OPTIMAL RECONSTRUCTION OF U.S. TEMPERATURE AND PRECIPITATION
SINCE 1895

A Thesis
Presented to the
Faculty of
San Diego State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in
Statistics

by
Carina Meta Mueller
Summer 2013
SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the
Thesis of Carina Meta Mueller:

Optimal Reconstruction of U.S. Temperature and Precipitation since 1895

Samuel Shen, Chair
Department of Mathematics and Statistics

Barbara Bailey
Department of Mathematics and Statistics

Trent Biggs
Department of Geography

07/26/2013
Approval Date
Copyright © 2013
by
Carina Meta Mueller
DEDICATION

To my beloved parents.
ABSTRACT OF THE THESIS

Optimal Reconstruction of U.S. Temperature and Precipitation since 1895
by
Carina Meta Mueller
Master of Science in Statistics
San Diego State University, 2013

This thesis studies optimal reconstruction and its errors of the United States surface air temperature and precipitation since 1895 and extensively uses Empirical Orthogonal Functions (EOF). The reconstruction is based on multivariate linear regression analysis using EOFs as the design matrix. Two datasets, named TOB and F52 from the United States Historical Climatology Network (USHCN), are used to build the model. Subsequently, the time series of these two datasets and the one obtained from our reconstruction are compared for error estimation through cross comparison analysis. The reconstructed data is very close to the slightly adjusted TOB data and has a larger deviation from the fully adjusted F52 dataset. The mean squared error (MSE) of the reconstructed temperature was computed. A lower MSE for temperature was observed in the summer months than in the winter months.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2 Empirical Orthogonal Functions</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Theoretical Computation</td>
<td>3</td>
</tr>
<tr>
<td>2.2 EOFs and Eigenvalues for USHCN temperature data</td>
<td>5</td>
</tr>
<tr>
<td>2.3 EOFs and Eigenvalues for USHCN precipitation data</td>
<td>16</td>
</tr>
<tr>
<td>3 Method of reconstruction</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Multivariate regression reconstruction method</td>
<td>23</td>
</tr>
<tr>
<td>3.2 Application of the Reconstruction Method</td>
<td>25</td>
</tr>
<tr>
<td>4 Area weighted averages and Error Estimation</td>
<td>27</td>
</tr>
<tr>
<td>4.1 Area weighted averages</td>
<td>27</td>
</tr>
<tr>
<td>4.2 Error Estimation</td>
<td>34</td>
</tr>
<tr>
<td>5 Conclusion and Discussion</td>
<td>42</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>43</td>
</tr>
<tr>
<td>APPENDIX</td>
<td></td>
</tr>
<tr>
<td>Codes</td>
<td>44</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Cumulative Percentage of Eigenvalues of $T_{max}$</td>
<td>16</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Mean values of $T_{min}$, $T_{avg}$ and $T_{max}$ anomalies by period (TOB data)</td>
<td>29</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1. EOF1 of $T_{avg}$ of January ......................................................... 6
Figure 2.2. EOF2 of $T_{avg}$ of January .......................................................... 6
Figure 2.3. EOF3 of $T_{avg}$ of January .......................................................... 7
Figure 2.4. EOF4 of $T_{avg}$ of January .......................................................... 7
Figure 2.5. EOF5 of $T_{avg}$ of January ......................................................... 8
Figure 2.6. EOF1 of $T_{avg}$ of February ........................................................ 8
Figure 2.7. EOF1 of $T_{avg}$ of March ........................................................... 9
Figure 2.8. EOF1 of $T_{avg}$ of April ............................................................. 9
Figure 2.9. EOF1 of $T_{avg}$ of Maı ............................................................... 10
Figure 2.10. EOF1 of $T_{avg}$ of June ........................................................... 10
Figure 2.11. EOF1 of $T_{avg}$ of July ........................................................... 11
Figure 2.12. EOF1 of $T_{avg}$ of August ......................................................... 11
Figure 2.13. EOF1 of $T_{avg}$ of September .................................................... 12
Figure 2.14. EOF1 of $T_{avg}$ of October ....................................................... 12
Figure 2.15. EOF1 of $T_{avg}$ of November ..................................................... 13
Figure 2.16. EOF1 of $T_{avg}$ of December ..................................................... 13
Figure 2.17. Eigenvalues versus Mode Number by Month ($T_{max}$) ................. 14
Figure 2.18. Cumulative Percentage of the Eigenvalues versus Mode Number ($T_{max}$) ............................................................ 15
Figure 2.19. EOF1 of precipitation data of January ............................................. 17
Figure 2.20. EOF2 of precipitation data of January .......................................... 17
Figure 2.21. EOF3 of precipitation data of January .......................................... 18
Figure 2.22. EOF1 of precipitation data of June .............................................. 18
Figure 2.23. EOF2 of precipitation data of June ............................................. 19
Figure 2.24. EOF3 of precipitation data of June ............................................. 19
Figure 2.25. Eigenvalues versus Mode Number by Month (Precipitation Data) .......... 20
Figure 2.26. Cumulative Percentage of the Eigenvalues versus Mode Number (Precipitation Data) ............................................................ 21
Figure 3.1. Number of Grid Boxes with Temperature Data ............................................ 22
Figure 3.2. Temperature anomalies of $T_{avg}$ in January 1897 .................................. 23
Figure 3.3. Reconstructed $T_{avg}$ of January 1897 ......................................................... 26
Figure 4.1. Annual Mean of $T_{avg}$ of TOB dataset ...................................................... 27
Figure 4.2. Annual Mean of $T_{min}$ of TOB dataset ...................................................... 28
Figure 4.3. Annual Mean of $T_{max}$ of TOB dataset ...................................................... 28
Figure 4.4. Annual Mean of $T_{avg}$ of reconstructed dataset ......................................... 30
Figure 4.5. Annual Mean of $T_{min}$ of reconstructed dataset ....................................... 30
Figure 4.6. Annual Mean of $T_{max}$ of reconstructed dataset ....................................... 31
Figure 4.7. Annual Mean of $T_{min}$ .............................................................. 32
Figure 4.8. Annual Mean of $T_{max}$ .............................................................. 33
Figure 4.9. Annual Mean of $T_{avg}$ .............................................................. 34
Figure 4.10. Square Root of MSE of $T_{min}$ over time ........................................ 39
Figure 4.11. Square Root of MSE of $T_{min}$ over time by month ............................... 40
Figure 4.12. Square Root of MSE of $T_{min}$ for each grid box ................................... 41
ACKNOWLEDGMENTS

First and foremost, I would like to express my deep gratitude to my advisor Professor Samuel Shen for his patience and guidance during my research. Throughout the entire year his door has always been open and his personal support and broad knowledge was very encouraging to me.

Second, I want to thank Professor Barbara Bailey not only for her help during my research but also for her commitment and enthusiasm in her role as advisor for my program of study.

Many thanks to Professor Trent Biggs for kindly serving as third member in my thesis committee.

Furthermore, my thanks go to the Fakultät für Mathematik und Wirtschaftswissenschaften der Universität Ulm and in particular to Professor Hans-Joachim Zwiesler for giving me the opportunity to study abroad at San Diego State University.

A big thank you to Jordan Schafer, I am very grateful for proofreading my thesis.

Last but not least I would like to give heartfelt thanks to my parents for their sincere support throughout my entire life and especially this year.
CHAPTER 1

INTRODUCTION

Our earth is warming. Every child knows that. Climate change and global warming have never been talked more about than they are today, even though this knowledge is not that new. In 1824 Jean-Baptiste Joseph Fourier explained the basics of what is now known as the greenhouse effect ([1], p.3). But the climate is a very complex system with very many influential variables. In order to understand the connections in between these we need to record the temperature, the precipitation and so on throughout time to be able to build a model to describe the observed data. Very roughly speaking this is what this thesis is about.

The dataset which was used for the computations in this thesis was provided by the United States Historical Climatology Network (USHCN). All the datasets can be downloaded on the following website: ftp://ftp.ncdc.noaa.gov/pub/data/ushcn/v2/monthly/

Concerning the temperature records one can find three different datasets. One for the minimum, one for the maximum and one for the average temperature of each station in each considered month. We will denote them by $T_{\text{min}}$, $T_{\text{max}}$ and $T_{\text{avg}}$. The average temperature here is not the actual mean temperature in the given month, but rather the average of minimum and maximum temperature. For each of these datasets then again there exist three different versions. There are the raw data, the TOB data and the F52 data. The TOB data is adjusted for the time of observation bias (TOB), i.e. the bias caused by a change in the time of day when the observations were recorded throughout the years. The adjustment method is described in Menne et al. (2009) [5]. The F52 data is fully adjusted for other biases and all missing values are interpolated. See Menne and Williams (2009) [4] for details of the adjustment method. In this thesis the F52 datasets are used for the computation of the Empirical Orthogonal Functions and the TOB data was applied for the reconstruction method. The raw data will not be used. Each dataset contains weather records from a total of 1218 different stations throughout the whole U.S. excluding Hawaiï and Alaska, whereas not every station has data throughout the whole considered time period from 1895-2012. The temperature is recorded in tenth of degrees Fahrenheit. However, in this thesis we use degree Celsius, i.e. the data had to be converted.

Regarding the precipitations dataset the USHCN provides the F52 dataset which contains unadjusted total monthly precipitation values with missing values estimated. The unit the precipitation is measured in is hundredths of inches. For this thesis the data has been converted to millimeters.
For both temperature and precipitation data some more transformations had to be made before the real analysis was conducted. Instead of using the absolute temperature and precipitation values, we used anomalies from a certain base period. This has an advantage in that it makes weather data from completely different regions, such as desert and mountains, comparable. The years 1961-1990 have been chosen as the base period according to some criteria in Lee (2009) [2]. Thereby, the temperature and precipitation, respectively, was averaged over this 30 years per month and per station. After that, the anomaly of each data value from this mean was computed and used for the following analysis.

Furthermore it is common practice in climate research to divide the observed part of the earth into grid boxes and average the weather data of all stations in this grid box in order to have one temperature or precipitation value per grid box. Therefore we place a grid of size $2.5^\circ \times 3.5^\circ$ over the U.S. starting in the southwest at coordinates (126 W, 22.5 N) to the northeast at coordinates (66.5 W, 50 N), which results in 187 grid boxes. However, only 120 of these grid boxes contain weather stations because the remaining ones are located over the ocean. This $2.5^\circ \times 3.5^\circ$ shape of the grid boxes results in almost square boxes because of the shape of the Earth, as is described in Wied (2012) [13].

The goal is now to find a model that describes the collected data. The method to find this model is called reconstruction and is based on multivariate regression techniques. Furthermore, we will have a look at the error that is connected to this method.

The second chapter introduces the Empirical Orthogonal Functions. The first section explains some theory on which the implemented code (that is computing the EOFs) is based. The second and third part contain the actual EOFs that are gained from the USHCN temperature and precipitation data, respectively.

In the third chapter we will expand on the reconstruction method. The first part explains then again the theory and the second part contains the results of applying the theory to the data.

In Chapter 4 we will start with examining the time series of the different temperature and precipitation datasets. Then the calculation of the mean squared error which the reconstruction method comes with will be outlined and the results of the error estimation will be discussed.
CHAPTER 2

EMPirical ORTHOGONAL FUNCTIONS

For the actual reconstruction of the temperature and precipitation data we will make use of the so called Empirical Orthogonal Functions. The purpose of this chapter is therefore to familiarize the reader with these and visualize them by means of the USHCN data.

2.1 THEORETICAL COMPUTATION

In order to study a spatial pattern and its change in time empirical orthogonal function (EOF) analysis or principal component analysis (PCA), as it is also referred to in literature, is often used in climate research [7]. The goal is to reduce dimensionality by identifying patterns of variation and expressing the data as a linear combination of these patterns. In Lehmann (2012) [3] we can find the derivation of the EOFs over a spatial field. This is what we are theoretically interested in because we can consider the temperature distribution in the U.S. as a continuous climate field. However, since it is not possible to measure the temperature/precipitation at every single location in the U.S. we use grid boxes -as mentioned earlier- to discretize the spatial field. For that reason it will be demonstrated in the following, how we actually proceed to come up with the EOFs in the discrete case.

Let $X$ be a $p$-dimensional random vector with expected value 0. Finding the first EOF is equivalent to finding a vector $e_1$ with $\| e_1 \| = 1$, which minimizes the error

$$
\epsilon_1 = E(\| X - (X \cdot e_1)e_1 \|^2)
$$

(2.1)

Here, $\cdot$ is used for the inner product and $\| \cdot \|$ denotes the Euclidean norm.

The error formula in 2.1 corresponds to the difference of the data and the projection of the data onto a subspace spanned by $e_1$. Minimizing $\epsilon_1$ is equivalent to maximizing the variance contained in this subspace since

$$
\epsilon_1 = E\left( (X - (X \cdot e_1)e_1)'(X - (X \cdot e_1)e_1) \right)
= E\left( \|X\|^2 - 2(X \cdot e_1)'X e_1 + (X \cdot e_1)(X \cdot e_1) \right)
= E\left( \|X\|^2 - (X \cdot e_1)(X \cdot e_1) \right)
= Var(X) - Var(X \cdot e_1)
= Var(X) - e_1' \Sigma e_1
$$

(2.2)
where $\Sigma$ is the covariance matrix of $X$. Obviously, the error $\epsilon_1$ gets smaller, when $Var(X \cdot e_1)$ increases. The maximum value, that this term can take on, is of course $Var(X)$. Since this will not happen, unless all data points are actually already located in the subspace spanned by $e_1$, we now need to minimize $\epsilon_1$ under the constraint $\|e_1\|^2 = 1$. This leads us to the following system of linear equations. The Lagrange multiplier is denoted by $\lambda$.

\[
\frac{d}{de_1}(-e_1^t \Sigma e_1 + \lambda(\|e_1\|^2 - 1)) = 2\Sigma e_1 + 2\lambda e_1 = 0
\]

\[
\|e_1\|^2 - 1 = 0
\]  

(2.3)  

(2.4)  

Solving this, results in:

\[
\Sigma e_1 = \lambda e_1
\]  

(2.5)  

That means that $e_1$ is an eigenvector of the covariance matrix $\Sigma$ of $X$ with corresponding eigenvalue $\lambda$. In order to minimize $\epsilon_1$ we will choose the eigenvector out of the $p$ different ones which corresponds to the largest eigenvalue, since it accounts for the largest variance of $X \cdot e_1$.

\[
Var(X \cdot e_1) = e_1^t \Sigma e_1 = e_1^t \lambda e_1 = \lambda
\]  

(2.6)  

That means we chose the largest eigenvalue because it maximizes $Var(X \cdot e_1)$. The vector $e_1$ is now our first EOF.

In order to find the second EOF, we repeat the described steps, minimizing

\[
\epsilon_2 = E(\|(X - (X \cdot e_1)e_1 - (X \cdot e_2)e_2\|^2)
\]  

(2.7)  

where $e_1$ is the first EOF that we just found and $e_2$ denotes the vector which minimizes $\epsilon_2$, i.e. the second EOF. It turns out that the second EOF is the eigenvector corresponding to the second largest eigenvalue. We can move on to find all $p$ EOFs.

It is worth mentioning that the EOFs are orthogonal, as their name suggests, since the eigenvectors of a symmetric matrix are orthogonal. Furthermore, the eigenvalues also have an interesting interpretation. Since

\[
Var(X \cdot e_i) = e_i^t \Sigma e_i = e_i^t \lambda_i e_i = \lambda_i, \quad i = 1, \ldots p
\]  

(2.8)  

the $i$-th eigenvalue is a measure of the variance explained by the $i$-th eigenvector. If we divide each eigenvalue by the sum of all eigenvalues we can then interpret the resulting values as percentage of variance explained by a certain corresponding eigenvector.


2.2 EOFs and Eigenvalues for USHCN Temperature Data

For the EOF computation we use the fully adjusted F52 dataset, which has no missing values. As mentioned earlier, we don’t use absolute temperature values but anomalies from the base period 1961-1990. Let’s denote the matrix containing the temperature anomalies for a given month \( k \) by \( T_k(i, t) \). The row index runs from 1 to 120 since there are 120 grid boxes and the column index goes up until 30 because of the 30 years of the base period. Before computing the covariance matrix, we need to area-weight the data matrix. Thereby, \( A(x) \) denotes the weight for grid box \( x \). Denoting the latitude of the \( x \)-th grid box by \( \phi_x \), the relative weight for grid box \( x \) is computed as the following:

\[
A(x) = \frac{\cos \left( \frac{\pi \phi_x}{180} \right)}{\sum_{i=1}^{N} \cos \left( \frac{\pi \phi_i}{180} \right)} \tag{2.9}
\]

See Wied (2012) [13] for further details on the derivation of the area weights. Using 2.9 we can now proceed by calculating the covariance matrix \( C_k \) of the area weighted temperature anomalies for a given month \( k \).

\[
\begin{bmatrix} C_k(i, j) \end{bmatrix}_{120 \times 120} = \frac{1}{30} \sum_{t=1}^{30} \sqrt{A(i)}T_k(i, t)T_k(j, t)\sqrt{A(j)} \tag{2.10}
\]

Solving the eigenvalue problem of this covariance matrix gives us the eigenvalues and the eigenvectors or EOFs for a given month \( k \). We call the eigenvector corresponding to the biggest eigenvalue first EOF mode, the one corresponding to the second biggest eigenvalue second EOF mode and so on. All 30 EOF modes were computed and normalized for the three different temperature datasets \( T_{min}, T_{max} \) and \( T_{avg} \) for all twelve months. Figures 2.1 - 2.5 show the pattern of the first five EOF modes of January of the \( T_{avg} \) dataset plotted on a map of the U.S. We notice that the first EOF has a pretty rough pattern. The orthogonality of the modes appears, since the patterns of two consecutive modes seem like the “opposite” of each other. Moreover, we notice that the higher the mode the less of a pattern -like the first mode. This is because the first mode explains around 51% of the variance, i.e. the first mode sketches quite roughly the main pattern. Higher modes explain very small fluctuations in the data.
Figure 2.1. EOF1 of $T_{avg}$ of January

Figure 2.2. EOF2 of $T_{avg}$ of January
Figure 2.3. EOF3 of $T_{avg}$ of January

Figure 2.4. EOF4 of $T_{avg}$ of January
Figure 2.5. EOF5 of $T_{avg}$ of January

Figure 2.6 - 2.16 show the first mode of the EOFs of the average temperature $T_{avg}$ of the other months February-December.

Figure 2.6. EOF1 of $T_{avg}$ of February
Figure 2.7. EOF1 of $T_{avg}$ of March

Figure 2.8. EOF1 of $T_{avg}$ of April
Figure 2.9. EOF1 of $T_{avg}$ of Mai

Figure 2.10. EOF1 of $T_{avg}$ of June
Figure 2.11. EOF1 of $T_{avg}$ of July

Figure 2.12. EOF1 of $T_{avg}$ of August
Figure 2.13. EOF1 of $T_{avg}$ of September

Figure 2.14. EOF1 of $T_{avg}$ of October
Figure 2.15. EOF1 of $T_{avg}$ of November

Figure 2.16. EOF1 of $T_{avg}$ of December

Figure 2.17 shows the eigenvalues of $T_{max}$ plotted versus the modes for each month. We can see that they decrease quite fast. That means that most of the variance in the data is explained by the first few modes. In order to express the data through a linear combination of the EOFs, just the first few EOFs are necessary to approximate the data quite well, i.e.
without making a significant error.

Figure 2.17. Eigenvalues versus Mode Number by Month ($T_{max}$)

Figure 2.18 shows the almost the same plot, just now we can see the cumulative percentage of the variance explained. For example the first eigenvalue for January explains approximately 53% of the variance in the data. The first two eigenvalues of January together account for around 72% of the variance and so on. These values can be found in Table 2.1. For the reconstruction method that is described in the next chapter, we will use as many modes such that the cumulated percentage of the eigenvalues is greater or equal to 90%. The dashed line in Figure 2.18 refers to this threshold. We need at least five modes to account for at least 90% of the variance in this dataset ($T_{max}$).
Figure 2.18. Cumulative Percentage of the Eigenvalues versus Mode Number ($T_{max}$)

We also notice in the previous two figures that in the warmer months June-August we need more modes to exceed this 90% threshold than in the other months.
Table 2.1. Cumulative Percentage of Eigenvalues of $T_{max}$

<table>
<thead>
<tr>
<th>Mode</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.53</td>
<td>0.53</td>
<td>0.54</td>
<td>0.48</td>
<td>0.40</td>
<td>0.38</td>
<td>0.33</td>
<td>0.34</td>
<td>0.49</td>
<td>0.44</td>
<td>0.41</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>0.72</td>
<td>0.70</td>
<td>0.72</td>
<td>0.69</td>
<td>0.57</td>
<td>0.59</td>
<td>0.55</td>
<td>0.61</td>
<td>0.69</td>
<td>0.79</td>
<td>0.70</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
<td>0.82</td>
<td>0.72</td>
<td>0.71</td>
<td>0.68</td>
<td>0.71</td>
<td>0.76</td>
<td>0.84</td>
<td>0.82</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>0.87</td>
<td>0.82</td>
<td>0.78</td>
<td>0.75</td>
<td>0.78</td>
<td>0.81</td>
<td>0.88</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>5</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>0.89</td>
<td>0.87</td>
<td>0.84</td>
<td>0.79</td>
<td>0.84</td>
<td>0.85</td>
<td>0.90</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>6</td>
<td>0.92</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
<td>0.89</td>
<td>0.87</td>
<td>0.83</td>
<td>0.87</td>
<td>0.89</td>
<td>0.92</td>
<td>0.91</td>
<td>0.93</td>
</tr>
<tr>
<td>7</td>
<td>0.94</td>
<td>0.93</td>
<td>0.94</td>
<td>0.93</td>
<td>0.90</td>
<td>0.86</td>
<td>0.89</td>
<td>0.90</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>0.91</td>
<td>0.88</td>
<td>0.91</td>
<td>0.92</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>9</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>0.90</td>
<td>0.92</td>
<td>0.93</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>10</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td>0.94</td>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>11</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>12</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>13</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>14</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>15</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>16</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>17</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>18</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>19</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>20</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>21</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>22</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>23</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>24</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>25</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>26</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>27</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>28</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>29</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>30</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

2.3 EOFs AND EIGENVALUES FOR USHCN PRECIPITATION DATA

The computation of the EOFs and eigenvalues of the precipitation works exactly the same. The precipitation values are area-weighted, the covariance matrix is computed and then the eigenvectors and eigenvalues of this covariance matrix are found. Figures 2.19 - 2.21 show the first three modes of the precipitation data of January. Figures 2.22 - 2.24 show the first three modes June.
Figure 2.19. EOF1 of precipitation data of January

Figure 2.20. EOF2 of precipitation data of January
Figure 2.21. EOF3 of precipitation data of January

Figure 2.22. EOF1 of precipitation data of June
Figure 2.23. EOF2 of precipitation data of June

Figure 2.24. EOF3 of precipitation data of June

The following Figure, number 2.25, shows the eigenvalues of the precipitation data plotted versus the EOF mode. In Figure 2.26 the cumulative percentage of the eigenvalues was entered on the y-axis. We notice, that compared to the equivalent plots of the temperature
data, more modes are needed here to reach the 90% of the variance explained. This is because in the precipitation data there is a lot more variance than in the temperature data.

Figure 2.25. Eigenvalues versus Mode Number by Month (Precipitation Data)
Figure 2.26. Cumulative Percentage of the Eigenvalues versus Mode Number (Precipitation Data)
CHAPTER 3

METHOD OF RECONSTRUCTION

As described in the introduction, the TOB dataset of the USHCN includes weather records since 1895. But there is also a lot of missing data, i.e. for some grid boxes there are no temperature data available for certain months and years. Figure 3.1 shows how the number of grid boxes with temperature data changes over time. As mentioned earlier for the precipitation data only the interpolated F52 dataset is available. Therefore all grid boxes are filled at all times.

As we can see, the number of grid boxes with data is increasing from 104 in January 1895 (not considering the fluctuations) until in March 1918 for the first time all 120 grid boxes contained data. The minimum number of 103 grid boxes with data was reached from April-June 1895. For around 90 years almost all grid boxes had temperature data. Lately, a lot of stations were closed, which leads to a decrease of the number of grid boxes with data since
Now we can have a look at the spatial distribution of the grid boxes with missing values for a certain time point. Figure 3.2 shows the temperature anomalies of the average temperature of January 1897. A total of twelve grid boxes are blank in this plot because of missing data for these grid boxes.

![Temperature anomalies of Tavg in January 1897](image)

**Figure 3.2. Temperature anomalies of Tavg in January 1897**

The purpose of the so-called reconstruction method, which is explained in the following section, is first to estimate the temperature data where it is missing. Secondly, we want to estimate the error that this method produces in order to compare different methods which are used in climate research to model weather data. The error computation will be discussed in further detail in the ensuing chapter.

### 3.1 Multivariate regression reconstruction method

First we want to recapitulate some linear multivariate regression theory. The basic model looks like the following

\[ Y = X\beta + \epsilon \]  

(3.1)
where \( Y \) is the \( n \)-dimensional vector of dependent variables, \( X \in \mathbb{R}^{n \times m} \) is called the design matrix and contains \( n \) row vectors - \( m \) observations each - of the independent variables. We want to relate \( X \) to \( Y \) by estimating \( \beta \) which is the \( m \)-dimensional parameter vector. The \( n \)-dimensional error vector \( \epsilon \) is assumed to be normally distributed.

\[
\epsilon \sim N(0, \sigma^2 I)
\]  

(3.2)

We use the least squares method to minimize the sum of squared errors regarding \( \beta \):

\[
\sum_{i=1}^{n}(Y_i - (X\beta)_i)^2 \rightarrow \min_{\beta \in \mathbb{R}^m}
\]

(3.3)

Assuming \( X'X \) is invertible, the resulting least squares estimator for \( \beta \) has the following form

\[
\hat{\beta} = (X'X)^{-1}X'Y
\]

(3.4)

Now we want to make use of the linear regression model to generate the temperature and precipitation data of the U.S. again using EOFs. Hence, the design matrix is represented by the matrix of EOFs. We use a slightly different notation and denote the design matrix by \( E_k \) where the index \( k \) stands for the month since the EOFs differ for each month.

Consequently, this implies that the design matrix varies with the according month we want to reconstruct. Each column of \( E_k \) contains one EOF mode. Furthermore, \( E_k(x, m) \) denotes the \( x \)-th element of the \( m \)-th column vector of \( E_k \). The range of \( x \) is \( \{1, \ldots, N\} \) for instance, where \( N \) denotes the number of grid boxes. The second dimension \( m \) of \( E_k \) ranges from 1 to \( M \) which is the maximum number of modes. In practice, we use only as many modes as are necessary to represent 90% of the variance of the data used for the EOF computation. This number depends on the month and is therefore denoted by \( M_k \).

Let \( D_t(x) \) be the temperature data for a given month in a specific year (time point \( t \)) and a certain grid box \( x \). Moreover \( A(x) \) denotes the weight for grid box \( x \). See formula 2.9 in Chapter 2 for the definition of the area weight.

Then we define:

\[
R_t(x) = D_t(x)\sqrt{A(x)}
\]

(3.5)

Thus, we have one data vector \( R_t \) which contains the area weighted temperature data for each grid box for a certain month in a specific year. Therefore, the \( N \)-dimensional (ideal case of no missing values) vector \( R_t \) depends on the location and the time. We can now set up the adapted regression model

\[
R_t(x) = E_k \beta_t + \epsilon
\]

(3.6)
where \( \beta_t \in \mathbb{R}^{M_k} \) is still the parameter vector, which contains the regression coefficients that are to be estimated. Furthermore, \( \epsilon \) is the reconstruction error, which is again assumed to be normally distributed with mean 0 and variance matrix \( \sigma^2 I \). Moreover, we can represent the reconstructed area-weighted temperature as

\[
\hat{R}_t(x) = E_k \hat{\beta}_t \tag{3.7}
\]

In the ideal case of no missing data, the dimensions of \( R, \hat{R} \) and \( \epsilon \) are \( N \times 1 \) where \( N \) is the number of grid boxes. We denote the reconstruction field by \( \Omega \). Hence, \( N = |\Omega| \). Since we have to deal with a lot of missing data the number of grid boxes with non-missing data for a given time point \( t \) will not be \( N \) in most cases. We denote the data domain at time point \( t \) by \( \Omega_t \) and \( N_t = |\Omega_t| \) which represents the dimension of \( R_t, \hat{R}_t \) and \( \epsilon_t \). We have to keep in mind that \( \Omega_t \) and therefore also \( N_t \) changes in time. If the data vector shrinks to dimensions \( N_t \times 1 \), we also have to trim \( E_k \) for the estimation of \( \beta_t \) to dimensions \( N_t \times M_k \) by deleting all the rows which correspond to missing values of certain grid boxes. We denote this trimmed design matrix by \( E_t \) since it depends not only on the month but also on the time point of the data to be estimated.

Applying the least squares linear regression formula that is shown in 3.4 for the general theory gives us an equation for the unbiased estimator - one that minimizes the sum of squared errors.

\[
\hat{\beta} = \min_{\beta \in \mathbb{R}^{M_k}} \sum_{x \in \Omega_t} (R_t(x) - (E_t \beta)(x))^2
\]

\[
= (E'_t E_t)^{-1} E'_t R_t \tag{3.9}
\]

\[
= G_t R_t \tag{3.10}
\]

We call \( G_t = (E'_t E_t)^{-1} E'_t \) the link matrix.

Note that \( E'_t E_t \) is invertible because its columns are linearly independent vectors.

Furthermore,

\[
\hat{R}_t(x) = E_k (E'_t E_t)^{-1} E'_t R_t \tag{3.11}
\]

is the final representation of the reconstructed temperature.

### 3.2 Application of the Reconstruction Method

The EOFs were calculated according to the method described in Chapter 2 using the F52 dataset. For the reconstruction we use these EOFs as the design matrix. In addition we use the TOB datasets of \( T_{min}, T_{max} \) and \( T_{avg} \) to built the data vector \( D_t(x) \). In appendix A.4
one can find the code to apply the described method to the USHCN datasets. Now we can have a look at the reconstructed temperature of January 1897 (Figure 3.3). If we compare this map with the map shown in Figure 3.2 we first notice, that now all grid boxes are filled. We estimated not only the temperature of the grid boxes that had data before but also the temperature of the ones with missing values. If we have a closer look, we notice that the reconstructed map seems somehow smoother than the original data, since nearby grid boxes have similar colors. Smoothing is a property of most modeling methods like linear regression, which the reconstruction method is based on. Nevertheless, we can recognize a pretty similar pattern in the original data and the reconstructed data. There is the cold east coast and also the west coast is pretty cool, whereas the midwest has high positive temperature anomalies. Only the area southwestern of the Great Lakes seems hotter in the original data than in the reconstructed data.

![Figure 3.3. Reconstructed Tavg of January 1897](image-url)
CHAPTER 4

AREA WEIGHTED AVERAGES AND ERROR ESTIMATION

In this chapter we will examine the annual mean temperature and precipitation over time. We will go over some results from Weithmann (2011) [12] and Wied (2012) [13] and compare the annual mean of the original TOB data with the F52 and the reconstructed data. Also, the estimation of the mean squared errors of the reconstruction method will be presented in the second part of this chapter.

4.1 AREA WEIGHTED AVERAGES

In order to look at the time series of the TOB data of $T_{avg}$, $T_{min}$ and $T_{max}$ we computed an area-weighted average (see 2.9 for definition of area weights) of all grid boxes and calculated the annual mean. Figure 4.1-4.3 show this yearly mean plotted versus time. Red bars stand for positive anomalies, while blue bars represent negative anomalies.

Figure 4.1. Annual Mean of $T_{avg}$ of TOB dataset
Figure 4.2. Annual Mean of $T_{min}$ of TOB dataset

Figure 4.3. Annual Mean of $T_{max}$ of TOB dataset
We can clearly see four distinct climate periods which have already been identified in
Shen et al. (2011) [9]. There is a colder period from 1895-1930 followed by a warmer one in
1931-1960. Subsequently, the years 1961-1985 have mainly negative temperature anomalies
and the most recent years from 1986-2012 are again warmer than the average from the base
period that we refer to. Weithmann (2011) [12] conducted a series of Kolmogorov-Smirnov
tests to investigate if the four periods, we identified by looking at the time series, differ
significantly. It turned out that all p-values are very small. Hence, he rejected all null
hypotheses of a common underlying distribution of the temperature time series at a
significance level of 5%. This means that we see significant differences not only if we
compare the two colder with the two warmer periods, but also if we compare the two colder
periods with each other and the two warmer ones with each other. The following table shows
the average anomaly values of $T_{min}$, $T_{avg}$ and $T_{max}$ for each of the four periods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{min}$</td>
<td>-0.21</td>
<td>0.17</td>
<td>-0.09</td>
<td>0.66</td>
</tr>
<tr>
<td>$T_{max}$</td>
<td>-0.14</td>
<td>0.33</td>
<td>-0.08</td>
<td>0.35</td>
</tr>
<tr>
<td>$T_{avg}$</td>
<td>-0.17</td>
<td>0.25</td>
<td>-0.09</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The ranking of the periods by temperature anomaly is obvious. If we have a closer
look we notice what Wied (2012) [13] already described. In the most recent period the
minimum temperature is higher than the maximum temperature. At first sight that doesn’t
seem right but the explanation for this phenomenon is the spatial distribution of the hot and
cold areas. The area-weighting of the grid boxes can then lead to the observed phenomenon.
Giving a bigger weight to areas with high minimal temperature (which by the way might be
even higher than the maximum temperature of other grid boxes) can result in a higher average
of the $T_{min}$ than $T_{max}$. In our case the western and southwestern parts of the U.S. contribute
the most to this phenomenon (see Wied (2012) [13]).

If we now look at the reconstructed temperature, we recognize the same pattern. The
four periods are clearly visible.
Figure 4.4. Annual Mean of $T_{avg}$ of reconstructed dataset

Figure 4.5. Annual Mean of $T_{min}$ of reconstructed dataset
But to investigate the differences of the three datasets (TOB, F52 and reconstructed temperature) we need to plot them in one graph so that we can recognize even small differences. Figure 4.7 shows the minimum temperature values for all three datasets. Figure 4.8 and 4.9 illustrate the same for $T_{max}$ and $T_{avg}$, respectively.
Figure 4.7. Annual Mean of $T_{min}$
Figure 4.8. Annual Mean of $T_{max}$
In general we can say that the deviations of the TOB data and the reconstructed data are very small. The data from the fully adjusted F52 dataset mostly differs more from the slightly adjusted TOB data than the reconstruction data which was computed in this thesis. The differences seem the highest for the maximum temperature and the lowest for the minimum temperature. For $T_{avg}$ the differences between F52 data and TOB data are also quite remarkable. We also notice that the deviations of the F52 data from the TOB data is not always in the same direction. For $T_{max}$ and $T_{avg}$ the F52 data underestimates the temperature during the first 100 years and in the more recent years it tends to overestimate the data. For $T_{min}$ we don’t really observe that pattern. Here, we can see lower F52 values in the years from 1920-1970. In the more recent years we can even say that the F52 values are rather lower than the TOB values. Throughout all $T_{min}$, $T_{max}$ and $T_{avg}$ the reconstructed annual means are very close to the TOB data values. We don’t observe any remarkable differences between the three data versions.
4.2 ERROR ESTIMATION

First we want to derive a formula to estimate the Mean Squared Error (MSE) of the reconstruction.

In theory we are looking at a continuous temperature field. Hence, the temperature \( R_t \) at time \( t \) and location \( x \) can be represented by an infinite linear combination of EOFs:

\[
R_t(x) = \sum_{m=1}^{\infty} \beta_m(t) E_k(x, m) \tag{4.1}
\]

Note that in theory there are infinitely many EOF modes. The reconstructed temperature \( \hat{R} \) at time \( t \) and location \( x \) can by written as a truncated linear combination

\[
\hat{R}_t(x) = \sum_{m=1}^{M} \hat{\beta}_m(t) E_k(x, m) \tag{4.2}
\]

The MSE is defined as the expected value of the squared difference of the estimator and the true value. We denote the MSE of the reconstruction by \( \epsilon^2 \) and according to the definition it is

\[
\epsilon^2(x) = < (\hat{R}_t(x) - R_t(x))^2 > \tag{4.3}
\]

Note that it depends on the location \( x \). In the following we want to find a way to express \( \epsilon^2 \) in terms of EOFs. For the derivation we need the following assumptions (see Shen et al. (2004) [8]):

\[
< (\beta_m - \hat{\beta}_m)(\beta_n - \hat{\beta}_n) > = 0 \quad \text{if} \quad m \neq n \tag{4.4}
\]
Using this assumption and the expansions introduced in 4.1 and 4.2 we can proceed as follows. Here, \( <.,.> \) denotes the ensemble average.

\[
\epsilon(x) = < (\hat{R}_t(x) - R_t(x))^2 > \\
= < R_t^2(x) - 2R_t(x)\hat{R}_t(x) + \hat{R}_t^2(x) > \\
= < R_t^2(x) > - < 2R_t(x)\hat{R}_t(x) > + < \hat{R}_t^2(x) > \\
= < \left( \sum_{m=1}^{\infty} \beta_m(t)E_k(x,m) \right)^2 > - \\
- 2 < \sum_{m=1}^{\infty} \beta_m(t)E_k(x,m) \sum_{m=1}^{M} \hat{\beta}_m(t)E_k(x,m) > + < \hat{R}_t^2(x) > \\
= < \left( \sum_{m=1}^{M} \beta_m(t)E_k(x,m) + \sum_{m=M+1}^{\infty} \beta_m(t)E_k(x,m) \right)^2 > - \\
- 2 < \left( \sum_{m=1}^{M} \beta_m(t)E_k(x,m) + \sum_{m=M+1}^{\infty} \beta_m(t)E_k(x,m) \right) \sum_{m=1}^{M} \hat{\beta}_m(t)E_k(x,m) > + \\
+ < \hat{R}_t^2(x) > \\
= < \left( \sum_{m=1}^{M} \beta_m(t)E_k(x,m) \right)^2 > - 2 \sum_{m=1}^{M} \beta_m(t)E_k(x,m) \sum_{m=M+1}^{\infty} \beta_m(t)E_k(x,m) \\
+ \left( \sum_{m=M+1}^{\infty} \beta_m(t)E_k(x,m) \right)^2 > - \\
- 2 < \left( \sum_{m=1}^{M} \beta_m(t)E_k(x,m) \right) \left( \sum_{m=1}^{M} \hat{\beta}_m(t)E_k(x,m) \right) > + \\
+ \left( \sum_{m=M+1}^{\infty} \beta_m(t)E_k(x,m) \right) \left( \sum_{m=1}^{M} \hat{\beta}_m(t)E_k(x,m) \right) > + < \hat{R}_t^2(x) >
The middle term of the first expected value as well as the second summand of the second expected value are 0 because of assumption 4.4.

\[
\epsilon^2 = \sum_{m=1}^{\mathcal{M}} <\beta_m^2(t) > E_k^2(x, m) + \sum_{m=\mathcal{M}+1}^{\infty} <\beta_m^2(t) > E_k^2(x, m) - \\
-2 \sum_{m=1}^{\mathcal{M}} \sum_{n=1}^{\mathcal{M}} <\beta_m(t) \hat{\beta}_n(t) > E_k(x, m)E_k(x, n) + \sum_{m=1}^{\mathcal{M}} <\hat{\beta}_m^2(t) > E_k^2(x, m) \\
= \sum_{m=1}^{\mathcal{M}} <\beta_m^2(t) > E_k^2(x, m) - 2 \sum_{m=1}^{\mathcal{M}} <\beta_m(t) \hat{\beta}_m(t) > E_k^2(x, m) + \\
+ \sum_{m=1}^{\mathcal{M}} <\hat{\beta}_m^2(t) > E_k^2(x, m) + \sum_{m=\mathcal{M}+1}^{\infty} <\beta_m^2(t) > E_k^2(x, m) \\
= \sum_{m=1}^{\mathcal{M}} <\beta_m^2(t) - 2\beta_m(t) \hat{\beta}_m(t) + \hat{\beta}_m^2(t) > E_k^2(x, m) + \\
+ \sum_{m=\mathcal{M}+1}^{\infty} <\beta_m^2(t) > E_k^2(x, m) \\
= \sum_{m=1}^{\mathcal{M}} <(\beta_m(t) - \hat{\beta}_m(t))^2 > E_k^2(x, m) + \sum_{m=\mathcal{M}+1}^{\infty} <\beta_m^2(t) > E_k^2(x, m) \tag{4.5}
\]

Now we can see that the MSE can be represented as a linear combination of the squared EOFs.

\[
\epsilon^2 = \sum_{m=1}^{\infty} \gamma_mE_k^2(x, m) \tag{4.6}
\]

where the coefficient

\[
\gamma_m = \begin{cases} 
< (\hat{\beta}_m(t) - \beta_m(t))^2 > , & m \leq \mathcal{M} \\
< \beta_m^2(t) > , & m > \mathcal{M} 
\end{cases} \tag{4.7}
\]

The second sum of the expression 4.5 is called truncation error and is never known [8]. Its size depends on the number of modes \(\mathcal{M}\) used which then again is determined by the EOFs or rather eigenvalues (see Chapter 3).

Now we want to use this representation of the MSE for in order to estimate the mean squared error of the reconstruction method. The discretization of the climate field yields the sum in formula 4.6 to be finite. The summation index runs until \(M_k\) which represents the number of EOFs necessary in month \(k\) to account for 90% of the variation of the data used for the EOF computation.

\[
\epsilon^2 = \sum_{m=1}^{M_k} \gamma_mE_k^2(x, m) \tag{4.8}
\]
According to Shen et al. (2004) [8] the coefficient $\gamma_m$ can be expressed by the following formula. The notation from Chapter 3 is used.

$$\gamma_m = \sum_{n=1}^{M_t} \lambda_n \left[ \delta_{mn} - \sum_{j=1}^{N_t} E_t(x, n) E_t(x, m) G_t(x, m) \right] + \sum_{j=1}^{N_t} e^2(E_t(x, m) G_t(x, m))^2 \quad (4.9)$$

where $G \in \mathbb{R}^{M_k \times N_t}$ is the link matrix of the linear regression and $e^2$ is the variance of the observed error for data $D_t(x)$. For the computations in this thesis, $e^2$ was set equal to 0.1 based on an empirical rule of thumb. Additional research has to be done to determine that value more precisely.

In what follows we want to present some results that we got from calculating the MSE for the reconstruction using the above formulas 4.8 and 4.9.

Figure 4.10 shows the square root of the MSE (averaged regarding the grid boxes) of the reconstruction of $T_{min}$ over time. We can see a separate line for each month. First of all we notice that the lines appear horizontal. Figure 4.11 shows each of these lines separately on a different scale. The upper left one shows the graph for January, the upmost one on the right side is February and so on. Now we can recognize slight fluctuations. If we have a look at the tick marks of the y-axis (which are in °C) we notice that these deviations are very small.

That’s why they don’t appear in the combined graph. The reason why the error doesn’t vary over time is, that the time is not directly related to the MSE. If we look back at the formulas used to compute the MSE we notice that the only parameter which is related to the time is the number of grid boxes with missing data. Assume there are no missing data throughout the entire considered time period, then the error would be constant in time.
Figure 4.10. Square Root of MSE of $T_{min}$ over time
Returning to Figure 4.10 we notice in addition, that there are quite big differences of the MSE comparing the different months. We want to know if certain grid boxes contribute more to this phenomenon than others. Figure 4.12 contains a line for each grid box. On the x-axis we can see the twelve months. The y-axis shows the square root of the MSE (averaged over all years) of $T_{min}$ as well. Obviously all grid boxes feature this "U"-shape - in varying strength. Apparently the MSE is lower in summer than in the winter months. The reason is that in summer the variation of the temperature data is lower than in winter. That affects the magnitude of the MSE.
Figure 4.12. Square Root of MSE of $T_{min}$ for each grid box
CHAPTER 5

CONCLUSION AND DISCUSSION

Empirical Orthogonal Functions are an important tool to describe a spatial pattern and its change in time. As a first step we showed in this thesis how to calculate them for an arbitrary data vector. It turned out that the EOFs are the eigenfunctions of the covariance matrix of the data vector. Therefore, the data can be represented as a linear combination of the EOFs, which we used for the method of reconstruction. Also the eigenvalues have a nice and useful interpretation. Their percentage of the total sum of all eigenvalues represents the proportion of the variance explained by the projection of the data into the subspace spanned by the according eigenvector. We also made use of that in the reconstruction. After actually having computed the EOFs for the temperature and precipitation data of USHCN and plotting them on a map of the U.S. we noticed, that the rough pattern that we observe for the first mode becomes more and more chaotic as the EOF mode increases. This is confirmed by examining the cumulative percentage of the eigenvalue of the total sum of all eigenvalues versus the EOF mode. Mostly, the first EOF mode accounts already for more than half of the variance. For the temperature data we reach 90% after around five EOFs already. Regarding the precipitation data we need some more modes - mostly more than 10.

The reconstruction method is based on multivariate linear regression techniques using the matrix of EOFs as the design matrix. This means, we were searching for a parameter vector which, multiplied by the EOF matrix, gives us an estimated value for the temperature. Thereby, we didn’t use all the EOF modes but only as many as necessary to explain 90% of the variance in the EOF data. The model that we got from that procedure can now be validated and then be used for prediction of missing and future data.

For validation, we need to look into the error which comes with the estimation. We found that the mean squared error oscillates with the months. During the winter months it is higher than during the summer months. Apart from that, there are just small fluctuations throughout the years which coincide with the number of grid boxes with missing data.

Further research can now be done to improve the error estimation and the reconstruction itself. The 90% rule of thumb to determine the number of EOF modes used for the reconstruction, can be questioned. A more data related rule might improve the quality of the estimation.
BIBLIOGRAPHY


APPENDIX

Codes
CODES

A.1 SAS Code for Data Preparation

The following SAS Code reads the longitude and the latitude of the four corners of all the 120 grid boxes. This file will be needed to do the data transformation.

libname path "PATH";

data path.gridpoints;

input gridbox max1 min1 max2 min2;
cards;

1.0000 84.0000 80.5000 25.0000 22.5000
2.0000 101.5000 98.0000 27.5000 25.0000
3.0000 84.0000 80.5000 27.5000 25.0000
4.0000 80.5000 77.0000 27.5000 25.0000
5.0000 101.5000 98.0000 30.0000 27.5000
6.0000 98.0000 94.5000 30.0000 27.5000
7.0000 94.5000 91.0000 30.0000 27.5000
8.0000 91.0000 87.5000 30.0000 27.5000
9.0000 87.5000 84.0000 30.0000 27.5000
10.0000 84.0000 80.5000 30.0000 27.5000
11.0000 115.5000 112.0000 32.5000 30.0000
12.0000 112.0000 108.5000 32.5000 30.0000
13.0000 108.5000 105.0000 32.5000 30.0000
14.0000 105.0000 101.5000 32.5000 30.0000
15.0000 101.5000 98.0000 32.5000 30.0000
16.0000 98.0000 94.5000 32.5000 30.0000
17.0000 94.5000 91.0000 32.5000 30.0000
18.0000 91.0000 87.5000 32.5000 30.0000
19.0000 87.5000 84.0000 32.5000 30.0000
20.0000 84.0000 80.5000 32.5000 30.0000
21.0000 122.5000 119.0000 35.0000 32.5000
22.0000 119.0000 115.5000 35.0000 32.5000
23.0000 115.5000 112.0000 35.0000 32.5000
24.0000 112.0000 108.5000 35.0000 32.5000
25.0000 108.5000 105.0000 35.0000 32.5000
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>105</td>
<td>101.5</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>101.5</td>
<td>98</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>98</td>
<td>94.5</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>94.5</td>
<td>91</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>91</td>
<td>87.5</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>87.5</td>
<td>84</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>84</td>
<td>80.5</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>80.5</td>
<td>77</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>77</td>
<td>73.5</td>
<td>35</td>
<td>32.5</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>122.5</td>
<td>119</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>119</td>
<td>115.5</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>115.5</td>
<td>112</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>112</td>
<td>108.5</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>108.5</td>
<td>105</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>105</td>
<td>101.5</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>101.5</td>
<td>98</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>98</td>
<td>94.5</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>94.5</td>
<td>91</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>91</td>
<td>87.5</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>87.5</td>
<td>84</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>84</td>
<td>80.5</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>80.5</td>
<td>77</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>77</td>
<td>73.5</td>
<td>37.5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>126</td>
<td>122.5</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>122.5</td>
<td>119</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>119</td>
<td>115.5</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>115.5</td>
<td>112</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>112</td>
<td>108.5</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>108.5</td>
<td>105</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>105</td>
<td>101.5</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>101.5</td>
<td>98</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>98</td>
<td>94.5</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>94.5</td>
<td>91</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>91</td>
<td>87.5</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>87.5</td>
<td>84</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>84</td>
<td>80.5</td>
<td>40</td>
<td>37.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>62.0000</td>
<td>80.5000</td>
<td>77.0000</td>
<td>40.0000</td>
<td>37.5000</td>
<td></td>
</tr>
<tr>
<td>63.0000</td>
<td>77.0000</td>
<td>73.5000</td>
<td>40.0000</td>
<td>37.5000</td>
<td></td>
</tr>
<tr>
<td>64.0000</td>
<td>126.0000</td>
<td>122.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>65.0000</td>
<td>122.5000</td>
<td>119.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>66.0000</td>
<td>119.0000</td>
<td>115.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>67.0000</td>
<td>115.5000</td>
<td>112.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>68.0000</td>
<td>112.0000</td>
<td>108.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>69.0000</td>
<td>108.5000</td>
<td>105.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>70.0000</td>
<td>105.0000</td>
<td>101.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>71.0000</td>
<td>101.5000</td>
<td>98.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>72.0000</td>
<td>98.0000</td>
<td>94.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>73.0000</td>
<td>94.5000</td>
<td>91.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>74.0000</td>
<td>91.0000</td>
<td>87.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>75.0000</td>
<td>87.5000</td>
<td>84.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>76.0000</td>
<td>84.0000</td>
<td>80.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>77.0000</td>
<td>80.5000</td>
<td>77.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>78.0000</td>
<td>77.0000</td>
<td>73.5000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>79.0000</td>
<td>73.5000</td>
<td>70.0000</td>
<td>42.5000</td>
<td>40.0000</td>
<td></td>
</tr>
<tr>
<td>80.0000</td>
<td>126.0000</td>
<td>122.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>81.0000</td>
<td>122.5000</td>
<td>119.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>82.0000</td>
<td>119.0000</td>
<td>115.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>83.0000</td>
<td>115.5000</td>
<td>112.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>84.0000</td>
<td>112.0000</td>
<td>108.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>85.0000</td>
<td>108.5000</td>
<td>105.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>86.0000</td>
<td>105.0000</td>
<td>101.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>87.0000</td>
<td>101.5000</td>
<td>98.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>88.0000</td>
<td>98.0000</td>
<td>94.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>89.0000</td>
<td>94.5000</td>
<td>91.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>90.0000</td>
<td>91.0000</td>
<td>87.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>91.0000</td>
<td>87.5000</td>
<td>84.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>92.0000</td>
<td>84.0000</td>
<td>80.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>93.0000</td>
<td>80.5000</td>
<td>77.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>94.0000</td>
<td>77.0000</td>
<td>73.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>95.0000</td>
<td>73.5000</td>
<td>70.0000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>96.0000</td>
<td>70.0000</td>
<td>66.5000</td>
<td>45.0000</td>
<td>42.5000</td>
<td></td>
</tr>
<tr>
<td>97.0000</td>
<td>126.0000</td>
<td>122.5000</td>
<td>47.5000</td>
<td>45.0000</td>
<td></td>
</tr>
</tbody>
</table>
The following SAS Code reads the data file which can be downloaded from the USHCN website, transforms the data and outputs a .txt file which can be read in Matlab in order to compute the EOFs and the reconstruction. It also needs the file stations.txt which can also be downloaded from the USHCN website. The path and filename of the dataset and the station file need to be customized as well as the path of the output file.

This program does the transformation of the temperature datasets. If the precipitation data is supposed to be transformed one can use the same program just the conversion of the units needs to be changed. The code is included but commented out.

```
DM 'CLEAR LOG; CLEAR OUTPUT;';

Libname path "PATH";
```
*read data USHCN data;
Data avg_in;
length no $11;
Infile "PATH\9641C_201301_tob_max.txt";
Input no $ m1 $ m2 $ m3 $ m4 $ m5 $ m6 $ m7 $ m8 $ m9 $ m10 $ m11 $ m12 $ avg $;
Run;

*delete letters, convert to numerical;
Data avg_prep (Keep=statno year month1-month12);
Set avg_in;
statno=substr(no,1,6)*1;
year=substr(no,8,4)*1;
Array mon m1-m12;
Array mon2 month1-month12;
do i=1 to dim(mon);
   If substr(mon(i),length(mon(i)),1) not in
       ("0","1","2","3","4","5","6","7","8","9")
   then substr(mon(i),length(mon(i)),1)="";
   mon2(i)=1*mon(i);
end;
run;

*transpose data;
proc transpose data=avg_prep out=prep;
var month1-month12;
by statno year;
run;

data prep2 (drop=_NAME_ where=(year<2013));
set prep;
month=substr(_NAME_,6,2)*1;
If COL1=-9999 then delete;
run;

*convert to Celsius;
data prep3;
set prep2;
temp=(((COL1/10)-32)*5/9);
drop COL1;
run;

/*FOR PRECIPITATION DATA: convert to ml;
data prep3;
set prep2;
temp=(COL1/100)*16.387064;
drop COL1;
run;
*/

*select base period;
Proc sql;
create table base1 as
select * from prep3
where year>1960 and year<1991;
run;
quit;

*calculate mean per station per month;
proc sql;
create table base2 as
select distinct statno, month,mean(temp) as mean_temp from base1
group by statno, month;
quit;

*merge mean with station data;
proc sql;
create table base3 as
select a.*, b.mean_temp from prep3 as a left join base2 as b
on a.statno=b.statno and a.month=b.month;
quit;
*calculate anomalies;
proc sql;
create table base4 as
select *, temp-mean_temp as anomaly from base3;
quit;

*read information about stations;
data stations;
infile "path\stations.txt";
input statno LATITUDE LONGITUDE;
longitude=longitude*(-1);
run;

*read information about grid;
data gridpoints;
set path.gridpoints;
run;

*assign station to gridbox;
proc sql;
create table comb1 as
select s.*, gridbox from stations s left join gridpoints g
on s.latitude>=min2 and latitude<max2 and
longitude>=min1 and longitude<max1
order by statno;
quit;

*combine temperature data with gridbox of station;
proc sql;
create table comb2 as
select a.*, b.gridbox from base4 as a, comb1 as b
where a.statno=b.statno;
run;
quit;

*average temperature anomalies within gridboxes;
proc sql;
create table comb3 as
select *, mean(anomaly) as avg_gridbox from comb2
group by gridbox, year, month;
quit;

*create full dataset;
data full;
do i=1 to 120;
do j=1895 to 2012;
do k=1 to 12;
gridbox=i;
year=j;
month=k;
output;
end;
end;
end;
drop i j k;
run;

*merge full dataset with anomalies;
proc sql;
create table full_temp as
select distinct a.*, b.avg_gridbox
from full as a left join comb3 as b
on a.gridbox=b.gridbox and a.year=b.year and a.month=b.month;
quit;

/*check
proc sql;
select * from full_temp
where avg_gridbox=. and year>1960 and year<1991;
quit;
*/
The following two programs are needed to read the file, that was produced with the SAS code in appendix A.1, in Matlab.

This Matlab function reshapeData.m is called by the readData script and needs a three input parameters. The dataset, the number of months of the considered time period and a flag which indicats if the EOF data is reshaped or the TOB data for the reconstruction.

```matlab
function[EOFdata]=reshapeData(EOFdataIn, NoMonths,flag);
EOFhelp=zeros(120,NoMonths);
for i=1:120
    EOFhelp(i,:)=EOFdataIn(NoMonths*(i-1)+1:NoMonths*i);
end;
```
if flag==1
    EOFdata=zeros(120,NoMonths);
    EOFdata=EOFhelp;
else
    EOFdata=zeros(120, 30*12);
    EOFdata=EOFhelp(:,65*12+1:95*12);
end;

The Matlab script 'readData.m' can just be called without any parameter as long as the files, that need to be read, are located in the right folder.

%read in prepared F52 data in order to compute EOFs
EOFdata=dlmread('data\temp_F52_avg.txt');
EOFavgData=reshapeData(EOFdata,NoMonths,0);
clear EOFdata;

EOFdata=dlmread('data\temp_F52_min.txt');
EOFminData=reshapeData(EOFdata,NoMonths,0);
clear EOFdata;

EOFdata=dlmread('data\temp_F52_max.txt');
EOFmaxData=reshapeData(EOFdata,NoMonths,0);
clear EOFdata;

EOFdata=dlmread('data\pcp_F52_avg.txt');
EOFpcpData=reshapeData(EOFdata,NoMonths,0);
clear EOFdata;

%read in prepared F52 data for whole period of time
EOFdata=dlmread('data\temp_F52_avg.txt');
F52avg=reshapeData(EOFdata,NoMonths,1);
clear EOFdata;

EOFdata=dlmread('data\temp_F52_min.txt');
F52min=reshapeData(EOFdata,NoMonths,1);
clear EOFdata;
EOFdata=dlmread('data\temp_F52_max.txt');
F52max=reshapeData(EOFdata,NoMonths,1);
clear EOFdata;

EOFdata=dlmread('data\pcp_F52_avg.txt');
F52pcp=reshapeData(EOFdata,NoMonths,1);
clear EOFdata;

%read in tob data for reconstruction;
data=dlmread('data\temp_tob_avg.txt');
TOBavg=reshapeData(data,NoMonths,1);
clear data;

data=dlmread('data\temp_tob_min.txt');
TOBmin=reshapeData(data,NoMonths,1);
clear data;

data=dlmread('data\temp_tob_max.txt');
TOBmax=reshapeData(data,NoMonths,1);
clear data;

### A.3 MATLAB Code: Compute EOFs

The following Matlab function `computeEOF.m` computes the EOFs and eigenvalues for all months for one dataset. The input parameter is the dataset and the output are two matrices one with the EOFs and one with the eigenvalues.

```matlab
function [EOF EV] = computeEOF(EOFdata)

%Base Period
basePeriod = 1961:1990;

%Number of years
M = length(basePeriod);

%isolate data for specific month
P=zeros(120,30,12);
for i = 1:12
```

P(:,:,i)=EOFdata(:,i:12:(M-1)*12+i);
end

% calculate AreaWeight and save it in diagonal matrix
W=diag(getgridarea);

EOF=nan(120,30,12);
EV=nan(30,12);
ver=nan(12,1);

% do all the computations for the 12 months
for j=1:12

% multiply by AreaWeight
Phat=W*P(:,:,j);

% Covariance in time
C = Phat'*Phat/M;

% Eigenvalues and Eigenvector of C
[U, D] = eig(C);

% EOF in space
V = Phat*U;

% Normalize EOF
VN = zeros(120, M);
for i = 1:M
    VN(:,i) = (V(:,i)./diag(W))/norm(V(:,i),2);
end

% permutation matrix
Permute = zeros(M, M);
for k = 1:M
    Permute(k,M-k+1) = 1;
end
EOFper = VN*Permute;
% make EOFs unique by making the first entry of each EOF positive
EOFsign=nan(120,30);
for i=1:30
    if EOFper(1,i)<0
        EOFsign(:,i)=-1.*EOFper(:,i);
    else EOFsign(:,i)=EOFper(:,i);
    end;
end;

%Permute eigenvalues
EigVals = zeros(1,M);
for i = 1:M
    EigVals(i) = D(M+1-i, M+1-i);
end

EOF(:,:,j)=EOFsign;
EV(:,j)=EigVals;

%Eigenvalue Verification
ver(j) = 1/30*sum(sum((Phat).^2));
end

%verification table
verification=nan(12,4);
%first column month indices
verification(:,1)=1:12;
verification(:,2)=sum(EV,1);
verification(:,3)=ver;
verification(:,4)=verification(:,2)-verification(:,3);

%save verification table
filename=strcat('figures/EigValVerification.xls');
xlswrite(filename,verification,inputname(1))
A.4 Matlab Code: Run Reconstruction

The following Matlab function 'reconstruction.m' performs the reconstruction. It needs the TOB data, the according EOFs and eigenvalues. In addition one has to indicate the year and month that is supposed to be reconstructed. The output contains the reconstructed temperature and the estimated mean squared error.

```matlab
function [Rhat, eps] = reconstruction(data, EOF, EV, year, month);

% data vector
vec = data(:, (year - 1895) * 12 + month);

% figure out where the missing values are
locnan = isnan(vec);

% figure out number of gridboxes with data
Nd = 120 - sum(locnan);

% calculate AreaWeights
a = getgridarea();

% multiply data with area weights
Y = a .* vec;

% compute EOFs
[EOF, EV] = computeEOF(EOFdata);

% number of modes needed for error estimation
% enough if cumulative percentage of Eigenvalues >= 0.9
id = find((cumsum(EV(:, month) / sum(EV(:, month))) >= 0.9) == 1);
NoModes = id(1);

% trim EOF (only m modes needed)
EOFtrim = EOF(:, 1:NoModes, month);

% delete rows of EOF where missing data
Zd = zeros(Nd, NoModes);
Yd = zeros(Nd, 1);
```
z=1;
for i=1:120
    if (locnan(i)==0)
        Zd(z,:)=EOFtrim(i,:);
        Yd(z)=Y(i);
        z=z+1;
    end
end

%regression
R=EOFtrim*inv(Zd'*Zd)*Zd'*Yd;

%divide by AreaWeight
Rhat=diag(1./a)*R;

%compute link matrix
G=inv(Zd'*Zd)*EOFtrim';
%G=inv(Zd'*Zd)*Zd';

%delete rows of EOF where missing data
Rtrim=zeros(Nd,1);
Gd=zeros(NoModes,Nd);
z=1;
for i=1:120
    if (locnan(i)==0)
        Rtrim(z)=Rhat(i);
        Gd(:,z)=G(:,i);
        z=z+1;
    end
end

%calculate regression variance
regvar=norm(Yd-Rtrim,2)/(Nd-NoModes-1);

I=eye(Nd);
%error estimation
eps_sumn=zeros(NoModes,1);
eps_summ=zeros(120,NoModes);
epsm=zeros(NoModes,1);
for m=1:NoModes
    for n=1:NoModes
        eps_sumn(n)=EV(n,month)*
        *(I(m,n)-sum(Zd(:,n).*Zd(:,m).*Gd(m,:')))^2+
        +(0.1*sum((Zd(:,m).*Gd(m,:')).^2));
    end;
    epsm(m)=sum(eps_sumn);
    eps_summ(:,m)=sum(eps_sumn)*(EOFtrim(:,m)).^2;
end;
eps=sum(eps_summ,2);