed (Fig. 7.4). It was shown that the QBO factors, the seasonal terms, and the trend term show the greatest impact. The solar cycle and the SST variabilities have a medium impact along with high-latitude large-scale blockings (BLOC1). Volcanic eruptions (AOD) only point to a small but more uncertain statistical importance. It was observed that relative impacts of external factors are very similar for FEM-VARX, MLP, and SVR, whereas those of LDA differ significantly from the model-averaged impact. Therefore, LDA should not be used for a study like this as it does not weight the external factors correctly.

Since the multi-layer perceptron (MLP) showed the best generalization performance it was used to predict the winter 2011/12 under reasonable assumptions about the external factors (Fig. 7.5). It predicts a disturbed and warm polar stratosphere, with a sudden stratospheric warming likely to take place during late January, early February 2012. This is in contrast to previous studies which expect a cold and less disturbed polar stratosphere given the same external factors (weak solar, QBO west, La Niña). However, standard analysis is based on linear models and does not consider more than a few external factors at the same time. The prediction is based on a nonlinear statistical method incorporating all significant external factors simultaneously.

After this statistical forecast was made, the winter 2011/12 has passed so that the statistical forecast could be compared to actual observations (Fig. 7.6). It was found that the prediction of warm and weak vortex conditions was correct. The prediction of a vortex breakdown was not. However, a strong minor warming was observed during mid-winter and the zonal flow slowed down and was the smallest from mid January to mid February. Hence, the forecast for the winter 2011/12 was correct to a large extent and shows the great potential in using nonlinear statistical models for the modeling and forecasting of polar stratospheric variability.

There are several improvements that could be made to this analysis. There may exist other, currently not included external factors that may improve the statistical forecasting of polar stratospheric variability. For instance, a different and possibly more regional blocking index should be tested within the current framework. In the current study, a set of factors was held constant and the optimal model architecture was computed for each statistical method. However, it would be favorable to optimize on the set of external factors plus the internal model setting. For each tested set of external factors, the optimal model setting would have to be estimated using information criteria or cross-validation. This is usually computationally expensive but may be feasible for a reasonable number of training events and external factors.

Instead of a linear lag correlation analysis, the lags should be computed separately for each statistical model, with a grid search technique using cross-validation. Unfortunately, these lag calculations would be computationally extremely expensive. It will also be interesting to decrease the temporal resolution of the considered time series to see if the modeling improves. Additionally,

the nonlinear interrelationships between external factors should be further investigated using the introduced methods.

Chapter 8

Conclusions and Outlook

This study focused at applying advanced, purely data-driven statistical approaches to the modeling and forecasting of stratospheric variability on different temporal and spatial domains. In contrast to simulations with chemistry-climate models, statistical methods are mathematically simpler, computationally less expensive, and weight external factors according to their importance. The statistical methods were trained with historical data from observations, reanalyses, and model simulations and were then used to forecast variability into the future. For the first time, four partly nonlinear, non-stationary statistical learning approaches were compared with respect to their ability to model and forecast stratospheric variability. In addition, the statistical importance of the external factors on each of the statistical methods was quantified. This work aims at answering three major questions raised in the introduction: 1) *What is gained by applying more complex statistical methods, as opposed to simple, linear methods?* 2) *Can statistical methods be used to forecast stratospheric behavior?* 3) *Which of the statistical approaches is best suited to tackle stratospheric problems?*

The stratosphere is governed by forcings and modes of variability that are shown to result in a complex, generally nonlinear and nonstationary response, particularly during winter when upward-propagating waves can interact with the mean flow (see Chapter 2). In this work and in contrast to previous studies, a wide class of statistical methods was applied that are potentially able to cope with these properties. The considered methods (see Chapter 3), are linear discriminant analysis (LDA), a cluster method based on finite elements (FEM-VARX), a neural network, namely the multi-layer perceptron (MLP), and the support vector machine (SVM). LDA is also referred to as multiple linear regression when applied to regression problems. LDA is a linear method which means that it is only able to derive linear relationships between external factors and its response. In contrast, MLP and SVM are able to infer functions that can nonlinearly map between external factors and response. LDA, MLP, and SVM are stationary methods meaning that their response does not depend on time, whereas FEM-VARX allows for non-stationary modeling due to a time-dependent jump process switching between persistent states called clusters.

The statistical methods used in this work, except LDA, depend on a set of tuning parameters referred to as the model architecture, which may be, e.g., the number of clusters in FEM-VARX or the number of hidden neurons in the MLP. For every application, an optimal model architecture

has to be derived for each statistical method. Once this optimal model is found, it is trained with historical data and can subsequently be used to recognize novel, previously unknown patterns. The latter is referred to as forecasting. There are two major branches of model selection (finding an optimal model architecture) applied in this study, information criteria and cross-validation. Both try to meet the principle of Occam's Razor which states that the simplest model is the preferred model if it contains just as much information as any of the more complicated models. Occam's Razor aims at finding the optimal trade-off between approximating the training data and generalizing to unseen data (forecasting). The model selection procedure to be used depends on the application and the statistical method itself (see Section 3.7). Instead of choosing only a single model, one may also select a set of models which are averaged to obtain better forecast results.

Aside from the advanced learning approaches, a number of mathematical tools were introduced in Chapter 3 and used all throughout this work, the most prominent being principal component analysis (PCA). PCA is utilized for reducing dimensionality in high dimensional fields, and to understand important variabilities in underlying data (see Section 3.1). The uncertainty of eigenvalues resulting from a PCA can be computed with bootstrap based on case resampling (see Section 3.8). Bootstrap is also used to compute uncertainties of correlation coefficients and statistical model responses. The difference between an altered and an equilibrium model response is used in this work to estimate the statistical importance (or impact) of external factors on modeled variabilities (see Section 3.9.2). For the simple case of LDA, this reduces to the regression coefficient times the standard deviation of the corresponding factor.

Due to their specific importance for indicating changes in dynamics and composition, stratospheric monthly temperature and ozone were statistically modeled (see Chapter 5). The data to be modeled were deseasoned and zonally averaged. The introduced statistical methods were trained with the help of external factors known to influence stratospheric processes, including variability in SSTs, the QBO, the annular modes (AM), the solar cycle (SFL), volcanic eruptions (AOD), and a trend term. The trend term was linear for temperature and Effective Stratospheric Chlorine ($\mathrm{ESC} = \mathrm{Chl} + 60 \cdot \mathrm{Br}$) for ozone (see Section 2.3). Variability was modeled in observations (RANDEL, NIWA), reanalyses (ERA, MERRA), and REF-B1 CCM simulations (EMAC, WACCM, MRI) (see Chapter 4). Since the dimensionality of the data to be modeled was large, i.e. global stratosphere in the zonal mean, PCA was used to efficiently reduce dimensionality, prior to the actual statistical modeling, while still retaining 90% of the overall variance (see Fig. 5.1, Fig. 5.2, and Table 5.1).

After estimating optimal model architectures (see Table 5.2), the regression performance of the different methods was assessed. It was found that SVR clearly outperformed the remaining methods. LDA, FEM-VARX, and MLP did not perform significantly different in terms of explained variance for both temperature and ozone (see Fig. 5.7, Fig. 5.8, and Table 5.3). An extensive analysis of the individual impacts of the different external factors on temperature and ozone variability was undertaken with respect to the different data sets. To summarize, the impacts on temperature (see Fig. 5.9) were rather similar across the data sets. The impacts for LDA were mostly close to the mean impact calculated across the different statistical models. The statistical models mostly agreed

and only showed large differences in the impacts for the linear trend term and the annular modes. For ozone (see Fig. 5.10), the ESC impact is more uncertain between the different statistical models, particularly for NIWA. The LDA impact for ozone was generally close to the mean impacts except for the annular modes and the seasonal influence (SI).

To finally quantify natural variability, the individual impacts were averaged across the statistical models and data sets to obtain a robust estimate including an uncertainty range (see Fig. 5.11). It was found that the most important factors for stratospheric temperature are the QBO (20%), SSTs (15%), and the two annular modes (15% each for NAM and SAM). The least important factors are SFL, AOD, and SI with each having around 7% impact. The impact of the trend term is of the same order as the SST impact. The impacts of SSTs, QBO, SFL, AOD, and SI on ozone are not significantly different, only showing larger uncertainties due to larger disagreements across the different ozone data sets. Only the impact of the annular modes (6% each) on ozone is significantly smaller and the ESC impact (29%) significantly larger. It was found that these impacts agree with previous work that investigated the impact of external factors on stratospheric variability [e.g. SPARC CCMVal, 2010, Chapter 8]. By adding up the different natural impacts, the overall natural variability can be quantified. For temperature, the natural factors combined explain $85\pm7\%$ of the total temperature variability, implying that there is at least 8% variability that is anthropogenically caused by the emissions of greenhouse gases and ozone depleting substances (see Section 2.3). For ozone, the natural factors are responsible for $71\pm10\%$ variability indicating that at least 19% is of anthropogenic origin due to the emissions of ozone depleting substances (see Section 2.3).

The regional importance depending on latitude and height of the external factors was also estimated (see Fig. 5.12). The typical well known structures could be recovered with QBO, SFL, and AOD being especially important for the tropics, whereas SSTs and the annular modes are more important for the extratropics. The solar impact is particularly large in the upper tropical stratosphere and the AOD impact in the lower tropical stratosphere. These impacts refer to the most prominent, direct influences of the external factors. As a final task, the statistical models were used to compute statistical forecasts up to the year 2100 under reasonable assumptions about the external factors. It was found that statistical models efficiently forecast long-term trends in temperature (see Fig. 5.13) and ozone (see Fig. 5.14). Even a simple method like LDA, performs just as good as the more complex statistical models. Overall, LDA leads to robust results for this problem and does not perform significantly worse than the other models. Even though SVR explains more variance, it does not lead to more robust impacts of the external factors. The wide application of LDA in the stratospheric community to regression problems with monthly data appears to be well justified. However, it should be noted that including all four methods in this analysis ensures that natural variability can be estimated including robust uncertainties. One challenge in this study is that the number of training events, i.e. the number of steps in time, is quite small. The training period from 1980 to 2005 consists of a total of only 312 months which is why FEM-VARX and MLP do not perform significantly better than the linear regression technique. This drastically changes when more training events become available.

A prominent example of dynamical wave-mean flow interactions during winter are sudden stratospheric warmings (SSWs). SSWs are dramatic extreme events characterized by a great temperature increase on daily time scales and a breakdown of the polar vortex. In Chapter 6 and published in Blume et al. [2012], statistical learning approaches were applied to the classification of SSWs in reanalysis data for 52 consecutive winters from 1958 to 2010. Three statistical methods were compared with respect to their ability to classify stratospheric warmings with the help of a polar cap temperature anomaly along with three important external factors (ENSO, QBO, SFL) known to influence the stratospheric vortex. The nonlinear MLP clearly outperformed (see Table 6.1) the linear methods (LDA, linear SVM) so it was further used to classify stratospheric warmings in major, minor, and final warming events. After deriving a training sample (see Fig. 6.2) and selecting the optimal model (see Fig. 6.3), the MLP was trained and evaluated with data from the ERA-40/ERA-interim and the NCEP/NCAR reanalysis. It was shown that the MLP nonlinearly combines the temperature anomaly along with ENSO, QBO, and SFL to classify stratospheric warmings correctly.

In agreement with previous work [e.g., Labitzke and Kunze, 2009b; Camp and Tung, 2007a,b; Mitchell et al., 2011], it was shown that vortex variability in reanalyses is mainly governed by the QBO, followed by ENSO and the solar cycle (see Table 6.3). In addition, nonlinear probability patterns for the major warming case were computed for solar minimum and maximum conditions (see Fig. 6.8). Important relationships between QBO, ENSO, and solar cycle could be recovered. In addition, it was shown how interrelationships are highly nonlinear, meaning that QBO-SFL relationships are nonlinearly modulated by ENSO. The SSW classification using a neural network was successfully applied to reanalyses data (see Table 6.4 and Fig. 6.5). An extensive compilation of major, minor, and final warming climatologies was compiled for monthly (see Fig. 6.6) and yearly (see Fig 6.7) distributions concerning frequency and intensity. It was shown that classification results are similar between ERA and NCEP data past 1979 but significant differences appear during the pre-satellite era before 1979.

The classification procedure using the nonlinear neural network was applied to the three CCM simulations EMAC, MRI, and WACCM (see Chapter 4.3) for the time of 1960 to 2005. It was shown how major warming frequencies (Table 6.6) and corresponding monthly distributions (Fig. 6.9) agree well with results presented in chapter 4 of SPARC CCMVal [2010]. Concerning the importance of the external factors, a very similar ranking as for the reanalyses was found for EMAC and WACCM assigning most impact to the QBO, followed by ENSO, and the solar cycle (Table 6.7). For MRI, the ENSO impact is largest, followed by the solar cycle, and the QBO. Using these external factors, probability distributions for the major warming state were calculated (Fig. 6.10). Even though some important relationships between the external factors and vortex variability could be recovered, the estimation of a robust probability pattern was not successful. This implies that interrelationships between QBO, ENSO, and SFL are significantly different among the CCMs.

In order to measure the sensitivity of the external factors to the frequencies of stratospheric warmings and to the neural network classification, two additional artificially forced WACCM simulations were analyzed. These simulations were equally forced as the WACCM REF-B1 run except that

1) The QBO is not nudged resulting in a constant easterly phase (WACCM-QBO), and 2) Climatological SSTs are prescribed resulting in neutral ENSO phase conditions (WACCM-SST). From the frequencies and the monthly distributions (Fig. 6.11) it was seen that including the QBO is more important than including SST variability to generate realistic SSW frequencies. The impact of the input factors on the neural network response for major warmings was computed (Table 6.9). Even though the QBO impact for WACCM-QBO and the ENSO impact for WACCM-SST were reduced compared to WACCM, they are not negligible which could be caused by the short data record (45 years) and a possible overfitting of the neural network.

Polar stratospheric variability is a prominent aspect of the middle atmosphere. The statistical modeling and forecasting of polar stratospheric variability in reanalysis data was performed in Chapter 7 and published in Blume and Matthes [2012]. The considered methods were LDA, FEM-VARX, MLP, and SVR which modeled geopotential and temperature anomalies representing polar variability of the middle stratosphere. It was shown that, with the help of external factors, FEM-VARX and MLP are able to model and satisfactorily hindcast the variabilities (Fig. 7.2). With the help of several model realizations, along with bootstrap sampling, reasonable confidence intervals could be calculated, enclosing most variability (Fig. 7.3). However, a degree of internal chaotic variability remains, seen in the sudden stratospheric warming of January 2009, which cannot be forecasted using the current statistical models and the current set of external factors. The statistical impact of each of the external factors on the statistical models was computed (Fig. 7.4). Relative impacts of external factors are similar for FEM-VARX, MLP, and SVR, whereas those of LDA differ significantly from the model-averaged impact. Therefore, LDA should not be used for a study like this as it does not weight the external factors correctly. Since the multi-layer perceptron (MLP) showed the best generalization performance, it was used to predict the winter 2011/12 under reasonable assumptions about the external factors (Fig. 7.5). It predicts a disturbed and warm polar stratosphere, with a sudden stratospheric warming likely to take place during late January, early February 2012. After the winter 2011/12 was observed (Fig. 7.6) it can be concluded that the prediction of warm and weak vortex conditions was correct. The prediction of a vortex breakdown was not. However, a strong minor warming was observed during mid-winter and the zonal flow slowed down and was at the smallest level from mid January to mid February. Hence, the forecast for the winter 2011/12 was correct to a large extent and shows the great potential in using nonlinear statistical models for the modeling and forecasting of polar stratospheric variability.

By applying statistical methods to the modeling of stratospheric variability, external forcings and internal modes are weighted according to their statistical importance. Doing a comparable study using a climate model is computationally not feasible. It was observed that the main drivers of stratospheric variability are the Quasi-Biennial Oscillation, variability in sea surface temperature (e.g. ENSO), the annular modes, and anthropogenic influence due to the emission of Greenhouse gases and ozone depleting substances. A generally smaller but still significant impact is caused by the 11-yr solar cycle and volcanic eruptions. It was shown that these impacts vary regionally, depending on height and latitude, and between different variables. For instance, The impact of the

solar cycle on polar stratospheric temperature is larger than that on the geopotential in this region. One assumptions being made by statistical methods is that the external factors (the features) are independent. However, factors such as the QBO and ENSO influence each other to a certain extent. Therefore, one needs to be careful when drawing conclusions from statistical impacts. This imposed uncertainty can be quantified by applying a range of statistical models to the same problem.

In this thesis, a standard linear (LDA) statistical method and three advanced nonlinear (MLP, SVM) and nonstationary (FEM-VARX) methods were applied. In Chapter 5, they were applied to model monthly temperature and ozone in the global stratosphere. The LDA regression performance was found to be similar to those of MLP and FEM-VARX, but SVR performed better. In terms of impact and forecast performance, LDA let to more robust results than the advanced models. Therefore, for the statistical modeling of stratospheric monthly data, the application of complex statistical models is not reasonable. This is caused by the small number of training events (monthly data). The application of the popular multiple linear regression analysis to the modeling of monthly stratospheric variability is therefore well justified. This changes when more training events become available. Chapter 6 classified stratospheric warmings using daily data and compared linear methods (LDA, LSVM) with a nonlinear method (MLP). The nonlinear neural network (MLP) performed significantly better than the linear methods for this application. Due to the much higher number of training events (factor of 30), the MLP was able to learn the nonlinear patterns and links that connect the external factors with stratospheric vortex variability. Due to the small number of external factors, it was possible to visualize the nonlinear patterns. Daily stratospheric vortex variability was also statistically modeled in Chapter 7, but with a regression approach and all significant external factors. The advanced models outperformed LDA on the training period, but only FEM-VARX and MLP were significantly better than LDA on the hindcast period. Even though the effect of overfitting is the smallest for LDA, making it a robust model, its forecast performance is very limited. In addition, the statistical impacts from LDA were significantly different from those computed with FEM-VARX, MLP, and SVR. The fact that FEM-VARX and MLP perform well reveals once more that the polar stratosphere is governed by processes that result in a non-stationary (FEM-VARX) and nonlinear (MLP) response. A method that can cope with both of these properties would be favorable. The three major questions stated in the beginning can now be answered.

What is gained by applying more complex statistical methods as opposed to simple, linear methods? If presented with monthly data, there is no significant information gain and a linear statistical method is sufficient and more robust. If there are many training events (e.g. daily data), than more complex (nonlinear, nonstationary) methods lead to more reasonable statistical impacts and better performances on the training and the validation period. However, more advanced methods come with a set of tuning parameters that need to be carefully selected using information criteria and cross-validation (model selection).

Can statistical methods be used to forecast stratospheric behavior? They can be used to forecast variability to a certain extent. It was found that long-term statistical forecasts agree well with those simulated by sophisticated CCMs. Also, statistical forecasts for the polar stratosphere on

the seasonal scale are very promising. However, there is internal variability that cannot be forecasted with the current methods and external factors.

Which of the statistical approaches is best suited to tackle stratospheric problems? It was shown that different statistical methods perform different for varying applications. It is always recommended to compare a range of statistical methods for a specific application. However, if it is known that relationships between external factors and response are linear than a standard linear method will suffice. If these relationships are known to be nonlinear and nonstationary, as the case for the polar stratosphere, than a nonlinear (e.g., MLP, SVM) or nonstationary (e.g., FEM-VARX) method should be applied. A sufficient amount of training events has to be available for the latter. From the applications in this thesis, the methods of FEM-VARX and MLP were most promising when dealing with daily data.

There is a range of possible improvements for future statistical modeling and forecasting of the kind presented in this thesis: 1) There may exist other, currently not included external factors that may improve the statistical modeling. In addition, other statistical models that were not considered in this thesis should be investigated. 2) In this study, a set of factors was held constant and the optimal model architecture was computed for each statistical method. However, it would be favorable, but computationally expensive, to optimize on both the set of external factors and the internal model setting by using information criteria or cross-validation. 3) Instead of a linear lag correlation analysis performed independently for each external factor, the lags should be computed separately for each statistical model across all factors with a grid search technique using cross-validation. Unfortunately, these lag calculations would be computationally expensive. 4) It would be interesting to see if other temporal resolutions (such as weekly) would improve modeling and forecasting.

There are three major branches to which the introduced statistical models can contribute, considering stratospheric variability: Quantifying impacts of external factors, understanding nonlinear interrelationships between external factors, and forecasting on seasonal, interannual, and decadal scales. It has been shown in this thesis that there is great potential for statistical models to cope with these challenges. They should not be thought of as to replace GCM or CCM simulations but rather to complement them. For instance, the importance of external forcings quantified with CCM sensitivity simulations can be validated and confirmed by applying statistical models. Moreover, an operational forecast for the polar stratosphere during winter, as made by the ECMWF, could be enhanced by performing statistical forecasts using nonlinear, nonstationary statistical methods that incorporate external factors known to influence the stratosphere.

Bibliography

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19, 716–723.

Andrews, D. G., J. R. Holton, and C. B. Leovy, 1987: *Middle Atmosphere Dynamics*. Academic Press, 489 pp.

Angell, J. K., 1997: Stratospheric warming due to Agung, El Chichón, and Pinatubo taking into account the Quasi-Biennial Oscillation. *J. Geophys. Res.*, 102 (D8), 9479–9485.

Ariew, R., 1976: Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony. Ph.D. thesis, Champaign-Urbana, University of Illinois.

Avriel, M., 2003: *Nonlinear Programming: Analysis and Methods*. Dover, 528 pp.

Baldwin, M. P. and T. J. Dunkerton, 1989: The stratospheric major warming of early december 1987. *J. Atmos. Sci.*, 46 (18), 2863–2884.

Baldwin, M. P. and T. J. Dunkerton, 2001: Stratospheric harbingers of anomalous weather regimes. *Science*, 294, 581–584.

Baldwin, M. P. and J. R. Holton, 1988: Climatology of the stratospheric polar vortex and planetary wave breaking. *J. Atmos. Sci.*, 45 (7), 1123–1142.

Baldwin, M. P. and D. W. J. Thompson, 2009: A critical comparison of stratosphere-troposphere coupling indices. *Q. J. Roy. Meteor. Soc.*, 135, 1661–1672.

Baldwin, M. P., et al., 2001: The Quasi-Biennial Oscillation. *Rev. Geophys.*, 39 (2), 179–229.

Ben-Hur, A., D. Horn, H. T. Siegelmann, and V. Vapnik, 2001: Support vector clustering. *J. Mach. Learn. Res.*, 2, 125–137.

Bishop, C., 1995: *Neural networks for pattern recognition*. Oxford: University Press, 482 pp.

Blume, C. and K. Matthes, 2012: Understanding and forecasting polar stratospheric variability with statistical models. *Atmos. Chem. Phys.*, 12, 5691–5701, doi: 10.5194/acp-12-5691-2012.

Blume, C., K. Matthes, and I. Horenko, 2012: Supervised learning approaches to classify sudden stratospheric warming events. *J. Atmos. Sci.*, 69 (6), 1824-1840, doi: 10.1175/JAS-D-11-0194.1.

Bodeker, G. E., I. S. Boyd, and W. A. Matthews, 1998: Trends and variability in vertical ozone and temperature profiles measured by ozonesondes at Lauder, New Zealand: 1986 to 1996. *J. Geophys. Res.*, 103, 28 661–28 681.

Bond, N. A. and D. E. Harrison, 2000: The Pacific Decadal Oscillation, air-sea interaction and central north Pacific winter atmospheric regimes. *Geophys. Res. Lett.*, 27 (5), 731–734.

Braman, K., R. Byers, and R. Mathias, 2002: The Multishift QR Algorithm. Part I: Maintaining Well-Focused Shifts and Level 3 Performance. *SIAM J. Matrix Anal. Appl*, 23, 929–947.

Burges, C. J. C., 1998: *A tutorial on support vector machines for pattern recognition*. 2d ed., Kluwer

Academic, 121–161 pp.

Burnham, K. P. and D. R. Anderson, 2002: *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 488 pp.

Butler, A. H. and L. M. Polvani, 2011: El Niño, La Niña, and stratospheric sudden warmings: A reevaluation in light of the observational record. *Geophys. Res. Lett.*, 38, L13807.

Cagnazzo, C. and E. Manzini, 2009: Impact of the stratosphere on the winter tropospheric teleconnections between ENSO and the North Atlantic and European region. *J. Climate*, 22 (5), 1223–1238.

Calvo, N., M. A. Giorgetta, R. Garcia-Herrera, and E. Manzini, 2009: Nonlinearity of the combined warm ENSO and QBO effects on the northern hemisphere polar vortex in MAECHAM5 simulations. *J. Geophys. Res.*, 114, D13109.

Camp, C. D. and K.-K. Tung, 2007a: The influence of the solar cycle and QBO on the late-winter stratospheric polar vortex. *J. Atmos. Sci.*, 64 (4), 1267–1283.

Camp, C. D. and K.-K. Tung, 2007b: Stratospheric polar warming by ENSO in winter: A statistical study. *Geophys. Res. Lett.*, 34, L04809.

Charlton, A. J. and L. M. Polvani, 2007: A New Look at Stratospheric Sudden Warmings. Part I: Climatology and Modeling Benchmarks. *J. Climate*, 20, 449–469.

Charney, J. G. and P. G. Drazin, 1961: Propagation of planetary-scale disturbances from the lower into the upper atmosphere. *J. Geophys. Res.*, 66 (1), 83–109.

Charney, J. G. and A. Eliassen, 1949: A numerical method for predicting the perturbations of the middle latitude westerlies. *Tellus*, 1 (2), 38–54.

Coughlin, K. and L. J. Gray, 2009: A continuum of sudden stratospheric warmings. *J. Atmos. Sci.*, 66 (2), 531–540.

Coughlin, K. and K.-K. Tung, 2001: QBO signal found at the Extratropical Surface through Northern Annular Modes. *Geophys. Res. Lett.*, 28 (24), 4563–4566.

Coughlin, K. and K.-K. Tung, 2004: Eleven-year solar cycle signal throughout the lower atmosphere. *J. Geophys. Res.*, 109, D21105.

Crooks, S. and L. Gray, 2005: Characterization of the 11-year solar signal using a multiple regression analysis of the ERA-40 dataset. *J. Climate*, 18, 996–1015.

Cullum, J. K. and R. A. Willoughby, 2002: *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*. Cambridge University Press, 257 pp.

Davison, A. and D. Hinkley, 1997: *Bootstrap methods and their applications*. Cambridge Series in Statistical and Probability Mathematics, 594 pp.

Delworth, T. L. and M. E. Mann, 2000: Observed and simulated multidecadal variability in the Northern Hemisphere. *Climate Dyn.*, 16, 661–676.

Deser, C., M. A. Alexander, S.-P. Xie, and A. S. Phillips, 2010: Sea Surface Temperature Variability: Patterns and Mechanisms. *Ann. Rev. Mar. Sci.*, 2, 115–143.

Dörnbrack, A., M. C. Pitts, L. R. Poole, Y. J. Orsolini, K. Nishii, and H. Nakamura, 2012: The 2009–2010 Arctic stratospheric winter – general evolution, mountain waves and predictability of an operational weather forecast model. *Atmos. Chem. Phys.*, 12 (8), 3659–3675, doi: 10.5194/acp-12-3659-2012.

Efron, B., 1979: Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7 (1), 1–26.

Efron, B. and R. Tibshirani, 1986: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1 (1), 54–75.

Elgaali, E. and L. Garcia, 2004: Neural network modeling of climate change impacts on irrigation water supplies in arkansas river basin. *Proceedings of the 24th Annual Hydrology Days*, Fort Collins, CO, USA.

Eyring, V., et al., 2007: Multimodel projections of stratospheric ozone in the 21st century. *J. Geophys. Res.*, 112, D16303.

Eyring, V., et al., 2008: Overview of the new CCMVal reference and sensitivity simulations in support of upcoming ozone and climate assessments and the planned SPARC CCMVal report. *SPARC Newsletter*, 30, 20–26.

Fawcett, T., 2006: An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27, 861–874.

Fortuin, J. P. F. and H. Kelder, 1998: An ozone climatology based on ozonesonde and satellite measurements. *J. Geophys. Res.*, 103 (D24), 31 709–31 734.

Frame, T. H. A. and L. J. Gray, 2010: The 11-Yr Solar Cycle in ERA-40 Data: An Update to 2008. *J. Climate*, 23 (8), 2213–2222.

Franzke, C., I. Horenko, A. J. Majda, and R. Klein, 2009: Systematic metastable atmospheric regime identification in an AGCM. *J. Atmos. Sci.*, 66 (7), 1997–2012.

Garcia, R. R., T. J. Dunkerton, and R. S. Liebermann, 1997: Climatology of the semiannual oscillation of the tropical middle stratosphere. *J. Geophys. Res.*, 102 (D22), 26,019–26,032.

Garcia, R. R., D. R. Marsh, D. E. Kinnison, B. A. Boville, and F. Sassi, 2007: Simulation of secular trends in the middle atmosphere, 1950-2003. *J. Geophys. Res.*, 112 (D9).

Garfinkel, C. I. and D. L. Hartmann, 2007: Effects of the El-Niño Southern Oscillation and the Quasi-Biennial Oscillation on polar temperatures in the stratosphere. *J. Geophys. Res.*, 112, D19112.

Gerber, E. P., C. Orbe, and L. M. Polvani, 2009: Stratospheric influence on the tropospheric circulation revealed by idealized ensemble forecasts. *Geophys. Res. Lett.*, 36 (24), L24801.

Gerber, E. P. and G. K. Vallis, 2005: A stochastic model for the spatial structure of annular patterns of variability and the Northern Atlantic Oscillations. *J. Climate*, 18, 2102–2118.

Gray, L. J., et al., 2010: Solar influences on climate. *Rev. Geophys.*, 48 (4), RG4001.

Haigh, J. D., 1994: The role of stratospheric ozone in modulating the solar radiative forcing of climate. *Nature*, 370, 544–546.

Haigh, J. D., 1996: The impact of solar variability on climate. *Science*, 272, 981–984.

Hamilton, J. D., 1994: *Time series analysis*. Princeton University Press, 820 pp.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson, 2007: Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Clim.*, 27 (9), 1119–1152.

Hassler, B., G. E. Bodeker, I. Cionni, and M. Dameris, 2009: A vertically resolved, monthly mean, ozone database from 1979 to 2100 for constraining global climate model simulations. *Int. J. Remote Sens.*, 30 (15–16), 4009–4018.

Haynes, P., 2005: Stratospheric dynamics. *Annu. Rev. Fluid Mech.*, 37, 263–293.

Holton, J. R. and H. C. Tan, 1980: The influence of the equatorial Quasi-Biennial Oscillation on the global atmospheric circulation at 50mb. *J. Atmos. Sci.*, 37, 2200–2208.

Holton, J. R. and H. C. Tan, 1982: The Quasi-Biennial Oscillation in the Northern Hemisphere

lower stratosphere. *J. Meteorol. Soc. Jpn.*, 60, 140–148.

Horenko, I., 2010a: On clustering of non-stationary meteorological time series. *Dyn. Atmos. Oceans*, 49 (2–3), 164–187.

Horenko, I., 2010b: On the identification of nonstationary factor models and their application to atmospheric data analysis. *J. Atmos. Sci.*, 67 (5), 1559–1574.

Horenko, I., 2011: Nonstationarity in multifactor models of discrete jump processes, memory and application to cloud modeling. *J. Atmos. Sci.*, 68 (7), 1493–1506.

Hurrell, J. W. and C. Deser, 2009: North Atlantic climate variability: The role of the North Atlantic Oscillation. *J. Mar. Syst.*, 78 (1), 28–41.

IPCC, 2001: *Climate Change 2001: The Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the IPCC*. Cambridge University Press.

IPCC, 2007: *Climate Change 2007 - The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Cambridge University Press.

Jain, A. et al., 1999: Data clustering: A review. *ACM Comput. Surv.*, 31 (3), 265–323.

Jain, A. et al., 2000: Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22 (1), 4–37.

Jaynes, E. T., 2003: *Probability theory: the logic of science*. Cambridge University Press, 758 pp.

Jöckel, P., et al., 2006: The atmospheric chemistry general circulation model ECHAM5/MESSy1: consistent simulation of ozone from the surface to the mesosphere. *Atmos. Chem. Phys.*, 6, 5067–5104.

Jolliffe, I. T., 2002: *Principal Component Analysis*. 2d ed., Springer Series in Statistics, 502 pp.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 364 pp.

Kalnay, E., et al., 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, 77 (3), 437–472.

Keerthi, S. S. and C.-J. Lin, 2003: Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput.*, 15, 1667–1689.

Kinnison, D., et al., 2007: Sensitivity of chemical tracers to meteorological parameters in the MOZART-3 chemical transport model. *J. Geophys. Res.*, 112, D20302.

Kirk, P. D. and M. P. Stumpf, 2009: Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics (Oxford, England)*, 25 (10), 1300–1306.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi, 1983: Optimization by simulated annealing. *Science*, 220 (4598), 671–680.

Kodera, K., 2002: Solar cycle modulation of the North Atlantic Oscillation: Implication in the spatial structure of the NAO. *Geophys. Res. Lett.*, 29 (8), 59.1–59.4.

Kodera, K., 2006: Influence of stratospheric sudden warming on the equatorial troposphere. *Geophys. Res. Lett.*, 33 (6), L06804.

Kodera, K. and Y. Kuroda, 2002: Dynamical response to the solar cycle. *J. Geophys. Res.*, 107 (D24).

Kohavi, R., 1995: A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. Int. Joint Conf. on Artificial Intelligence*, Montreal, QC, Canada, IJCAI, Vol. 2,

1137–1143.

Kurkova, V., 1992: Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5, 501–506.

Kuroda, Y., 2010: High initial-time sensitivity of medium-range forecasting observed for a stratospheric sudden warming. *Geophys. Res. Lett.*, 37, L16804.

Labitzke, K., 1965: On the mutual relation between stratosphere and troposphere during periods of stratospheric sudden warmings in winter. *J. Appl. Meteorol.*, 4, 91–99.

Labitzke, K., 1972: Temperature changes in the mesosphere and stratosphere connected with circulation changes in winter. *J. Atmos. Sci.*, 29, 756–766.

Labitzke, K., 1987: Sunspots, the QBO, and the stratospheric temperatures in the north polar region. *Geophys. Res. Lett.*, 14, 535–537.

Labitzke, K. and M. Kunze, 2005: Stratospheric temperatures over the arctic: comparison of three data sets. *Meteor. Z.*, 14, 65–74.

Labitzke, K. and M. Kunze, 2009a: On the remarkable arctic winter in 2008/09. *J. Geophys. Res.*, 114, D00I02.

Labitzke, K. and M. Kunze, 2009b: Variability in the stratosphere: The Sun and the QBO. Climate and Weather of the Sun-Earth System (CAWSES): Selected Papers from the 2007 Kyoto Symposium, Terrapub, Tokyo, 257–278.

Labitzke, K., M. Kunze, and S. Brönnimann, 2006: Sunspots, the QBO, and the stratosphere in the north polar region - 20 years later. *Meteor. Z.*, 15 (3), 355–363.

Labitzke, K. and B. Naujokat, 2000: The lower arctic stratosphere in winter since 1952. *SPARC Newsletter No.15*, 11–14.

Labitzke, K. and H. van Loon, 1988: Associations between the 11-year solar cycle, the QBO and the atmosphere. part I: the troposphere and stratosphere in the northern hemisphere winter. *J. Atmos. Terr. Phys.*, 50 (3), 197–206.

Labitzke, K. and H. van Loon, 1999: *The Stratosphere: Phenomena, History, and Relevance*. Springer, New York, 179 pp.

Lean, J. L., 2000: Short term, direct indices of solar variability. *Space Sci. Rev.*, 94 (1-2), 39–51.

Liddle, A., 2008: Information criteria for astrophysical model selection. *Mon. Not. R. Astron. Soc.*, 377 (1), L47–L78.

Linkin, M. E. and S. Nigam, 2008: The North Pacific Oscillation-West Pacific teleconnection pattern: Mature-phase structure and winter impacts. *J. Climate*, 21 (9), 1979–1997.

Lu, B. et al., 2009: Nonlinear Representation of the Quasi-Biennial Oscillation. *J. Atmos. Sci.*, 66 (1), 1886–1904.

MacDonald, G. M. and R. A. Case, 2005: Variations in the Pacific Decadal Oscillation over the past millennium. *Geophys. Res. Lett.*, 320, L08703.

Manzini, E., M. Giorgetta, M. Esch, L. Kornblueh, and E. Roeckner, 2006: The influence of sea surface temperatures on the northern winter stratosphere: Ensemble simulations with the MAECHAM5 model. *J. Climate*, 19, 3863–3881.

Marques de Sá, J. P., 2001: *Pattern recognition: concepts, methods, and applications*. Springer, 318 pp.

Martius, O., L. M. Polvani, and H. C. Davies, 2009: Blocking precursors to stratospheric sudden warming events. *Geophys. Res. Lett.*, 36, L14806.

Matsuno, T., 1971: A dynamical model of the stratospheric sudden warming. *J. Atmos. Sci.*, 28, 1479–1494.

Matthes, K., U. Langematz, L. Gray, K. Kodera, and K. Labitzke, 2004: Improved 11-Year Solar Signal in the Freie Universität Berlin Climate Middle Atmosphere Model (FUB-CMAM). *J. Geophys. Res.*, 109, D06101.

Matthes, K., D. Marsh, R. Garcia, D. Kinnison, F. Sassi, and S. Walters, 2010: Role of the QBO in modulating the influence of the 11 year solar cycle on the atmosphere using constant forcings. *J. Geophys. Res.*, 115, D18110.

Matthewman, N. J., J. G. Esler, A. J. Charlton-Perez, and L. M. Polvani, 2009: A New Look at Stratospheric Sudden Warmings. Part III: Polar Vortex Evolution and Vertical Structure. *J. Climate*, 22 (6), 1566–1585.

McIntyre, M. E., 1982: How well do we understand the dynamics of stratospheric warmings? *J. Meteor. Soc. Japan*, 60, 37–65.

McQuarrie, A. D. R. and C.-L. Tsai, 1998: *Regression and Time Series Model Selection.* World Scientific, 455 pp.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953: Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21 (6), 1087–1092.

Mitchell, D. M., A. J. Charlton-Perez, and L. J. Gray, 2011: Characterizing the Variability and Extremes of the Stratospheric Polar Vortices Using 2D Moment Analysis. *J. Atmos. Sci.*, 68, 1194–1213.

Mitchell, D. M., L. J. Gray, and A. J. Charlton-Perez, 2011: The structure and evolution of the stratospheric vortex in response to natural forcings. *Geophys. Res. Lett.*, 116, D15110.

Mohanakumar, K., 2008: *Stratosphere Troposphere Interactions*. Springer, 433 pp.

Montgomery, D., E. A. Peck, and G. G. Vining, 2006: *Introduction to linear regression analysis*. 4th ed., Wiley-Interscience, 612 pp.

Morgenstern, O., et al., 2010: Review of the formulation of present-generation stratospheric chemistry-climate models and associated external forcings. *J. Geophys. Res.*, 115.

Nao, I., A. Naomi, and H. Sachiko, 2006: Estimation of the surface area density of polar stratospheric clouds by using the support vector machines. *IPSJ SIG Tech. Reports*, 2006 (95), Z0031B.

Naoe, H. and K. Shibata, 2010: Equatorial Quasi-Biennial Oscillation influence on northern winter extratropical circulation. *J. Geophys. Res.*, 115, D19102.

Naujokat, B., K. Labitzke, R. Lenschow, K. Petzoldt, and R.-C. Wohlfahrt, 1988: The stratospheric winter 1987/88: An unusually early major midwinter warming. *Beilage zur Berliner Wetterkarte*, SO 6/88, 20 pp.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling error in the estimation of empirical orthogonal function. *Mon. Weather Rev.*, 110, 699–706.

Picard, R. R. and R. D. Cook, 1984: Cross-validation of regression models. *J. Amer. Statist. Assoc.*, 79 (387), 575–583.

Priestley, M. B., 1983: *Spectral Analysis and Time Series*. Academic Press, 890 pp.

Randel, W. J., R. R. Garcia, N. Calvo, and D. Marsh, 2009: ENSO influence on zonal mean

temperature and ozone in the tropical lower stratosphere. *Geophys. Res. Lett.*, 36, L15822.

Randel, W. J. and A. M. Thompson, 2011: Interannual variability and trends in tropical ozone derived from SAGE II satellite data and SHADOZ ozonesondes. *J. Geophys. Res.*, 116, D07303.

Randel, W. J. and F. Wu, 2007: A stratospheric ozone profile data set for 1979-2005: Variability, trends, and comparisons with column ozone data. *J. Geophys. Res.*, 112, D06313.

Randel, W. J., et al., 2004: The SPARC intercomparison of middle-atmosphere climatologies. *J. Climate*, 17 (5), 986–1003.

Randel, W. J. et al., 2009: An update of observed stratospheric temperature trends. *J. Geophys. Res.*, 114, D02107.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, 108 (D14).

Richter, J. H., K. Matthes, N. Calvo, and L. J. Gray, 2011: Influence of the Quasi-Biennial Oscillation and El Niño-southern oscillation on the frequency of sudden stratospheric warmings. *J. Geophys. Res.*, 116, D20111.

Richter, J. H., F. Sassi, and R. R. Garcia, 2010: Toward a Physically Based Gravity Wave Source Parameterization in a General Circulation Model. *J. Atmos. Sci.*, 67 (1), 136–156.

Rienecker, M. M. et al., 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, 24 (14), 3624–3648.

Ripley, B. D., 1996: *Pattern Recognition and Neural Networks*. Cambridge University Press, 415 pp.

Robock, A., 2000: Volcanic eruptions and climate. *Rev. Geophys.*, 38, 191–219.

Robock, A. and J. P. Mao, 1995: The volcanic signal in surface-temperature observations. *J. Climate*, 8 (5, Part 1), 1086–1103.

Roeckner, E., et al., 2003: The atmospheric general circulation model ECHAM5. PART I: model description. Tech. Rep. 349, MPI-M, Hamburg.

Roeckner, E., et al., 2004: The atmospheric general circulation model ECHAM5. PART II: The atmospheric general circulation model ECHAM5 Part II: Sensitivity of simulated climate to horizontal and vertical resolution. Tech. Rep. 354, MPI-M, Hamburg.

Salby, M. L. and P. F. Callaghan, 2006: Relationship of the Quasi-Biennial Oscillation to the stratospheric signature of the solar cycle. *J. Geophys. Res.*, 111 (D6).

Sato, M., J. E. Hansen, M. P. McCormich, and J. B. Pollack, 1993: Stratospheric aerosol optical depths, 1850-1990. *J. Geophys. Res.*, 98 (D12), 22 987–22 994.

Scherhag, R., 1952: Die explosionsartigen Stratosphärenerwärmungen des Spätwinters 1951/52. *Berichte des deutschen Wetterdienstes in der US-Zone*, 6, Nr.38, 51–63.

Schlesinger, M. E. and N. Ramankutty, 1994: An oscillation in the global climate system of period 65-70 years. *NATURE*, 367 (6465), 723–726.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, 6 (2), 461–464.

Shibata, K. and M. Deushi, 2008: Long-term variations and trends in the simulation of the middle atmosphere 1980-2004 by the chemistry-climate model of the Meteorological Research Institute. *Ann. Geophys.*, 26 (5), 1299–1326.

Shibata, K., M. Deushi, T. T. Sekiyama, and H. Yoshimura, 2005: Development of an MRI chemical

transport model for the study of stratospheric chemistry. *Papers Met. Geophys.*, 55, 75–119.

Shibata, K., H. Yoshimura, M. Ohizumi, M. Hosaka, and M. Sugi, 1999: A simulation of troposphere, stratosphere and mesosphere with an MRI/JMA98 GCM. *Papers Met. Geophys.*, 55 (1), 15–53.

Simmons, A., S. M. Uppala, D. Dee, and S. Kobayashi, 2006: ERA-Interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter*, 110, 26–35.

Smola, A. J., N. Murata, B. Schölkopf, and K.-R. Müller, 1998: Asymptotically optimal choice of ε-loss for support vector machines. *Proceedings of the Internat. Conf. on Artific. Neural Net.*, Springer, 105–110, Perspectives in Neural Computing.

Solomon, S., 1999: Stratospheric ozone depletion: A review of concepts and history. *Rev. Geophys.*, 37 (3), 275–316.

Spackman, K. A., 1989: Signal detection theory: valuable tools for evaluating inductive learning. *Proceedings of the sixth international workshop on Machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 160–163.

SPARC CCMVal, 2010: SPARC Report on the Evaluation of Chemistry-Climate Models, V. Eyring, T. G. Shepherd, D. W. Waugh (Eds.). *SPARC Report No. 5, WCRP-132, WMO/TD-No. 1526*, `www.atmosp.physics.utoronto.ca/SPARC`, C. **Blume** contributed to chapter 8.

Staehelin, J., N. R. P. Harris, C. Appenzeller, and J. Eberhard, 2001: Ozone trends: A review. *Rev. Geophys.*, 39 (2), 231–290.

Stix, M., 2004: *The Sun: An Introduction*. 2d ed., Springer, 490 pp.

Stolarski, R. S., A. R. Douglass, P. A. Newman, S. Pawson, and M. R. Schoeberl, 2010: Relative contribution of greenhouse gases and ozone-depleting substances to temperature trends in the stratosphere: A chemistry-climate model study. *J. Climate*, 23 (1), 28–42.

Thompson, D. W. J., 2003: Stratospheric connection to northern hemisphere wintertime weather: Implications for prediction. *J. Climate*, 15, 1421–1428.

Thompson, D. W. J. and S. Solomon, 2002: Interpretation of recent southern hemisphere climate change. *Science*, 296 (5569), 895–899.

Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.*, 78, 2771–2777.

Turkheimer, F. E., R. Hinz, and V. J. Cunningham, 2003: On the undecidability among kinetic models: from model selection to model averaging. *J. Cereb. Blood Flow Metab.*, 23, 490–498.

Uppala, S. M., et al., 2005: The ERA-40 re-analysis. *Q. J. Roy. Meteor. Soc.*, 131, 2961–3012.

Vapnik, V., 1995: *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 188 pp.

Varotsos, C., 2004: The extraordinary events of the major, sudden stratospheric warming, the diminutive antarctic ozone hole, and its split in 2002. *Environ. Sci. Pollut. R.*, 11, 405–411.

von Storch, H. and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.

Wallace, J. M., R. L. Panetta, and J. Estberg, 1993: A phase-space representation of the equatorial stratospheric Quasi-Biennial Oscillation. *J. Atmos. Sci.*, 50 (12), 1751–1762.

Walter, A. and C.-D. Schönwiese, 2002: Attribution and detection of anthropogenic climate change using a backpropagation neural network. *Meteor. Z.*, 11 (5), 335–343.

White, W. B. and Z. Liu, 2008: Non-linear alignment of El Niño to the 11-yr solar cycle. *Geophys. Res. Lett.*, 35, L19607.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*, International Geophysics Series, Vol. 59. Academic Press, 467 pp.

WMO, 2007: *Scientific Assessment of Ozone Depletion: 2006*. World Meteorological Organization, Geneva, 572 pp., Global Ozone Research and Monitoring Project-Report No. 50.

WMO, 2011: *Scientific Assessment of Ozone Depletion: 2010*. World Meteorological Organization, Geneva, 516 pp., Global Ozone Research and Monitoring Project-Report No. 52.

Woollings, T. and B. Hoskins, 2008: Simultaneous Atlantic-Pacific blocking and the Northern Annular Mode. *Q. J. Roy. Meteor. Soc.*, 134, 1635–1646.

Yoshida, K. and K. Yamazaki, 2011: Tropical cooling in the case of stratospheric sudden warming in January 2009: focus on the tropical tropopause layer. *Atmos. Chem. Phys.*, 11 (13), 6325–6336.

Zhang, G., B. E. Patuwo, and M. Y. Hu, 1998: Forecasting with artificial neural networks: The state of the art. *Int. J. Forecasting*, 14, 35–62.

Abbreviations

AIC	Akaike Information Criterion
AM	Annular Mode
AMO	Atlantic Multidecadal Oscillation
AOD	Aerosol Optical Depth
BIC	Bayesian Information Criterion
BLOC	Index representing high-latitude blockings
BLOC1	First principal component of 500 hPa geopotential anomalies from 35° N to 85° N
BLOC2	Second principal component after BLOC1
CCM	Chemistry-Climate Model
CCMVal	Chemistry-Climate Model Validation Activity
CFC	Chlorofluorocarbon
EMAC	ECHAM5/MESSy Atmospheric Chemistry model
ENSO	El Niño- Southern Oscillation
EOF	Empirical orthogonal function
ERA	ECMWF re-analysis
ESC	Effective stratospheric chlorine
EV	Explained variance
FEM-VARX	Finite element method plus vector auto-regression with external factors
FVX	FEM-VARX
GHG	Greenhouse gas
LDA	Linear discriminant analysis
LSVM	Linear support vector machine
MERRA	Modern Era Retrospective Analysis for Research and Applications
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MPE	Mean Prediction Error
MRI	Meteorological Research Institute model
NAM	Northern Annular Mode
NAO	North-Atlantic Oscillation
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCEP	National Center for Environmental Prediction

NIWA	National Institute of Water and Atmospheric Research
NOAA	National Oceanic and Atmospheric Administration
NPO/WP	North Pacific Oscillation- West Pacific teleconnection pattern
ODS	Ozone depleting substances
PCA	Principal Component Analysis
PDO	Pacific Decadal Oscillation
QBO	Quasi-Biennial Oscillation
QBO1	First principal component of equatorial stratospheric zonal mean zonal wind anomalies
QBO2	Second principal component after QBO1
ROC	Receiver operating characteristic
SAM	Southern Annular Mode
SAO	Semiannual Oscillation
SFL	11-yr solar cycle
SI	Seasonal influence
SPARC	Stratospheric Processes And their Role in Climate
SSE	Sum of squared errors
SST	Sea surface temperature
SSW	Sudden Stratospheric Warming
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression
TRE	Linear trend term
TSI	Total solar irradiance
UV	Ultra-violet
WACCM	Whole Atmosphere Community Climate Model
WCRP	World Climate Research Program
WMO	World Meteorological Organization

Acknowledgments

I would like to begin by thanking the people whom have supported me in not only my academic pursuits but in all my life decisions, my family. Without them this work would have not been possible and I am forever grateful to them for their support. In particular, my loving wife Amanda, with her endless patients, has helped me through difficult times and was always there when I needed her.

Warm acknowledgments to my supervisor Katja Matthes whose guidance and support helped make this thesis all that it is. Katja gave me the opportunity to do research in the fields of atmospheric science and statistical modeling. I would also like to thank Illia Horenko and Olga Kaiser for their valuable contributions and insight into numerous mathematical aspects used in my work. Sincere thanks to the current and former members of the NATHAN research group, in particular Christof Petrick, Lisa Neef, Christopher Kadow, and Felicitas Hansen for enlightening discussions and advice. I appreciate the great support from all my former colleagues of the German Research Centre of Geosciences in Potsdam.

I want to continue by thanking the members of the Stratosphere Research Group at the Institute for Meteorology, in particular Anne Kubin, Markus Kunze, Ulrike Langematz, and Karin Labitzke who have acquainted me with stratospheric insights and supported me in many ways. I am grateful to Ulrike for preparing a review of this thesis. Many thanks to Greg Bodeker, Christiane Jablonowski, Edwin Gerber, and Kuni Kodera for their scientific advice.

I would also like to thank my colleagues for many stimulating discussions during lunch and coffee breaks, in particular Peter Névir, Jens Grieger, Henning Rust, Tobias Pardowitz, and Tim Kruschke. Special thanks to Thomas Bergmann for his great technical support and unconditional good mood.

My final acknowledgments shall go to the great work of numerous developers of operating systems, programming languages, computing libraries, and applications that I have happily used during my research. These are Linux, Debian, Ubuntu, C++, C, Python, PHP, HTML, NCL, MATLAB, ROOT, TMVA, LIBSVM, Intel MKL, OpenMP, NetCDF, Armadillo, GCC, Make, Emacs, LaTeX, Aspell, CDO, SSH, SVN, rsync, Gimp, Ghostscript, and many more.

Selbstständigkeitserklärung

Hiermit erkläre ich an Eides Statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe angefertigt, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Berlin, 30. April 2012

"The CV is not included in the online version for reasons of data protection."

"The CV is not included in the online version for reasons of data protection."