Dimension Reduced Modeling of Spatio-Temporal Processes with Applications to Statistical Downscaling

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Jenný Brynjarsdóttir, B.S., M.S., M.S.

Graduate Program in Statistics

The Ohio State University

2011

Dissertation Committee:

L. Mark Berliner, Advisor
Peter F. Craigmile
Steven N. MacEachern
Abstract

The field of spatial and spatio-temporal statistics is increasingly faced with the challenge of very large datasets. Examples include data obtained from remote sensing satellites, global weather stations, outputs from climate models and medical imagery. The classical approach to spatial and spatio-temporal modeling is extremely computationally expensive when the datasets are large. Dimension-reduced modeling approach has proved to be effective in such situations. In this thesis I focus on the problem of modeling two spatio-temporal processes where the primary goal is to predict one process from the other and where the datasets for both processes are large.

I outline a general dimension-reduced Bayesian hierarchical approach to modeling of two spatio-temporal processes. The spatial structures of both processes are modeled in terms of a low number of basis vectors, hence reducing the spatial dimension of the problem. The temporal evolution of the spatio-temporal processes is then modeled through the coefficients (i.e. amplitudes) of the basis functions. I demonstrate that known multivariate statistical methods, Maximum Covariance Analysis (MCA) and Canonical Correlation Analysis (CCA), can be used to obtain basis vectors for dimension-reduced modeling of two spatio-temporal processes. Furthermore, I present a new method of obtaining data-dependent basis vectors that is geared to the goal of predicting one process from the other. The new basis vectors are called
Maximum Covariance Patterns (MCPs) and an orthogonal version is called Orthogonal Maximum Covariance Patterns (OMCPs). I apply these methods to a statistical downscaling example, where surface temperatures on a coarse grid over the Antarctic are downscaled onto a finer grid.
Dedicated to Elli, Anna Laufey and Tómas Ingi
Acknowledgments

First and foremost I thank my advisor, Dr. Mark Berliner. I feel very fortunate to have enjoyed his guidance and mentoring during my journey at Ohio State. Mark is the best advisor I could have asked for. I thank him for his encouragement, patience, generosity and good humor. I thank him for introducing me to the world of hierarchical Bayesian modeling and sharing with me his brilliant insights and vision.

I also thank Dr. Cressie, who played a big role in my studies. Dr. Cressie generously offered good advice and pointed to new areas, some of which became part of this thesis. In fact, his paper (Wikle and Cressie, 1999) ignited the spark that ultimately lead to this thesis. Furthermore, I very much enjoyed being a part of the SSES program.

I thank Dr. Calder for her generous advise and support. I have enjoyed her excellent teachings in many classes and our interactions through journal clubs and conferences have been inspiring. I also thank my committee members, Dr. Craigmile and Dr. MacEachern, for their valued input to my research.

I am grateful for all the good teachers in the Department of Statistics. The journey to a PhD is not without challenges and the excellent faculty of the Department are very supportive. In particular, I thank Dr. Jason Hsu for convincing me to come to Ohio State and for his support, especially during my first years in the program.
Also, a big shout-out to my fellow students Dr. Candace Berrett, the best conference buddy ever, and Dr. Josh Svenson, my favorite office mate.

I could not have done this without my family. To my wonderful parents, Brynjar and Anna: Thank you for your never-ending love and support. To my loving husband, Elli: Thank you for always believing in me. To my children Anna Laufey and Tómas Ingi: Thank you for all that you are.
Vita

1999 ......................... B.S. Mathematics, University of Iceland

2001-2002 .................... Research Associate, Marine Research Institute, Reykjavik, Iceland

2002 ......................... M.S. Industrial Engineering, University of Iceland

2002-2003 .................... Data Analyst, Statistics Department, deCODE genetics, Reykjavik, Iceland

2003-2005 .................... Research Associate, Science Institute, University of Iceland

2006-2008, 2010-2011 ....... Graduate Teaching Associate, Department of Statistics, The Ohio State University

Summer 2007 .................. Graduate Research Associate, Center of Biostatistics, The Ohio State University Medical Center

Summers 2008, 2010 .......... Graduate Teaching Associate, Mathematical Biosciences Institute, The Ohio State University

2008 ......................... M.S. Statistics, The Ohio State University

2008-2010, summer 2011 ...... Graduate Research Associate, Department of Statistics, The Ohio State University

Summer 2010 .................. Graduate Research Associate, Department of Geography, The Ohio State University

2010 ......................... Awarded the D. Ransom & Marian Whitney Award for outstanding effort as a Research Associate, Department of Statistics, The Ohio State University

2011 ......................... Awarded the D. Ransom & Marian Whitney Award for outstanding thesis researcher, Department of Statistics, The Ohio State University
Publications


Fields of Study

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiv</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Dimension reduced spatio-temporal modeling</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Dimension reduced representation</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1 Joint dimension reduction of two random vectors</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1.1 Canonical Correlation Analysis</td>
<td>6</td>
</tr>
<tr>
<td>1.2.1.2 Maximum Covariance Analysis</td>
<td>8</td>
</tr>
<tr>
<td>1.2.1.3 Some similarities and differences between CCA and MCA</td>
<td>9</td>
</tr>
<tr>
<td>1.2.2 Maximum/minimum autocorrelation factors</td>
<td>10</td>
</tr>
<tr>
<td>1.2.3 Principal Oscillation Patterns</td>
<td>12</td>
</tr>
<tr>
<td>1.3 Statistical Downscaling</td>
<td>15</td>
</tr>
<tr>
<td>2. A Hierarchical and Dimension Reduced Modeling Framework for Two Spatio-Temporal Processes</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Two spatial processes</td>
<td>17</td>
</tr>
<tr>
<td>2.1.1 Extension to more than two spatial processes</td>
<td>23</td>
</tr>
</tbody>
</table>


2.1.2 A Bayesian hierarchical model for two dimension reduced spatial processes ........................................... 24
2.1.3 Mean Square Error of the dimension reduced approach ................................................................. 25
2.2 Two spatio-temporal processes .......................................................... 28
  2.2.1 A Bayesian hierarchical model for two dimension reduced spatio-temporal processes ....................... 34
2.3 Two spatio-temporal processes misaligned in time ................................................................. 36
2.4 Spatio-temporal processes with continuous spatial domains .......................................................... 38
  2.4.1 Extension to two space-time processes ................................................................. 40
2.5 Bayesian Filtering .............................................................. 45
  2.5.1 Kalman Filter .............................................................. 47

3. Data-Dependent Basis Vectors ....................................................... 51
  3.1 Empirical Orthogonal Functions ................................................................. 51
  3.2 Maximum Covariance Analysis ................................................................. 55
  3.3 Canonical Correlation Analysis ................................................................. 57
  3.4 Maximum Covariance Patterns ................................................................. 59
    3.4.1 Proportion of total variance explained ................................................................. 63
    3.4.2 Covariance of EOF and MCP amplitudes across pairs ................................................................. 66
    3.4.3 MCP for more than two spatio-temporal processes ................................................................. 67
  3.5 Orthogonal Maximum Covariance Patterns ................................................................. 68
    3.5.1 Cost of orthogonality of OMCPs ................................................................. 71

4. Statistical Downscaling of Temperatures over the Antarctic ................................................................. 73
  4.1 The Polar MM5 and ERA-40 projects ................................................................. 74
  4.2 Exploratory data analysis of Polar MM5 and ERA-40 data ................................................................. 76
  4.3 Centered data, OMCPs and EOF patterns ................................................................. 80
  4.4 Exploratory analysis on estimated amplitudes ................................................................. 90
  4.5 Bayesian hierarchical model ................................................................. 101
    4.5.1 Data model ................................................................. 101
    4.5.2 Process model ................................................................. 102
    4.5.3 Parameter model ................................................................. 105
      4.5.3.1 Hyperparameters ................................................................. 111
  4.6 Predicting Polar MM5 temperatures with and without new ERA-40 data ................................................................. 115
    4.6.1 Forecasting without new ERA-40 data ................................................................. 115
    4.6.2 Downscaling ERA-40 data ................................................................. 118
  4.7 Results ................................................................. 120
    4.7.1 Forecasting the next year, without new ERA-40 data ................................................................. 129
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7.2</td>
<td>Statistical downscaling of ERA-40 data onto the Polar MM5 grid 2001-2002</td>
<td>141</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Statistical downscaling of ERA-40 data onto the Polar MM5 grid 2000-2001</td>
<td>147</td>
</tr>
<tr>
<td>5.</td>
<td>Discussion</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>Appendices</td>
<td>156</td>
</tr>
<tr>
<td>A.</td>
<td>Alternative Proofs and Useful Theorems</td>
<td>156</td>
</tr>
<tr>
<td>A.1</td>
<td>Some useful theorems</td>
<td>156</td>
</tr>
<tr>
<td>A.1.1</td>
<td>Normal-Normal updates</td>
<td>156</td>
</tr>
<tr>
<td>A.1.2</td>
<td>Kalman Filter as Bayesian Filtering</td>
<td>160</td>
</tr>
<tr>
<td>A.1.3</td>
<td>Singular Value Decomposition</td>
<td>164</td>
</tr>
<tr>
<td>A.1.4</td>
<td>Density of the square-root of an IG random variable</td>
<td>165</td>
</tr>
<tr>
<td>A.2</td>
<td>Alternative proofs for MCPs and OMCPs</td>
<td>165</td>
</tr>
<tr>
<td>A.2.1</td>
<td>Alternative proof of Theorem 3.4.1</td>
<td>165</td>
</tr>
<tr>
<td>A.2.2</td>
<td>Alternative proof of Theorem 3.5.1</td>
<td>166</td>
</tr>
<tr>
<td>B.</td>
<td>Exploratory Data Analysis</td>
<td>169</td>
</tr>
<tr>
<td>B.1</td>
<td>Polar MM5 and ERA-40 2-meter temperature fields</td>
<td>169</td>
</tr>
<tr>
<td>B.2</td>
<td>Centered Polar MM5 and ERA-40 2-meter temperature fields</td>
<td>177</td>
</tr>
<tr>
<td>B.3</td>
<td>Exploratory analysis of estimated EOF and OMCP amplitudes</td>
<td>184</td>
</tr>
<tr>
<td>B.3.1</td>
<td>Plots of $a_{l,t}$ vs. $b_{m,t}$</td>
<td>185</td>
</tr>
<tr>
<td>B.3.2</td>
<td>Plots of $b_{m,t}$ vs. $b_{m-1,t}$</td>
<td>196</td>
</tr>
<tr>
<td>B.4</td>
<td>EOFs and OMCPs for data through fall of 2001</td>
<td>207</td>
</tr>
<tr>
<td>C.</td>
<td>Gibbs Sampler</td>
<td>211</td>
</tr>
<tr>
<td>C.1</td>
<td>Full conditional distributions</td>
<td>212</td>
</tr>
<tr>
<td>C.2</td>
<td>Full conditional distributions for predictive model</td>
<td>219</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>221</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Proportion of total sample variance explained by the first four EOFs for each season and the first four OMCPs for each month.</td>
<td>91</td>
</tr>
<tr>
<td>4.2</td>
<td>Results from regressing estimated amplitudes $\hat{a}<em>{l,t,k}$ on $\hat{b}</em>{m1,t,k}$, $\hat{b}<em>{m2,t,k}$ and $\hat{b}</em>{m3,t,k}$ for each season $l$ and group $k$. Columns 2 through 5 show the estimated regression coefficients and stars that indicate whether the coefficient is significantly different from zero. The last column shows the $R^2$ when only the last month is included in the regression.</td>
<td>95</td>
</tr>
<tr>
<td>4.3</td>
<td>Correlation coefficients between residuals from the regression of $\hat{a}<em>{l,t,k}$ on $\hat{b}</em>{m1,t,k}$, $\hat{b}<em>{m2,t,k}$ and $\hat{b}</em>{m3,t,k}$ (see also Table 4.2) for each season.</td>
<td>95</td>
</tr>
<tr>
<td>4.4</td>
<td>Results from regressing estimated OMCP amplitudes, $\hat{b}<em>{m,t,k}$, on the OMCP amplitudes the month before, $\hat{b}</em>{m-1,t,1}$, $\hat{b}<em>{m-1,t,2}$, $\hat{b}</em>{m-1,t,3}$ and $\hat{b}_{m-1,t,4}$, for July through December. Columns 2 through 6 show the estimated regression coefficients and stars that indicate whether the coefficient is significantly different from zero.</td>
<td>98</td>
</tr>
<tr>
<td>4.5</td>
<td>Correlation coefficients between residuals from the regression of $\hat{b}<em>{m,t,k}$ on $\hat{b}</em>{m-1,t,2}$ for $k_2 = 1, 2, 3, 4$, (see also Table 4.4) for each month.</td>
<td>99</td>
</tr>
<tr>
<td>4.6</td>
<td>Results from regressing estimated OMCP amplitudes, $\hat{b}<em>{m,t,k}$, on the OMCP amplitudes the month before, $\hat{b}</em>{m-1,t,k}$, and the amplitude for same month but the year before, $\hat{b}_{m,t-1,k}$. Columns 2 through 6 show the estimated regression coefficients and stars that indicate whether the coefficient is significantly different from zero.</td>
<td>100</td>
</tr>
<tr>
<td>4.7</td>
<td>The $d_{l,k}$ and $c_l$ constant used for the $R_l$ matrices, see equations (4.16) and (4.17).</td>
<td>108</td>
</tr>
</tbody>
</table>
4.8 The $d_{m,k}$ and $c_m$ constant used for the $S_m$ matrices, see equations (4.19) and (4.18) ................................. 108

4.9 Percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals, forecast without new ERA-40 data. ................................. 140

4.10 Percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals, for downscaling 2001-2002 with new ERA-40 data. ................................. 148

4.11 Proportion of total variance explained by the first four EOFs for each season and the first four OMCPs for each month, using only data through fall 2000. ................................. 149

4.12 Percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals, for downscaling 2000-2001 with new ERA-40 data. ................................. 152

C.1 List of unknown parameters in the model. Note that there are 89 season of data included in the model. The last column shows the number of parameters for the chosen number of basis vectors, $K_Y = 4$ and $K_X = 4$. Note that, in comparison, the total number of observations is $14641 \times 89 + 2736 \times 89 \times 3 = 2,033,561$. ................................. 212
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The reduced Gaussian grid (left) used in the ERA-40 project and the regular 2.5° latitude-longitude grid (right) on which the data are provided. These Figures were obtained from the ERA-40 archive at NCAR, <a href="http://dss.ucar.edu/pub/era40/">http://dss.ucar.edu/pub/era40/</a></td>
</tr>
<tr>
<td>4.2</td>
<td>The ERA-40 grid, a regular 2.5° × 2.5° latitude-longitude grid (grey points) and the stereographic Polar MM5 grid (blue points).</td>
</tr>
<tr>
<td>4.3</td>
<td>Another look at the ERA-40 grid, a regular 2.5° × 2.5° latitude-longitude grid (grey points), and the stereographic Polar MM5 grid (blue points) using a stereographic projection. The Polar MM5 grid covers the South Georgia Island (the upper left corner of the blue grid) but just misses the Falkland Islands.</td>
</tr>
<tr>
<td>4.4</td>
<td>Temperature fields, ERA-40 (circles) and Polar MM5 (squares) in Kelvin, year 1986. Note that 273.15K = 0°C. Bottom panel: Elevation values (in meters) for each Polar MM5 grid point.</td>
</tr>
<tr>
<td>4.5</td>
<td>Polar MM5 2-meter temperatures from all 24 years versus elevation for land locations only. The land use variable (i.e. ocean or land/permanent ice) was provided with the Polar MM5 data. We also show boxplots for every 200 meter interval where the width of the box is proportional to the amount of data in the interval.</td>
</tr>
<tr>
<td>4.6</td>
<td>Average Polar MM5 2-meter temperatures over land/permanent ice (left) and ocean (right) locations, for each season.</td>
</tr>
<tr>
<td>4.7</td>
<td>Average ERA-40 2-meter temperatures over all locations, for each month. The colors denote the four seasons and the line type is used to denote the first, second and third month within a season.</td>
</tr>
</tbody>
</table>
4.8 Centered temperature fields, ERA-40 (circles) and Polar MM5 (squares) in Kelvin, year 1986. The color palettes are separate for the ERA-40 and Polar MM5 data. 84

4.9 First five groups of OMCPs (circles) and EOF patterns (squares) for summer season \((v_{12,k}, v_{1,k}, v_{2,k} \text{ and } u_{s,k} \text{ for } k = 1, \ldots, 5)\). 86

4.10 First five groups of OMCPs (circles) and EOF patterns (squares) for fall season \((v_{3,k}, v_{4,k}, v_{5,k} \text{ and } u_{f,k} \text{ for } k = 1, \ldots, 5)\). 87

4.11 First five groups of OMCPs (circles) and EOF patterns (squares) for winter season \((v_{6,k}, v_{7,k}, v_{8,k} \text{ and } u_{w,k} \text{ for } k = 1, \ldots, 5)\). 88

4.12 First five groups of OMCPs (circles) and EOF patterns (squares) for spring season \((v_{9,k}, v_{10,k}, v_{11,k} \text{ and } u_{p,k} \text{ for } k = 1, \ldots, 5)\). 89

4.13 Proportion of total sample variance explained by the EOFs for each season (left) and OMCPs for each month (right). The colors denote the four seasons and the line type is used to denote the first (solid), second (dashed) and third (dotted) month within a season. 90

4.14 Proportion of the sample covariance between amplitudes lost by orthogonalizing the MCPs. The colors denote the four seasons and the line type is used to denote the first, second and third month within a season. 91

4.15 First five estimated Polar MM5 EOF amplitudes for summer (all 22 years), \(\hat{a}_{s,t,k_Y}\), plotted against the first five estimated ERA-40 OMCP amplitudes for the first summer month (December, all 22 years), \(\hat{b}_{12,t,k_X}\), for every combination of \(k_Y = 1, \ldots, 5 \text{ and } k_X = 1, \ldots, 5\). 93

4.16 Estimated December OMCP amplitudes plotted against estimated November OMCP amplitudes, for every combination of \(k_1 = 1, \ldots, 5 \text{ and } k_2 = 1, \ldots, 5\). 97

4.17 Density function of \(IG(2,1)\), the prior distribution used for \(r_l\) and \(s_m\) (left), and the implied density for \(\sqrt{r_l}\) and \(\sqrt{s_m}\). 114
4.18 Estimated posterior means (“-”) and 90% credible intervals (vertical lines) for the Polar MM5 amplitudes, $a_{t,t}$, for all 89 seasons. The top panel shows the first amplitude for each season, the bottom panel shows the fourth amplitude. ........................................... 122

4.19 Estimated posterior means (“-”) and 90% credible intervals (vertical lines) for the Polar MM5 amplitudes, $b_{m,t}$, for all 268 months. The top panel shows the first amplitude for each season, the bottom panel shows the fourth amplitude. ........................................... 123

4.20 Estimated posterior means of the Polar MM5 amplitudes, $a_{t,t}$, for all 89 seasons. ................................................................. 124

4.21 Estimated posterior means of the ERA-40 amplitudes, $b_{m,t}$, for all 268 months. ................................................................. 124

4.22 Estimated posterior densities of the square root of the $r_t$ and $s_m$ parameters. ................................................................. 125

4.23 Estimated posterior densities of the diagonal elements of $H_m$ for each month. ................................................................. 127

4.24 Estimated posterior densities of the 16 elements of the transition matrix $B_m$ for each month. ................................................................. 128

4.25 Estimated posterior densities of the standard deviations of the $C_l$ matrices, i.e. the square root of the diagonal elements of $C_l$. ................................................................. 130

4.26 Estimated 90% posterior credible intervals and posterior means (circles) for the correlations in $C_l$. The labels, $j,k$, indicate the $(j,k)$th element of $C_l$. ................................................................. 131

4.27 Estimated posterior densities of the standard deviations of the $D_m$ matrices, i.e. the square root of the diagonal elements of $D_m$. ................................................................. 132

4.28 Estimated 90% posterior credible intervals and posterior means (circles) for the correlations in $D_m$. The labels, $j,k$, indicate the $(j,k)$th element of $D_m$. ................................................................. 133

4.29 ERA-40 and Polar MM5 temperature anomalies, in degrees Kelvin, for the last year. ................................................................. 135
4.30 Forecasted amplitudes, $b_{m,t,k}$ (left) and $a_{t,t,k}$ (right), without new ERA-40 data. The circles show the predicted means, the thick and think lines show the middle 50% and 90% prediction intervals, respectively. The x-marks show the estimated amplitudes. 136

4.31 Forecast without new ERA-40 data. The maps show estimated posterior predicted mean ERA-40 and Polar MM5 temperature-anomaly fields, $U_{t\,\tilde{a}_{t,t}}$ (squares) and $V_{m\,\tilde{B}_{m,t}}$ (circles), for winter 2001 - fall 2002 in degrees Kelvin. 137

4.32 Forecast without new ERA-40 data. The maps show estimated posterior predicted standard deviations for ERA-40 and Polar MM5 temperature fields, in degrees Kelvin. 138

4.33 $MSE(U_{t\,\tilde{a}_{t,t}})$ (upper panel) for Polar MM5 temperatures and $MSE(V_{m\,\tilde{B}_{m,t}})$ (lower panel) for ERA-40 data. Red points represent MSE for forecasted fields without new ERA-40 data. 139

4.34 Estimated $b_{m,t,k}$ amplitudes (left) and predicted $a_{t,t,k}$ amplitudes (right) for downscaling 2001-2002 without new ERA-40 data. The circles show the predicted means, the thick and think lines show the middle 50% and 90% prediction intervals, respectively. The x-marks show the estimated amplitudes. 142

4.35 Downscaling 2001-2002 with new ERA-40 data. The maps show the estimated posterior mean ERA-40 fields and the estimated posterior predicted mean of Polar MM5 temperature-anomaly fields for winter 2001 - fall 2002 in degrees Kelvin. 143

4.36 Downscaling 2001-2002 with new ERA-40 data. The maps show the estimated posterior standard deviations of ERA-40 and Polar MM5 temperature-anomaly fields for winter 2001 - fall 2002 in degrees Kelvin. 144

4.37 Estimated posterior predicted mean Polar MM5 temperature fields using no new ERA-40 data (first column), estimated posterior mean of downscaled Polar MM5 temperature fields using the ERA-40 data from the last year (second column) and the actual Polar MM5 temperature data. 146
4.38 $MSE(U_i \bar{a}_{t,t})$ (upper panel) for Polar MM5 temperatures and $MSE(V_m \bar{b}_{m,t})$ (lower panel) for ERA-40 data. Red points represent MSE for down-scaled fields with new ERA-40 data 2001-2002.

4.39 Estimated $b_{m,t,k}$ amplitudes (left) and predicted $a_{l,t,k}$ amplitudes (right) for downscaling 2000-2001 without new ERA-40 data. The circles show the predicted means, the thick and thin lines show the middle 50% and 90% prediction intervals, respectively. The x-marks show the estimated amplitudes.

4.40 Estimated posterior predicted standard deviations (first column) and means (second column) of Polar MM5 temperature fields from downscaling the 2000-2001 ERA-40 data and the actual Polar MM5 temperature data (third column).

4.41 $MSE(U_i \bar{a}_{t,t})$ (upper panel) for Polar MM5 temperatures and $MSE(V_m \bar{b}_{m,t})$ (lower panel) for ERA-40 data. Red points represent MSE for down-scaled fields with new ERA-40 data 2000-2001.

B.1 Antarctica overview map, obtained from Landsat Image Mosaic of Antarctica (LIMA), http://lima.usgs.gov.


B.8 Centered temperature fields ERA-40 (circles) and Polar MM5 (squares)
in degrees Kelvin, year 1979-1982. ................................. 178
B.9 Centered temperature fields ERA-40 (circles) and Polar MM5 (squares)
in degrees Kelvin, year 1983-1986. ................................. 179
B.10 Centered temperature fields ERA-40 (circles) and Polar MM5 (squares)
in degrees Kelvin, year 1987-1990. ................................. 180
B.11 Centered temperature fields ERA-40 (circles) and Polar MM5 (squares)
in degrees Kelvin, year 1991-1994. ................................. 181
B.12 Centered temperature fields ERA-40 (circles) and Polar MM5 (squares)
in degrees Kelvin, year 1995-1998. ................................. 182
B.13 Centered temperature fields ERA-40 (circles) and Polar MM5 (squares)
in degrees Kelvin, year 1999-2002. ................................. 183
B.14 Estimated EOF amplitudes for summer plotted against estimated OMCP
amplitudes for January for every combination of $k_Y = 1, \ldots, 5$ and
$k_X = 1, \ldots, 5$. .......................................................... 185
B.15 Estimated EOF amplitudes for summer plotted against estimated OMCP
amplitudes for February for every combination of $k_Y = 1, \ldots, 5$ and
$k_X = 1, \ldots, 5$. .......................................................... 186
B.16 Estimated EOF amplitudes for fall plotted against estimated OMCP
amplitudes for March for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. .................................................. 187
B.17 Estimated EOF amplitudes for fall plotted against estimated OMCP
amplitudes for April for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. .................................................. 188
B.18 Estimated EOF amplitudes for fall plotted against estimated OMCP
amplitudes for May for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. .................................................. 189
B.19 Estimated EOF amplitudes for winter plotted against estimated OMCP
amplitudes for June for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. .................................................. 190
B.20 Estimated EOF amplitudes for winter plotted against estimated OMCP amplitudes for July for every combination of \(k_Y = 1, \ldots, 5\) and \(k_X = 1, \ldots, 5\). ............................................................................. 191

B.21 Estimated EOF amplitudes for winter plotted against estimated OMCP amplitudes for August for every combination of \(k_Y = 1, \ldots, 5\) and \(k_X = 1, \ldots, 5\). ............................................................................. 192

B.22 Estimated EOF amplitudes for spring plotted against estimated OMCP amplitudes for September for every combination of \(k_Y = 1, \ldots, 5\) and \(k_X = 1, \ldots, 5\). ............................................................................. 193

B.23 Estimated EOF amplitudes for spring plotted against estimated OMCP amplitudes for October for every combination of \(k_Y = 1, \ldots, 5\) and \(k_X = 1, \ldots, 5\). ............................................................................. 194

B.24 Estimated EOF amplitudes for spring plotted against estimated OMCP amplitudes for November for every combination of \(k_Y = 1, \ldots, 5\) and \(k_X = 1, \ldots, 5\). ............................................................................. 195

B.25 Estimated January OMCP amplitudes plotted against estimated December OMCP amplitudes, for every combination of \(k_1 = 1, \ldots, 5\) and \(k_2 = 1, \ldots, 5\). ............................................................................. 196

B.26 Estimated February OMCP amplitudes plotted against estimated January OMCP amplitudes, for every combination of \(k_1 = 1, \ldots, 5\) and \(k_2 = 1, \ldots, 5\). ............................................................................. 197

B.27 Estimated March OMCP amplitudes plotted against estimated February OMCP amplitudes, for every combination of \(k_1 = 1, \ldots, 5\) and \(k_2 = 1, \ldots, 5\). ............................................................................. 198

B.28 Estimated April OMCP amplitudes plotted against estimated March OMCP amplitudes, for every combination of \(k_1 = 1, \ldots, 5\) and \(k_2 = 1, \ldots, 5\). ............................................................................. 199

B.29 Estimated May OMCP amplitudes plotted against estimated April OMCP amplitudes, for every combination of \(k_1 = 1, \ldots, 5\) and \(k_2 = 1, \ldots, 5\). ............................................................................. 200
B.30 Estimated June OMCP amplitudes plotted against estimated May OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. ......................................................... 201

B.31 Estimated July OMCP amplitudes plotted against estimated June OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. ........................................ 202

B.32 Estimated August OMCP amplitudes plotted against estimated July OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. ......................................................... 203

B.33 Estimated September OMCP amplitudes plotted against estimated August OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. ......................................................... 204

B.34 Estimated October OMCP amplitudes plotted against estimated September OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. ......................................................... 205

B.35 Estimated November OMCP amplitudes plotted against estimated October OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. ......................................................... 206

B.36 First five groups of OMCPs and EOF patterns for summer season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001. ......................................................... 207

B.37 First five groups of OMCPs and EOF patterns for fall season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001. ......................................................... 208

B.38 First five groups of OMCPs and EOF patterns for winter season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001. ......................................................... 209

B.39 First five groups of OMCPs and EOF patterns for spring season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001. ......................................................... 210
Chapter 1: Introduction

In this age of great data abundance, scientists are often faced with the challenge of learning from very large datasets. Climate science is one of the scientific fields where datasets are getting increasingly larger. This is, of course, partly due to increasing technological advances (e.g. satellites, climate models etc.) and partly due to the nature of the subject being studied. There are a large number of interacting processes that make what we call “climate” and the spatial domains tend to be large (e.g. the Earth). The relevant datasets in climate studies therefore tend to have many spatial observations, and depending on the process of interest, the temporal frequency may be high as well.

Climatologists have many multivariate statistics methods up their sleeve that they use to study climatic systems, some of which are reviewed in Section 1.2. The methods I focus on here are essentially dimension reduction techniques. They aim at explaining high-dimensional process in terms of only a few patterns (or basis functions). These methods are used by climatologists to gain understanding of the physical processes at work. Statisticians interest in these methods is usually as tools for building models for high-dimensional data, as is briefly discussed in Section 1.1. These two views are, of course, not mutually exclusive.
In this thesis I add to the arsenal of dimension-reduction techniques for studying and modeling high-dimensional data. In Chapter 3 I present and discuss a technique I call Maximum Covariance Patterns (MCPs). The new method is geared towards the case where two spatio-temporal processes are of interest and where we wish to predict one process from the other. Statistical downscaling is an example of a situation where we want to predict one (high-resolution) process from another (low-resolution) process. I briefly review statistical downscaling in the setting of climatology in Section 1.3. In Chapter 4 I pursue an example of statistical downscaling using the new approach of MCPs.

The statistical framework of this thesis is hierarchical and Bayesian. In Chapter 2 I present a general Bayesian hierarchical modeling framework that incorporates dimension reduction of the spatial dimension of two spatio-temporal processes. I also show how a model for the original processes implies a model for the coefficients (i.e. amplitudes) of the basis vectors used to obtain the dimension reduction.

The thesis is organized as follows. In the remainder of this chapter, I review dimension reduced approaches to statistical modeling of spatio-temporal processes (Section 1.1), statistical dimension reduction approaches (Section 1.2) and statistical downscaling (Section 1.3). A general Bayesian hierarchical modeling framework is presented in Chapter 2. In Chapter 3 I demonstrate how the multivariate methods, Empirical Orthogonal Functions (EOFs), Maximum Covariance Analysis (MCA) and Canonical Correlation Analysis (CCA) can be useful as tools for dimension reduction of space-time processes. Furthermore, in Sections 3.4 and 3.5 I develop new ways of obtaining basis vectors for dimension reduction, Maximum Covariance Patterns (MCP) and Orthogonal MCPs (OMCPs), respectively. In Chapter 4 I pursue an
application of statistical downscaling of surface temperatures over the Antarctic, using EOFs and OMCPs in a Bayesian hierarchical model. Chapter 5 contains discussion.

1.1 Dimension reduced spatio-temporal modeling

The field of spatial and spatio-temporal statistics is increasingly faced with the challenge of very large or massive data sets. Examples include data obtained from remote sensing satellites, global weather stations, output from climate models and medical imagery. The traditional approach in spatial statistics (Cressie, 1993; Banerjee et al., 2004; Gelfand et al., 2010) to modeling point referenced spatial processes, involves modeling the covariance at every pair of spatial locations. When the number of spatial locations is huge, the corresponding covariance matrices become prohibitively high-dimensional and the needed manipulations of these covariance matrices (multiplications, inverse, storing) become too computationally expensive to do in practice. The dimension reduced approach involves representing the spatial (or spatio-temporal) process in terms of a lower dimensional latent process. There have been many different approaches to such dimension-reduced representations in the literature.

Consider a spatio-temporal process $Y_t(s)$ at spatial location $s \in D_Y \subset \mathbb{R}^d$, and time $t$ where $t = 1, \ldots, T$. Let $Y_t = (Y_t(s_1), \ldots, Y_t(s_{N_Y}))'$ where $s_1, \ldots, s_{N_Y}$ index $N_Y$ locations where the process is observed. A common general approach to spatio-temporal modeling is to apply a linear dynamical spatio-temporal model (DSTM; Cressie and Wikle, 2011), for example

$$Y_t = MY_{t-1} + \epsilon_{Y_t}, \quad t = 1, \ldots, T,$$  \hspace{1cm} (1.1)
where $\epsilon_t \overset{iid}{\sim} N(0, \Sigma)$ and $M$ is an unknown $N_Y \times N_Y$ dimensional transition matrix. A dimension-reduced approach usually involves two stages. First, the process $Y_t$ is assumed to have a dimension-reduced representation of the form

$$Y_t = Ua_t + \eta_t.$$  

(1.2)

The matrix $U$ is a known $N_Y \times K_Y$ dimensional matrix, $a_t$ is a $K_Y$ dimensional random vector and $K_Y << N_Y$. The error vector, $\eta_t$, is often ignored or assumed to have mean zero and a covariance matrix of a simple form, e.g. $\Sigma_\eta = \sigma^2 I$ where $\sigma^2$ is a scalar. Secondly, a linear dynamical spatio-temporal model is assumed for the latent process $a_t$:

$$a_t = Ha_{t-1} + \xi_t$$  

(1.3)

where usually $\xi_t \overset{iid}{\sim} N(0, \Sigma_\xi)$. The transition matrix $H$ is unknown, but is of much smaller dimension than $M$. Furthermore, the covariance matrix $\Sigma_\xi$ is of much smaller dimension than $\Sigma_\epsilon$. Dimension-reduced approach not only eliminates the need to model the full covariance matrix $\Sigma_\epsilon$, but also allows for nonstationarity which is difficult to achieve with traditional methods. Further discussion of strengths and issues of dimension-reduced modeling is presented in Chapter 2.

Different approaches to dimension-reduced spatial and spatio-temporal modeling are essentially different approach to choosing the matrix $U$. One common approach is to use empirical orthogonal functions (EOFs), see for example Wikle and Cressie (1999) and Berliner et al. (2000). A more detailed discussion of the use of EOFs in this setting is presented in Section 3.1. Many other approaches and examples can be found in the statistical literature. For example, Stroud et al. (2001) use locally weighted
mixtures of linear regressions; Wikle et al. (2001) and Kang et al. (2010) use multi-
resolution (wavelet) bases; Xu et al. (2005) use Fourier basis functions. Finally, Wikle
(2010) gives a general overview of dimension-reduced spatial modeling and Section
7.1.3 in Cressie and Wikle (2011) contains a detailed discussion of dimension-reduced
spatio-temporal modeling.

1.2 Dimension reduced representation

Here I discuss various linear dimension reduction techniques. These techniques
aim at describing the most important aspects of multivariate datasets or processes.
That is, they aim at separating the “signal” from the “noise”. They are all based on
some variation of eigen techniques.

In Section 1.2.1 I discuss two approaches that involve dimension reduction of two
fields: Canonical Correlation Analysis (CCA) and Maximum Covariance Analysis
(MCA). In Section 1.2.2 I review max/min autocorrelation factors (MAF) and in
Section 1.2.3 I review Principal Oscillation Patterns (POPs). CCA, MCA and POPs
are routinely used in climatology.

1.2.1 Joint dimension reduction of two random vectors

We start with establishing some notation. Let \( \mathbf{Y} \) and \( \mathbf{X} \) be \( N_Y \) and \( N_X \)
dimensional vectors, respectively. These random vectors can represent any random
phenomenon but here we will usually think of \( \mathbf{Y} \) and \( \mathbf{X} \) as being spatially referenced,
that is, \( \mathbf{Y} \) and \( \mathbf{X} \) represent some spatial processes observed at \( N_Y \) and \( N_X \) spatial
locations, respectively. We sometimes refer to \( \mathbf{Y} \) and \( \mathbf{X} \) as fields. We partition the
joint covariance matrix of \( Y \) and \( X \) as

\[
\Sigma = \begin{pmatrix}
\Sigma_{YY} & \Sigma_{YX} \\
\Sigma_{XY} & \Sigma_{XX}
\end{pmatrix}.
\] (1.4)

We assume that \( T \) observations of the fields \( Y \) and \( X \) are organized in an \( N_Y \times T \) matrix \( Y \) and an \( N_X \times T \) matrix \( X \). We also assume that both matrices are temporally centered i.e., \( \sum_{t=1}^{T} y_{it} = 0 \) and \( \sum_{t=1}^{T} x_{jt} = 0 \) for all \( i \) and \( j \). We obtain the sample covariance matrix as

\[
S = \begin{pmatrix}
S_{YY} & S_{YX} \\
S_{XY} & S_{XX}
\end{pmatrix} = \frac{1}{(T-1)} \begin{pmatrix}
YY' & YX' \\
XY' & XX'
\end{pmatrix}.
\] (1.5)

If the \( T \) observations are a random sample of \( Y \) and \( X \) then \( S \) is an unbiased estimator of \( \Sigma \) (see, for example, Johnson and Wickern, 2002, page 122). However, in practice we usually do not have a random sample but rather observations at \( T \) time points. Also, we note that in statistical climatology literature it is customary to use \( \frac{T-1}{T} S \) (e.g., von Storch and Zwiers, 2001, page 299).

1.2.1.1 Canonical Correlation Analysis

CCA is a method that quantifies the association between \( Y \) and \( X \) by finding pairs of linear combinations that have the highest correlation, among those that are uncorrelated with former pairs. Specifically, the objective is to find pairs of linear combinations,

\[
A_1 = u'_1 Y \quad \text{and} \quad B_1 = v'_1 X
\]

\[...
\]

\[
A_d = u'_d Y \quad \text{and} \quad B_d = v'_d X, \quad d = \min(N_Y, N_X),
\] (1.6)
so that $\rho_1 = \text{Cor}(A_1, B_1)$ is maximized and so that for all $i = 2, \ldots, d$, $A_i$ and $B_i$ have the maximum correlation ($\rho_i$) among all linear combinations that are uncorrelated with $A_1, \ldots, A_{i-1}$ and $B_1, \ldots, B_{i-1}$. The variables $A_i$ and $B_i$ are called the $i^{th}$ canonical variables, $\rho_i$ is called the $i^{th}$ canonical correlation coefficient and the vectors $u_i$ and $v_i$ are called $i^{th}$ canonical vectors.

The canonical vectors can be found by first finding the singular value decomposition (SVD) of the following matrix$^1$:

$$\Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} = \hat{U}_{NY \times NY} D_{NY \times NX} \hat{V}_X'^{NY}$$

(1.7)

where $\hat{U}$ and $\hat{V}$ are orthogonal and $D$ is a diagonal matrix (Johnson and Wickern, 2002, Ch. 10). Some properties of the SVD are given in Appendix Section A.1.3. The canonical vectors, $u_i$ and $v_i$, are the columns of

$$U = \Sigma_{YY}^{-1/2} \hat{U} \quad \text{and} \quad V = \Sigma_{XX}^{-1/2} \hat{V}$$

(1.8)

and the elements of $D$ are the canonical correlation coefficients $\rho_1, \ldots, \rho_d$. The columns of

$$\Sigma_{YY} U \quad \text{and} \quad \Sigma_{XX} V$$

(1.9)

are called canonical correlation (CC) patterns and the canonical variables $A_i$ and $B_i$ are sometimes called CC coordinates (von Storch and Zwiers, 2001, Ch. 14.1.3). Sample canonical vectors are obtained by substituting the covariance and cross-covariance matrices in (1.7) - (1.9) with their sample counterparts. For a more complete overview of CCA and further references see, Chapter 14 in von Storch and Zwiers (2001) and Chapter 5 in Cressie and Wikle (2011).

$^1$Here $\Sigma_{YY}^{-1/2} = PE^{-1/2}P'$ where $PEP'$ is the spectral decomposition of $\Sigma_{YY}$ and $E^{-1/2} = \text{diag}(1/\sqrt{E_{11}}, \ldots, 1/\sqrt{E_{NY \times NX}})$. 

7
1.2.1.2 Maximum Covariance Analysis

Akin to CCA, MCA seeks to quantify the association between two sets of variables but the emphasis is on covariances instead of correlations. The method also seeks pairs of linear combinations as in (1.6), but the objective is to maximize the covariance $c_i = \text{Cov}(A_i, B_i)$ subject to the constraints that the vectors are orthogonal and of unit length, i.e. $u_i' u_i = v_i' v_i = 1$ and $u_i' u_j = v_i' v_j = 0$ for $i \neq j$. The variables $A_i$ and $B_i$ are often called singular variables and $u_i$ and $v_i$ are called SVD vectors, but here I call them Maximum Covariance (MC) variables and MC patterns, respectively.

The solution to MCA is found via SVD of the cross-covariance matrix:

$$\Sigma_{YX} = UDV'$$  (1.10)

(von Storch and Zwiers, 2001). The $i$th columns of $U$ and $V$ contain the $i$th pair of the MC vectors and the diagonal matrix $D$ contains the canonical covariances $c_1, \ldots, c_d$. Note that the $U$ and $V$ matrices here are not the same as the $U$ and $V$ matrices in the CCA solution in (1.8). Sample MC vectors are obtained via SVD of the sample cross-covariance matrix $S_{YX}$.

We note that MCA is often called “SVD analysis”, which refers to the fact that the solution to MCA can be obtained by singular value decomposition (SVD) of the cross-covariance matrix. However, the objective of the analysis is to find orthogonal vectors that maximize covariance and so the name Maximum Covariance Analysis is more descriptive than the name of the mathematical procedure used to obtain the solution. The name “SVD analysis” is widely used in literature but MCA may be catching on, it is used for example in the book by von Storch and Zwiers (2001) and by Salim et al. (2005) and Frankignoul and Sennéchael (2007).
1.2.1.3 Some similarities and differences between CCA and MCA

Both CCA and MCA seek to summarize the association between $Y$ and $X$. CCA and MCA are sometimes looked on as competing methods but as argued by Cherry (1996), they are different techniques with different goals. The main difference being that CCA is a method of analysing correlation structure whereas MCA is a method of analysing covariance structure. MCA was first used in climatology by Prohanska (1976), but was popularized in this field by Bretherton et al. (1992) and Wallace et al. (1992). Examples of CCA in climatic research include Zorita et al. (1992), Juneng and Tangang (2008) and Gyllistras et al. (1994).

Another difference between CCA and MCA is that the maximization problems are set up with different types of constraints. In CCA the correlation of the CC variables, $A_i$ and $B_i$, is maximized subject to a constraint on the CC variables, that they are uncorrelated between pairs; i.e. $\text{Cor}(A_i, A_j) = \text{Cor}(B_i, B_j) = \text{Cor}(A_i, B_j) = 0$ for $i \neq j$. The resulting CC vectors in (1.8) are generally not orthogonal nor are the CC patterns in (1.9). In MCA, on the other hand, the covariance of the MC variables is maximized subject to constraints that the MC vectors, $u_i$ and $v_i$, are orthonormal; i.e. $u'_i u_i = v'_i v_i = 1$ and $u'_i u_j = v'_i v_j = 0$ for $i \neq j$. The resulting MC patterns are therefore orthonormal. However, the MC variables ($A_i$ and $B_i$) are generally not uncorrelated, that is although $\text{Cor}(A_i, B_j) = 0$ for $i \neq j$, generally $A_i$ and $A_j$ will be correlated for $i \neq j$ and so will $B_i$ and $B_j$.

A common aspect of both CCA and MCA is that one needs to exercise caution in interpreting the results. Difficulties in interpretation have been documented in the statistical literature for CCA, see for example Rencher (1992). Also, Bretherton et al. (1992) compare MCA and CCA and favor MCA over CCA on grounds of its
simplicity and ease of interpretation. Newman and Sardeshmukh (1995) raise questions about the usefulness of MCA and argue that MCA is only capable of detecting coupled patterns under very special circumstances. In light of these papers, Cherry (1996) set out to illustrate some difficulties with interpreting the results of both CCA and MCA. In particular, the use of so-called homogeneous and heterogeneous covariance or correlation maps. These maps are constructed from the covariances or correlations between the original fields and the CC/MC variables from the same field (homogeneous map) and from the other field (heterogeneous map).\footnote{For CCA the homogeneous covariance maps are the same as the CA patterns, i.e. the columns in $U$ and $V$, but that does not hold for MCA.} Cherry (1996) showed through simulated examples, that there is a potential for spurious patterns and correlations to show up in both MCA and CCA, and increasingly so for small sample sizes. That is, MCA and CCA can produce highly correlated and seemingly coupled patterns even when there is no relationship between or within the original fields.

\subsection{1.2.2 Maximum/minimum autocorrelation factors}

Switzer and Green (1984) propose a procedure they call max/min autocorrelation factors (MAF) for analysis of multivariate spatial data. Let $Y(s) = (Y_1(s), \ldots, Y_p(s))'$ be a $p$-dimensional, second-order stationary spatial process, where $s$ is a location on a regular spatial grid. For a given spatial lag $\delta$, the MAF procedure finds uncorrelated linear combinations, called factors, that maximize autocorrelation at the lag $\delta$. Specifically, MAF finds $p$ factors

$$A_i(s) = u_i'Y(s) \quad i = 1, \ldots, p,$$  \hspace{1cm} (1.11)
so that the first autocorrelation $\text{Cor}(A_1(s), A_1(s + \delta))$ is maximized and for $i = 2, \ldots, p$, $\text{Cor}(A_i(s), A_i(s + \delta))$ is maximized under the constraint that the factors are uncorrelated, i.e. $\text{Cov}(A_i(s), A_j(s)) = 0$ for $i \neq j$, and have zero cross-correlation at spatial lag $\delta$, $\text{Cov}(A_i(s), A_j(s + \delta)) = 0$ for $i \neq j$.

Let $\Sigma = \text{Cov}(Y(s))$ be the point-wise covariance matrix. Note that $\Sigma$ does not depend on $s$ since the multivariate spatial process is assumed to be second-order stationary. Let $C(\delta) = \text{Cov}(Y(s), Y(s + \delta))$ be the autocovariance matrix for spatial lag $\delta$. The MAF vectors, $u_i$, that provide the solution to the optimization problem described above, can be obtained by first obtaining the eigenvectors $\tilde{u}_i$ of

$$
\Sigma^{-1/2}C(\delta)\Sigma^{-1/2}.
$$

Then the MAF vectors are $u_i = \Sigma^{-1/2}\tilde{u}_i$. The eigenvalues of $\Sigma^{-1/2}C(\delta)\Sigma^{-1/2}$ are the auto-correlations, $\text{Cor}(A_i(s), A_i(s + \delta))$.

MAF explores the spatial structure of the multivariate spatial process but is limited to lag $\delta$. There have been extensions introduced that attempt to include a wider range of spatial lags, see Grunsky and Agterberg (1992); Bailey and Krzanowski (2000); Bailey et al. (2005) Krzanowski and Bailey (2007).

The MAF can be seen as a point-wise dimension reduction technique. The dimension reduction is for the $p$ dimensional space but not the spatial dimension. That can be useful when $p$ is large or when the elements of $Y(s)$ are highly correlated and could be represented as a lower-dimension spatial process. Another point-wise dimension reduction that is sometimes used is principal components analysis (or EOFs, see Section 3.1) of the point-wise covariance matrix $\Sigma$. That approach represents the $Y(s)$ process in lower dimensions but does not include the spatial aspect of the process as MAF does. However, Cressie and Helterbrand (1994) show that point-wise principal
components and MAF can be seen as special cases of a larger class of spatial principal components analysis.

Finally, note that MAF can be formulated as CCA (Bailey and Krzanowski, 2000). This can be seen if we consider \( \mathbf{Y}(s) \) and \( \mathbf{Y}(s + \delta) \) as two groups of equally many random variables with the same covariance matrix, i.e. in the notation in Section 1.2.1.1 we have \( \Sigma_{YY} = \Sigma_{XX} \). Then the formulation of MAF is the same as the formulation for CCA. Note that in this case the two CC vectors in each pair are the same, i.e. \( \mathbf{u}_i = \mathbf{v}_i \).

### 1.2.3 Principal Oscillation Patterns

Principal Oscillation Patterns (POPs) are used to study the underlying dynamics of a first-order dynamical system through a spectral decomposition of the propagator matrix (Hasselman, 1988; von Storch et al., 1988). POP analysis have been routinely used to diagnose the space-time variability of a climate system, see von Storch et al. (1988) and Xu and von Storch (1990) for examples. Overviews of POP analysis can be found in von Storch et al. (1995), von Storch and Zwiers (2001) and Cressie and Wikle (2011).

Consider first a deterministic first-order linear dynamical system

\[
\mathbf{y}_t = M \mathbf{y}_{t-1} \quad t = 1, 2, \ldots \quad (1.13)
\]

where \( \mathbf{y}_t, t = 1, 2, \ldots \) are \( N_Y \) dimensional real vectors. The general solution of this system is given by

\[
\mathbf{y}_t = M^t \mathbf{y}_0 \quad t = 1, 2, \ldots \quad (1.14)
\]

where the vector \( \mathbf{y}_0 \) is a starting value and \( M^t \) is the \( t^{th} \) power of \( M \). The characteristics of the linear dynamical system in (1.13) can be studied through the eigen
The decomposition of the propagator matrix $M$:

$$M = WAV'$$

(1.15)

where $W$ and $V$ are the right and left normalized eigenvectors and $\Lambda$ is a diagonal matrix with the eigenvalues $\lambda_k$, $k = 1, \ldots, N_Y$, on the diagonal. Note that $M$ is typically not symmetric and so the eigenvalues can be complex numbers. Also, the eigenvalues of $M$ are assumed to be distinct. We have that $W^{-1} = V'$ and therefore we have $M' = W\Lambda V'$. The dynamical system in (1.13) is critically dependent on the eigenvalues $\lambda_k$ (see, for example, Section 3.2.1 in Cressie and Wikle, 2011).

Applying this eigen decomposition of the transition matrix of first-order dynamic system for a spatio-temporal process is called POP analysis. Here the vectors $y_t$ are spatially referenced, i.e. observations of a spatio-temporal process, and the right eigenvectors, $w_k$, are called principal oscillation patterns (POPs). The POPs are linearly independent (but not necessarily orthogonal) and so they form a basis for $\mathbb{R}^{N_Y}$. Therefore, each $y_t$ can be expressed in terms of POPs:

$$y_t = Wa_t \quad \Rightarrow \quad a_t = W^{-1}y_t = V'y_t.$$

(1.16)

The elements of the $N_Y$ dimensional vector $a_t$ are called POP coefficients. Substituting (1.16) and (1.15) into the dynamical system in (1.13) yields

$$Wa_t = MWa_{t-1}$$

$$\Rightarrow \quad a_t = W^{-1}WAV'Wa_{t-1} = \Lambda a_{t-1}.$$

(1.17)

The POP coefficients therefore evolve according to a first-order difference equation and assuming that $a_{0,k} = 1$, we get $a_{t,k} = \lambda_k^t$.

The oscillations of $y_t$ over time can be are studied through the complex POPs. Note that the complex POPs occur in conjugate pairs. Each complex POP is the
sum of the real and imaginary parts, \( w_k = w_k^R + iw_k^I \) where \( i = \sqrt{-1} \). The general evolution of a dampened POP (i.e. \( |\lambda_k| < 1 \)) can be described schematically as
\[
\cdots \rightarrow w_k^R \rightarrow -w_k^I \rightarrow -w_k^R \rightarrow w_k^I \rightarrow w_k^R \rightarrow \cdots
\]
with a period \( 2\pi/\phi_k \), where \( \phi_k \) is the angle formed by \( \lambda_k \) in the complex plane.

Now consider a stochastic first-order dynamic system
\[
Y_t = MY_{t-1} + \eta_t
\]
where \( \eta_t \) is a vector of zero-mean random errors. In practice, the transition matrix \( M \) is unknown and needs to be estimated before applying the POP analysis. Assuming second-order stationarity of the process, \( M \) is estimated as
\[
\hat{M} = S_1 S^{-1}
\]
where \( S \) is the estimated covariance matrix and \( S_1 \) is the estimated lag 1 covariance matrix
\[
S = \hat{\text{Cov}}(Y_t) \quad \text{and} \quad S_1 = \hat{\text{Cov}}(Y_t, Y_{t-1}).
\]
Note that \( S \) and \( S_1 \) do not depend on \( t \) if the process is second-order stationary. The sample POPs are obtained by the eigendecomposition of \( \hat{M} \). von Storch et al. (1988) give a criterion for distinguishing between POPs that contain useful information in the sense that the evolution in (1.18) holds for the sample POPs.

There is a connection between POPs and CCA. If we view \( Y_t \) and \( Y_{t-1} \) as two random vectors then the CC patterns are the same as the POPs (Cressie and Wikle, 2011, Section 5.6.2). In essence, POP can be viewed as a time-lagged CCA while MFA can be viewed as a spatially lagged CCA.
1.3 Statistical Downscaling

Downscaling refers to the situation where one wishes to obtain information about a process (e.g. temperatures) at small scales from data at larger scales. This is a very common situation with meteorological and climatic data. Global numerical weather forecast models and climate models produce data (model output) on large scales, represented on low-resolution grids, while the interest is often in local climate processes at high-resolution grids.

A general circulation model (GCM) is a mathematical model of the general circulation of global atmospheric or oceanic processes. Atmospheric and oceanic GCMs are key components in global climate models that are used to study planetary climate and project future climate under varying scenarios of future concentrations of greenhouse gasses (Solomon et al., 2007, for example). The spatial resolution of GCMs is typically between 1 and 5 degrees in latitude and longitude. Increasing the spatial resolution of GCMs is hampered by computational resources. Downscaling of GCM output is of particular interest when assessing the impact of future climate change (conditional on a given emission scenarios) on local climate and local resource systems. For these impact studies a higher-resolution of model output is needed.

Regional climate models (RCMs) are a numerical (deterministic) approach to downscaling. RCMs model finer-scale climate processes and have a much higher spatial resolution than GCM but only cover a small region. RCMs are driven by a GCM, that is GCM output is used as input at the edges of an RCM. The NARCCAP (Mearns et al., 2009) project is an example of this, which involves running a set of RCMs driven by a set of GCMs over most of North America.
Statistical downscaling involves a statistical model for the same endeavor, that is predicting at small spatial scales based on data (or GCM output) at larger scales. There are many examples of statistical downscaling in the climate literature and with large datasets, such as GCM model outputs, the dimension reduction methods of EOF, CCA and MCA have proved to be useful. For example, von Storch et al. (1993) use large-scale North Atlantic sea level pressure (both observations and GCM outputs) to explain small-scale rainfall over the Iberian Peninsula. They propose using the leading CCA vector as a basis vector and use a regression between amplitudes (CC coefficients) to predict future rainfall amplitude. Gyllistras et al. (1994) and Cui et al. (1995) use similar methods and Widmann et al. (2003) use MCA basis vectors for downscaling. Overview and comparisons of methods used for statistical downscaling in climate research can be found in Wilby et al. (1998) and Zorita and von Storch (1999). In Chapter 4 of this thesis we apply new approaches to dimension reduction (Maximum Covariance Patterns) for statistical downscaling of surface temperatures over the Antarctic.

The statistical downscaling approach is related to what is called “change-of-support” problems in spatial statistics, see for example Cressie (1993) and Gotway and Young (2002). Wikle and Berliner (2005) provide a Bayesian perspective.
Chapter 2: A Hierarchical and Dimension Reduced Modeling Framework for Two Spatio-Temporal Processes

In this Chapter I discuss a general hierarchical modeling framework that incorporates dimension reduction of the spatial dimensions of two spatio-temporal processes. I also show how a model for the original processes implies a model for the coefficients of the basis vectors used to obtain the dimension reduction. To better convey ideas I first consider two spatial processes in Section 2.1 and then expand to two spatio-temporal processes in Sections 2.2 and 2.3. In Section 2.4 I consider the case where two spatio-temporal processes are defined on a continuous spatial domain and in Section 2.5 I present a Bayesian Filtering approach for a special case of the model.

2.1 Two spatial processes

First we consider two spatial processes, say $Y(s)$ and $X(c)$ where $s$ and $c$ are locations in spatial domains $D_Y$ and $D_X$, $s \in D_Y \subset \mathbb{R}^{d_Y}$ and $c \in D_X \subset \mathbb{R}^{d_X}$. Let $Y$ and $X$ be $N_Y$ and $N_X$ dimensional random vectors containing the processes at $N_Y$ and $N_X$ locations, i.e. $Y_i = Y(s_i)$, for $i = 1, \ldots, N_Y$ and $X_j = X(c_j)$, for $j = 1, \ldots, N_X$. The corresponding $N_Y$ and $N_X$ observations of $Y$ and $X$ are denoted by $Y_{\text{obs}}$ and $X_{\text{obs}}$, respectively. In the statistical downscaling example in Chapter 4 $Y$ is a high resolution (Polar MM5) temperature field and $X$ is a coarse resolution (ERA-40)
temperature field. Our goal is to build a model for the joint distribution of $Y$ and $X$, $[Y, X|\theta]$. The vector $\theta$ represents unknown model parameters and the notation $[X]$ stands for “the probability distribution of the random variable $X$” (Gelfand and Smith, 1990). In a Bayesian framework we model the joint distribution of the random vectors $Y$ and $X$ and the parameters $\theta$ which can be broken down as follows

$$[Y, X, \theta] = [Y, X|\theta][\theta].$$  \hspace{1cm} (2.1)

When the objective is primarily to predict $Y$ from $X$, we would be interested in obtaining the conditional distribution $[Y|X]$. In that case it is helpful to model (2.1) hierarchically, that is to break down the joint distribution $[Y, X|\theta]$ into the conditional and marginal parts

$$[Y, X|\theta] = [Y|X, \theta][X|\theta].$$  \hspace{1cm} (2.2)

Such hierarchical modeling approach is extremely useful even if we are not interested in predicting $Y$ from $X$. The reason is that often the modeling of the two individual parts on the right hand side is more easily obtained than the modeling of the joint distribution on the left hand side. For example, we might have have some idea about a physical/biological/chemical/etc. relationship between $Y$ and $X$ and so the modeling of the $Y$ process given the $X$ process could rely on such knowledge. Further discussion of physical-statistical modeling can be found in Berliner (2003) and for two interesting examples of this see Royle et al. (1998) who model wind and pressure fields in a hierarchical way (wind depends on pressure) and Brynjarsdóttir and Berliner (2011) where a model of borehole temperatures given surface temperatures over time is based on the physical model of heat conduction (the heat equation).
Another strength of using the hierarchical approach is that it eliminates the need to specify a joint covariance model for \( Y \) and \( X \) directly. The specification of a valid joint covariance model (i.e. where the covariance matrix is always positive definite) is a well known challenge in modeling multivariate spatial processes. Royle and Berliner (1999) proposed using the hierarchical approach for multivariate spatial processes which avoids this challenge by specifying the conditional and marginal covariance structure that together imply a valid joint covariance model (see also Cressie and Wikle, 2011, Chapter 4.1.5). Also, as Royle and Berliner (1999) demonstrate, some traditional multivariate spatial modeling such as KED (kriging with external drift) are special cases of this hierarchical approach. For further discussions of the strengths of hierarchical modeling see, for example, Berliner (2000) and Cressie et al. (2009).

We now focus on the conditional distribution \( [Y|X, \theta] \). A general, possibly non-linear, model with additive errors can be represented as

\[
Y = f(X) + \varepsilon(X) \tag{2.3}
\]

where \( f(\cdot) \) is an unknown function \( \mathbb{R}^{N_X} \rightarrow \mathbb{R}^{N_Y} \) and \( \varepsilon(X) \) is a random vector of zero-mean errors, that (possibly) depends on \( X \), with some covariance matrix \( \Sigma_\varepsilon \). If the relationship between \( Y \) and \( X \) is linear we can write \( f(X) = FX \) where \( F \) is an unknown \( N_Y \times N_X \) dimensional matrix. Also, from now on we will assume that the error vector does not depend on \( X \). Then the linear version of the model in (2.3) is

\[
Y = FX + \varepsilon. \tag{2.4}
\]

Note that the general models in (2.3) and (2.4) do not necessarily match locations in \( Y \) and \( X \). That is, the observation \( Y_i \) at location \( s_i \) does not necessarily have to be dependent on an observation of \( X \) at that same location. In fact, \( Y \) and \( X \)
do not need to be observed at the same locations and the spatial domains, $D_Y$ and $D_X$, can in principle be quite different. For example, $Y$ and $X$ could be observed on two different grids over the same spatial area or be from two non-overlapping spatial areas.

Note also that if we desire predictions of $Y(s)$ at new locations then $F$ (or $f$) and the covariance structure $\Sigma_c$ need to be modeled in terms of the locations, e.g. their coordinates or the distance between locations, so that the model structure can be extended to new locations.

In the most general form, each $Y_i$ can be modeled as dependent on the entire vector $X$, but such a model is usually too rich in unknowns to be fitted, e.g. $F$ in (2.4) would have $N_Y \times N_X$ unknowns. In practice, $F$ (or $f$) is usually parametrized, which restricts the modeling class.

When $N_Y$ and $N_X$ are large the above models become very difficult to work with (unless the parameters in $F$ or $f$ are very restricted). Hence we consider dimension reduction of $Y$ and $X$. Let $\mathcal{U}$ and $\mathcal{V}$ be $N_Y \times N_Y$ and $N_X \times N_X$ dimensional matrices that contain basis vectors that span $\mathbb{R}^{N_Y}$ and $\mathbb{R}^{N_X}$, respectively. Then any vector in $\mathbb{R}^{N_Y}$ or $\mathbb{R}^{N_X}$ can be written as a linear combination of the corresponding basis vectors. In particular, the observed vectors $Y^{\text{obs}}$ and $X^{\text{obs}}$ can be written as

$$Y^{\text{obs}} = \sum_{k=1}^{N_Y} u_k \alpha_k \quad \text{and} \quad X^{\text{obs}} = \sum_{k=1}^{N_X} v_k \beta_k \quad (2.5)$$

where $u_k$ and $v_k$ are column vectors of $\mathcal{U}$ and $\mathcal{V}$, respectively. The scalars $\alpha_k$ and $\beta_k$ are the coordinates of the vectors $Y^{\text{obs}}$ and $X^{\text{obs}}$ with respect to the bases $\mathcal{U}$ and $\mathcal{V}$, respectively. A dimension reduced representation of $Y^{\text{obs}}$ and $X^{\text{obs}}$ is obtained by
truncating the sums to only the first $K_Y$ and $K_X$ basis vectors, giving the approximations

$$\mathbf{Y}^{\text{obs}} \approx \sum_{k=1}^{K_Y} \mathbf{u}_k \alpha_k = U(\alpha_1, \ldots, \alpha_{K_Y})'$$

and

$$\mathbf{X}^{\text{obs}} \approx \sum_{k=1}^{K_X} \mathbf{v}_k \beta_k = V(\beta_1, \ldots, \beta_{K_X})'$$  \hspace{1cm} (2.6)$$

where $U$ and $V$ contain the first $K_Y$ and $K_X$ column vectors of $\mathbf{U}$ and $\mathbf{V}$. These approximations of data vectors suggest a dimension reduced model for the random vectors $\mathbf{Y}$ and $\mathbf{X}$,

$$\mathbf{Y} = \sum_{k=1}^{K_Y} \mathbf{u}_k a_k + \eta_Y = Ua + \eta_Y \quad \text{and} \quad \mathbf{X} = \sum_{k=1}^{K_X} \mathbf{v}_k b_k + \eta_X = Vb + \eta_X$$  \hspace{1cm} (2.7)$$

where $a$ and $b$ are $K_Y$ and $K_X$ dimensional unknown vectors and the $\eta$’s are random vectors modeled to account for the left over structure. In this setting, elements of the vectors $a$ and $b$ are called *amplitudes*. Note that the modeled amplitudes in $a$ and $b$ need not coincide with the coordinates $\alpha_k$ and $\beta_k$. The dimension reduction will be effective if the first $K_Y$ and $K_X$ basis vectors from $U$ and $V$ capture the most essential structure of $\mathbf{Y}$ and $\mathbf{X}$. In Chapter 3 we will discuss a few options for choosing $U$ and $V$.

Using the dimension reduced models in (2.7), we can write the models in (2.3) and (2.4) as

$$Ua + \eta_Y = f(Vb + \eta_X) + \varepsilon \quad \text{and}$$

$$Ua + \eta_Y = FVb + F\eta_X + \varepsilon, \quad \hspace{1cm} (2.9)$$
respectively. If the $K_Y$ column vectors in $U$ are orthogonal then we have $U'U = I_{K_Y}$ and we can write the above models as

$$a = U' f(V b + \eta_X) + U' \varepsilon - U' \eta_Y \quad \text{and} \quad (2.10)$$

$$a = U' F V b + U' F \eta_X + U' \varepsilon - U' \eta_Y . \quad (2.11)$$

If the column vectors in $U$ are not orthogonal we can still find a matrix, say $U^*$, such that $U^* U = I_{K_Y}$ since the $K_Y$ columns in $U$ are linearly independent. In that case we simply replace $U'$ in (2.10) and (2.11) with $U^*$.

Pooling the error terms in the linear model in (2.11) into $e = U' F \eta_X + U' \varepsilon - U' \eta_Y$ and defining $H = U' F V$ we obtain the following linear model for the amplitudes

$$a = H b + e . \quad (2.12)$$

We have shown that the conditional linear model for $[Y|X, \theta]$ in (2.4) and the dimension reduced models in (2.7) imply a conditional linear model between the amplitude vectors $a$ and $b$. This model for vectors of amplitudes crystallizes the gain of using dimension reduction for $Y$ and $X$. The original linear model in (2.4) has a model matrix $F$ of dimensions $N_Y \times N_X$ while the corresponding model for the amplitudes (2.12) has a model matrix $H$ of dimensions $K_Y \times K_X$. If $K_Y$ and $K_X$ are very small compared to $N_Y$ and $N_X$ the model in (2.12) will be much more manageable. Note that we have not assumed any parametrization of $F$ and so we have not restricted the model in (2.4), beyond assuming linearity. On the other hand, parametrizing $H$ is not straightforward. Furthermore, the variance structure, that is the distribution of $e$, is potentially more complicated than the variance structure in (2.4). Finally, modeling $H$ and the distribution of $e$ directly means that we will not be able to make inference on $F$, $\eta_X$, $\varepsilon$ or $\eta_Y$.  

22
2.1.1 Extension to more than two spatial processes

The hierarchical and dimension reduced modeling framework for two spatial processes discussed above can be generalized to more than two spatial processes. Let $Y_j$ be a random vector that contains the values of a spatial process at $N_j$ locations, for $j = 1, \ldots, n$. The joint distribution of $Y_1, \ldots, Y_n$, given model parameters $\theta$, can be written hierarchically in many ways. As an example we consider the following breakdown of the joint distribution:

$$[Y_n, \ldots, Y_1|\theta] = [Y_n|Y_{n-1}, \ldots, Y_1, \theta] [Y_2|Y_1, \theta] [Y_1|\theta]. \tag{2.13}$$

Parallel to (2.4), suppose we have a linear model for each of the parts in (2.13)

$$Y_j = F_j \left( \begin{array}{c} Y_{j-1} \\ \vdots \\ Y_1 \end{array} \right) + \varepsilon_j, \quad \text{for } j = 2, \ldots, n \quad \text{and} \quad Y_1 = \mu + \varepsilon_1. \tag{2.14}$$

Furthermore, parallel to (2.7), suppose we have a dimension reduction model for each $Y_j$,

$$Y_j = U_j a_j + \eta_j, \quad \text{for } j = 1, \ldots, n. \tag{2.15}$$

Assuming that $U_j$ has orthonormal column vectors for each $j$ then each model in (2.14) can be written as

$$a_j = U_j' F_j \text{diag}(U_{j-1}, \ldots, U_1) \left( \begin{array}{c} a_{j-1} \\ \vdots \\ a_1 \end{array} \right) + U_j' F_j \left( \begin{array}{c} \eta_{j-1} \\ \vdots \\ \eta_1 \end{array} \right) + U_j' \varepsilon_j - \eta_j, \quad \text{for } j = 2, \ldots, n$$

$$a_1 = U_1' \mu + U_1' \varepsilon_1 - \eta_1. \tag{2.16}$$

where $\text{diag}(U_{j-1}, \ldots, U_1)$ is a block diagonal matrix with $U_{j-1}, \ldots, U_1$ on the diagonal. This shows that as for two spatial processes, a hierarchical linear model for $[Y_n, \ldots, Y_1|\theta]$ leads via dimension reduction to a hierarchical linear model for the amplitude vectors, $[a_n, \ldots, a_1|\theta]$. 

23
2.1.2 A Bayesian hierarchical model for two dimension reduced spatial processes

Now we look at how a dimension reduction representation of two spatial processes fits into the Bayesian hierarchical framework in (2.1) and (2.2). We now have two new parameter vectors \( a \) and \( b \) and (2.1) becomes

\[
[Y, X, a, b, \theta] = [Y, X|a, b, \theta] [a, b|\theta] [\theta].
\] (2.17)

We will refer to the three parts on the right hand side as the data model, process model and parameter model following the terminology of Berliner (1996). Note that we did not include the basis vectors in \( U \) and \( V \) in the above notation as they are assumed to be known constants.

The data model can be further broken down as

\[
[Y, X|a, b, \theta] = [Y|X, a, b, \theta] [X|a, b, \theta] \]
\[
= [Y|a, \theta] [X|b, \theta].
\] (2.18) (2.19)

The assumptions made in the last step are that given the amplitudes in \( a \) (and \( \theta \)), \( Y \) is independent of \( X \) and \( b \), and also that given the amplitudes in \( b \) (and \( \theta \)), \( X \) is independent of \( a \). These assumptions of conditional independence are reasonable if the dimension reduction models in (2.7) are a good representation of the structure in \( Y \) and \( X \). In other words, if \( a \) (and \( U \)) captures most of the variation in \( Y \) there is only a small part left to be explained by \( X \) or \( b \). Similarly, if \( b \) (and \( V \)) captures most of the variation in \( X \) then \( a \) has little to contribute. The parts \([Y|b, \theta]\) and \([X|a, \theta]\) are essentially the dimension reduction part and we assume that

\[
[Y|a, \theta] \text{ is } D(Ua, R) \quad \text{and} \quad [X|b, \theta] \text{ is } D(Vb, S)
\] (2.20)
where $\mathcal{D}(\mu, \Sigma)$ stands for some distribution with mean $\mu$ and variance-covariance matrix $\Sigma$. Note that both the amplitude vectors $a$ and $b$ and the variance matrices $R$ and $S$ are unknown. We remark that by taking the mean to be $Ua$ we are assuming that the dimension reduction errors $\eta_Y$ have mean zero. We also assume $\eta_X$ has mean zero. Furthermore, note that the covariance matrices $R$ and $S$ are of dimensions $N_Y \times N_Y$ and $N_X \times N_X$ and due to these high dimensions both $R$ and $S$ require parametrization, for an example see Section 4.5. Finally, we note that the representation in (2.20) masks the many sources of variation that are at play when modeling $Y$ and $X$ conditional on $a$ and $b$. These sources could include measurement error, sub-scale spatial variation (i.e. the nugget effect) as well as error introduced by the dimension reduction. However, we will assume that all these sources of variation are modeled through $R$ and $S$.

The process model, $[a, b|\theta]$, models the association between the two amplitude vectors and it can be developed hierarchically as

$$\begin{align*}
[a, b|\theta] &= [a|b, \theta][b|\theta].
\end{align*}$$

Recalling the linear model for the amplitudes in (2.12) a candidate for $[a|b, \theta]$ is $\mathcal{D}(Hb, C)$ where $C$ represents the covariance matrix of $e$.

### 2.1.3 Mean Square Error of the dimension reduced approach

It is a well-known fact that when we want to predict a random vector $Y$ based on a random vector $X$ the best predictor in terms of mean square error (MSE) is $E(Y|X)$ (Rao, 1973, Section 4g.1). To see this, consider a predictor $h(X)$. Recall
that the squared norm of a vector $\mathbf{x}$ is $||\mathbf{x}||^2 = \mathbf{x}'\mathbf{x}$. The MSE of this predictor is

$$E(||\mathbf{Y} - h(\mathbf{X})||^2) = E(||\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}) + E(\mathbf{Y}|\mathbf{X}) - h(\mathbf{X})||^2)$$

$$= E(||\mathbf{Y} - E(\mathbf{Y}|\mathbf{X})||^2) + E(||E(\mathbf{Y}|\mathbf{X}) - h(\mathbf{X})||^2)$$

$$+ 2E((E(\mathbf{Y}|\mathbf{X}) - h(\mathbf{X}))(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))').$$

(2.22)

The third term is equal to zero since

$$E((E(\mathbf{Y}|\mathbf{X}) - h(\mathbf{X}))(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X})))$$

$$= E((E(\mathbf{Y}|\mathbf{X}) - h(\mathbf{X}))'E((\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))|\mathbf{X})) = 0.$$  

(2.23)

Therefore

$$E(||\mathbf{Y} - h(\mathbf{X})||^2) = E(||\mathbf{Y} - E(\mathbf{Y}|\mathbf{X})||^2) + E(||E(\mathbf{Y}|\mathbf{X}) - h(\mathbf{X})||^2)$$

$$\geq E(||\mathbf{Y} - E(\mathbf{Y}|\mathbf{X})||^2)$$

$$= E(\text{tr}(\text{Var}(\mathbf{Y}|\mathbf{X}))).$$

(2.24)

The lower bound is attained then $h(\mathbf{X}) = E(\mathbf{Y}|\mathbf{X})$.

In our case we do not use $E(\mathbf{Y}|\mathbf{X})$, but rather a dimension reduced approach. The question becomes how much is added to the mean squared error, i.e. how large is the term $E(||E(\mathbf{Y}|\mathbf{X}) - h(\mathbf{X})||^2)$ for the predictor $h(\mathbf{X})$ that we use.

Assume that we have the linear model in (2.4). Then the optimal predictor is

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{F}\mathbf{X}.$$  

(2.25)

Due to the high dimensionality of $\mathbf{F}$ we are not using this predictor. Under the dimension reduced model have $E(\mathbf{Y}|\mathbf{a}) = \mathbf{U}\mathbf{a}$. However, the amplitude vectors $\mathbf{a}$ are not observed. Assuming the linear model in (2.12), we get that $E(\mathbf{a}|\mathbf{b}) = \mathbf{H}\mathbf{b}.$ We
can therefore consider the following predictor for $Y$:

$$E(Y|b) = E\left(E(Y|a, b)|b\right)$$

$$= E(Ua|b) = UE(a|b) = UHb \ . \quad (2.26)$$

In practice, the amplitude vector $b$ is not observed either. However, we consider the following estimator of $b$ using the observable vector $X$:

$$\hat{b}(X) = V'X \ , \quad (2.27)$$

assuming that the columns of $V$ are orthonormal. (If the columns of $V$ are not orthonormal we would replace $V'$ with a matrix $V^*$ for which $V^*V = I$ holds.) The predictor we use can therefore be written as

$$h(X) = UHV'X \ . \quad (2.28)$$

The increase in the MSE from using the predictor in (2.28) can now be seen to be

$$E\left(||E(Y|X) - h(X)||^2\right) = E\left(\left(FX - UHV'X\right)'\left(FX - UHV'X\right)\right)$$

$$= E\left(X'(F - UHV')'(F - UHV')X\right) \ . \quad (2.29)$$

This term depends on the difference between $F$ and $UHV'$, which can be viewed as a basis approximation error. The closer $UHV'$ is to $F$ the better the predictor in (2.28) is, in terms of MSE.

We now consider the expression of MSE for the predictor $E(Y|b) = E(Ua|b) = UHb$. Here we also take into account the dimension reduced model $Y = Ua + \eta_Y$ (see equation 2.7). The MSE can be written as

$$E_{Y,b} \left(||Y - E(Ua|b)||^2\right) = E_{a,\eta_Y,b} \left(||Ua + \eta_Y - E(Ua|b)||^2\right)$$

$$= E_{a,\eta_Y,b} \left(||Ua - E(Ua|b)||^2\right) + 2E_{a,\eta_Y,b} \left(\eta_Y'(Ua - E(Ua|b))\right)$$

$$+ E_{a,\eta_Y,b} \left(||\eta_Y||^2\right) \ . \quad (2.30)$$
The subscripts of the expectation operators are included to emphasize the random variables the expectation refers to. We now take a closer look at the second term in (2.30). Assuming that \([a|\eta, b] = [a|b]\) we get
\[
E_{a,\eta, b} (\eta' (Ua - E(Ua|b))) = E_{\eta, b} (\eta' E_{a|\eta, b} (Ua - E(Ua|b|b, \eta)))
= E_{\eta, b} (\eta' E_{a|b} (Ua - E(Ua|b)))
= E_{\eta, b} (\eta' 0) = 0 .
\] (2.31)

Assuming that \(U'U = I\), the first term in (2.30) is
\[
E_{a,\eta, b} (\|Ua - E(Ua|b)\|^2) = E_{a, b} ((Ua - UE(a|b))'(Ua - UE(a|b)))
= E_{a, b} ((a - E(a|b))'U'U(a - E(a|b)))
= E_{a, b} ((a - E(a|b))'(a - E(a|b))) .
\] (2.32)

Recall that \(E(\eta_Y) = 0\). The third term in (2.30) can therefore be expressed as
\[
E_{a, \eta, b} (\|\eta_Y\|^2) = E_{\eta} ((\eta_Y)'\eta_Y)) = \sum_{i=1}^{N_Y} Var(\eta_{Y_i}) .
\] (2.33)

We have shown that under the dimension reduction model in (2.7), the MSE for the predictor \(E(Y|b)\) is the MSE of the optimal predictor of \(a\) from \(b\), \(E(a|b)\), plus the sum of the variances of the dimension reduction error terms \(\eta_{Y_i}\):
\[
E (\|Y - E(Ua|b)\|^2) = E (\|a - E(a|b)\|^2) + \sum_{i=1}^{N_Y} Var(\eta_{Y_i}) .
\] (2.34)

This expression highlights that to get a low MSE we need both a good dimension reduction, i.e. small variance of \(\eta_{Y_i}\), and good prediction of the \(a\) amplitudes.

### 2.2 Two spatio-temporal processes

Now we consider two spatio-temporal processes \(Y(s, t)\) and \(X(c, t)\) where \(s\) and \(c\) are locations in the spatial domains \(D_Y\) and \(D_X\), respectively; \(s \in D_Y \subset \mathbb{R}^{d_Y}\) and
The processes $Y$ and $X$ might not be observed at the same time, for example in the statistical downscaling example in Chapter 4, the ERA-40 data ($X$) are monthly data while the Polar MM5 data ($Y$) are seasonal. However, for ease of notation we first assume that both $Y$ and $X$ are observed at the same discrete time points $t = 1, \ldots, T$ and consider the case of misaligned temporal observations in Section 2.3.

Let $Y_t$ and $X_t$ be $N_Y$ and $N_X$ dimensional random vectors that contain $Y(s_{i(t)}, t)$ and $X(c_{j(t)}, t)$ for $i(t) = 1, \ldots, N_Y$ and $j(t) = 1, \ldots, N_X$. If we have the same number of observations at every time point then the $Y_t$ and $X_t$ vectors will be $N_Y$ and $N_X$ dimensional for every $t$.

Our goal is to model jointly $Y_t$ and $X_t$ for all time points, via the hierarchical approach. The joint distribution is

$$\left[ Y_{1:T}, X_{1:T}, \theta \right] = \left[ Y_{1:T}, X_{1:T} \mid \theta \right] \left[ \theta \right]$$

where the vector $\theta$ represents all unknown parameters and the index notation ‘1 : $T$’ stands for all time points from 1 to $T$ i.e., $Y_{1:T}$ and $X_{1:T}$ are data matrices. Next, we can write $\left[ Y_{1:T}, X_{1:T} \mid \theta \right]$ as a product of conditional distributions according to the natural chronological order, that is:

$$\left[ Y_{1:T}, X_{1:T} \mid \theta \right] = \left[ Y_T, X_T \mid Y_{1:T-1}, X_{1:T-1}, \theta \right] \cdots \left[ Y_2, X_2 \mid Y_1, X_1, \theta \right] \left[ Y_1, X_1 \mid \theta \right]$$

$$= \left[ Y_T \mid Y_{1:T-1}, X_{1:T}, \theta \right] \left[ X_T \mid Y_{1:T-1}, X_{1:T-1}, \theta \right]$$

$$\cdots \left[ Y_2 \mid Y_1, X_{1:2}, \theta \right] \left[ X_2 \mid Y_1, X_1, \theta \right] \left[ Y_1 \mid X_1, \theta \right] \left[ X_1 \mid \theta \right]$$

(2.36)

where in the second step we have for each time point $t$ further broken down the joint distribution of $Y_t$ and $X_t$ as

$$\left[ Y_t, X_t \mid Y_{1:t-1}, X_{1:t-1}, \theta \right] = \left[ Y_t \mid Y_{1:t-1}, X_{1:t}, \theta \right] \left[ X_t \mid Y_{1:t-1}, X_{1:t-1}, \theta \right].$$

(2.37)
Note that no simplifying assumptions about conditional independence have been made at this point.

Consider the two parts on the right hand side of (2.37). As for (2.3) we define a general model for \[ Y_t | Y_{1:t-1}, X_{1:t}, \theta \], with additive errors as

\[
Y_t = f_t(X_{0:t}, Y_{0:(t-1)}) + \varepsilon_y, \quad t = 1, \ldots, T ,
\] (2.38)

where \( f_t(\cdot) \) is an unknown function from \( \mathbb{R}^{\sum_{i=0}^t N_{X_i} + \sum_{i=0}^{t-1} N_{Y_i}} \) to \( \mathbb{R}^{N_{Y_t}} \). Due to the recursive nature of the hierarchical model we introduced two initial parameter vectors \( X_0 \) and \( Y_0 \). These can be viewed as the two spatial processes at time 0, i.e. from a time point before the processes are observed. In the notation of (2.36) and (2.37), \( X_0 \) and \( Y_0 \) are a part of the parameter vector \( \theta \). The linear form of (2.38) can be written as

\[
Y_t = F_t \begin{bmatrix} X_0 \\ \vdots \\ X_t \\ Y_0 \\ \vdots \\ Y_{t-1} \end{bmatrix} + \varepsilon_y, \quad t = 1, \ldots, T .
\] (2.39)

The model matrix \( F_t \) is \( (\sum_{i=0}^t N_{X_i} + \sum_{i=0}^{t-1} N_{Y_i}) \times N_{Y_t} \) dimensional. This form is usually too general so often we make do with a simpler structure. For example, assuming the \( Y_t \) process is conditionally independent of past \( X \) and \( Y \), the above models can be reduced to

\[
Y_t = f_t(X_t) + \varepsilon_y \quad \text{or} \quad Y_t = F_t X_t + \varepsilon_y .
\] (2.40, 2.41)

The model matrix, \( F_t \), is \( N_{Y_t} \times N_{X_t} \) dimensional for all \( t \).
We also need to formulate the distributions \([X_t | Y_{1:t-1}, X_{1:t-1}, \theta]\). We envision general models of the form

\[ X_t = g_t(X_{0:t-1}, Y_{0:(t-1)}) + \varepsilon_{X_t}, \quad t = 1, \ldots, T \text{ or} \]

\[ X_t = G_t \begin{bmatrix} x_0 \\ \vdots \\ x_{t-1} \\ y_0 \\ \vdots \\ y_{t-1} \end{bmatrix} + \varepsilon_{X_t}, \quad t = 1, \ldots, T, \quad (2.43) \]

where \(g_t(\cdot)\) are unknown functions from \(\mathbb{R}^{\sum_{i=0}^{t-1} N_{X_i} + \sum_{i=0}^{t-1} N_{Y_i}}\) to \(\mathbb{R}^{N_{X_t}}\) and \(G_t\) are unknown transition matrices of dimensions \((\sum_{i=0}^{t-1} N_{X_i} + \sum_{i=0}^{t-1} N_{Y_i}) \times N_{X_t}\). If we assume a first-order Markovian structure, then \([X_t | Y_{1:t-1}, X_{1:t-1}, \theta]\) only depends on \(X_{t-1}:\)

\[ X_t = g_t(X_{t-1}) + \varepsilon_{X_t}, \quad t = 1, \ldots, T \text{ or} \]

\[ X_t = G_t X_{t-1} + \varepsilon_{X_t}, \quad t = 1, \ldots, T. \quad (2.45) \]

The model in (2.45) is often called a first-order linear dynamical spatio-temporal model (for discussion and further references see Cressie and Wikle, 2011).

Now we turn attention to how we can use a dimension reduced approach within the linear models (2.39) and (2.43). Let \(U_t\) and \(V_t, t = 0, \ldots, T\), be vector bases that span \(\mathbb{R}^{N_{Y_t}}\) and \(\mathbb{R}^{N_{X_t}}\), respectively and let \(U_t\) and \(V_t\) be matrices that contain the first \(K_{Y_t}\) and \(K_{X_t}\) column vectors of \(U_t\) and \(V_t\), respectively. As in (2.7) we define dimension reduction models for both \(Y_t\) and \(X_t:\)

\[ Y_t = \sum_{k=1}^{K_{Y_t}} u_{t,k} a_{t,k} + \eta_{Y_t} = U_t a_t + \eta_{Y_t}, \quad t = 0, \ldots, T \quad \text{and} \quad (2.46) \]

\[ X_t = \sum_{k=1}^{K_{X_t}} v_{t,k} b_{t,k} + \eta_{X_t} = V_t b_t + \eta_{X_t}, \quad t = 0, \ldots, T. \quad (2.47) \]

Note that included in (2.46) and (2.47) are dimension reduction model for the initial processes \(X_0\) and \(Y_0\) so that we also have initial amplitude vectors, \(a_0\) and \(b_0\). The
linear model for \([\mathbf{Y}_t | \mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t}, \theta]\) in (2.39) can be written as

\[
U_t \mathbf{a}_t + \eta_{Y_t} = F_t \begin{pmatrix}
V_0 b_0 + \eta_{X_0} \\
\vdots \\
V_t b_t + \eta_{X_t} \\
U_0 a_0 + \eta_{Y_0} \\
\vdots \\
U_{t-1} a_{t-1} + \eta_{Y_{t-1}}
\end{pmatrix} + \varepsilon_{Y_t} = F_t W_t \begin{pmatrix}
b_0 \\
\vdots \\
b_t \\
a_0 \\
\vdots \\
a_{t-1}
\end{pmatrix} + F_t \begin{pmatrix}
\eta_{X_0} \\
\vdots \\
\eta_{X_t} \\
\eta_{Y_0} \\
\vdots \\
\eta_{Y_{t-1}}
\end{pmatrix} + \varepsilon_{Y_t} \quad (2.48)
\]

where \(W_t\) is the block diagonal matrix

\[
W_t = diag(V_0, \ldots, V_t, U_0, \ldots, U_{t-1}) .
\] (2.49)

If the column vectors of \(U_t\) are orthonormal, we get the following model for the \(a_t\) amplitude vectors:

\[
a_t = U_t' F_t W_t \begin{pmatrix}
b_0 \\
\vdots \\
b_t \\
a_0 \\
\vdots \\
a_{t-1}
\end{pmatrix} + U_t' F_t \varepsilon_{Y_t} - U_t' \eta_{Y_t} .
\] (2.50)

So, the linear model in (2.39) together with the dimension reduction models in (2.46) and (2.47) imply a model for the \(a_t\) amplitudes for each time point \(t\):

\[
a_t = U_t' F_t V_t \begin{pmatrix}
b_0 \\
\vdots \\
b_t \\
a_0 \\
\vdots \\
a_{t-1}
\end{pmatrix} + \varepsilon_{Y_t} \quad (2.51)
\]

where \(\varepsilon_{Y_t}\) represents a collection of error terms. Note that the simpler model \(\mathbf{Y}_t = F_t \mathbf{X}_t + \varepsilon_{Y_t}\), corresponds to the following model for the amplitudes:

\[
a_t = U_t' F_t V_t b_t + \varepsilon_{Y_t} = H_t b_t + \varepsilon_{Y_t}\] (2.52)

where \(H_t = U_t' F_t V_t\). As in Section 2.1, if the column vectors of \(U_t\) are not orthonormal we can still find a matrix \(U_t^*\) so that \(U_t^* U_t = I_{K_Y}\), and replace \(U_t\) in (2.50) - (2.52) with \(U_t^*\).
Similarly, the linear model for \([X_t | Y_{1:t-1}, X_{1:t-1}, \theta]\) in (2.43) together with the dimension reduced models in (2.46) and (2.47) can be written as

\[
V_t b_t + \eta_{X,t} = G_t \widetilde{W}_t \left( \begin{array}{c} b_0 \\ \vdots \\ b_{t-1} \\ a_0 \\ \vdots \\ a_{t-1} \end{array} \right) + \widetilde{W}_t \left( \begin{array}{c} \eta_{X_0} \\ \vdots \\ \eta_{X_{t-1}} \\ \eta_{Y_0} \\ \vdots \\ \eta_{Y_{t-1}} \end{array} \right) + \epsilon_{X_t}
\]

(2.53)

where \(\widetilde{W}_t\) is the block diagonal matrix

\[
\widetilde{W}_t = \text{diag}(V_0, \ldots, V_{t-1}, U_0, \ldots U_{t-1})
\]

(2.54)

If the column vectors of \(V_t\) are orthonormal, we get the following model for the \(b_t\) amplitude vectors

\[
b_t = V_t' G_t \widetilde{W}_t \left( \begin{array}{c} b_0 \\ \vdots \\ b_{t-1} \\ a_0 \\ \vdots \\ a_{t-1} \end{array} \right) + V_t' \epsilon_{X_t} - V_t' \eta_{X_t}
\]

(2.55)

otherwise we replace \(V_t'\) with a matrix \(V_t^*\) where \(V_t^* V_t = I_{K_X}\). So again, the linear model in (2.43) together with the dimension reduction models in (2.46) and (2.47) imply a model for the \(b_t\) amplitude vectors:

\[
b_t = V_t' G_t \widetilde{W}_t \left( \begin{array}{c} b_0 \\ \vdots \\ b_{t-1} \\ a_0 \\ \vdots \\ a_{t-1} \end{array} \right) + e_{X_t}
\]

(2.56)

where \(e_{X_t}\) represents a collection of error terms. Note that a first-order Markovian model \(X_t = G_t X_{t-1} + \epsilon_{X_t}\), corresponds to the following model for the amplitudes:

\[
b_t = V_t' G_{t-1} b_{t-1} + e_{X_t} = B_t b_{t-1} + e_{X_t}
\]

(2.57)

The \(B_t\) matrices are called \textit{transition matrices}.
2.2.1 A Bayesian hierarchical model for two dimension reduced spatio-temporal processes

Parallel to Section 2.1.2, we consider the dimension reduced models in (2.46) and (2.47) and the models for amplitude vectors within the Bayesian Hierarchical framework in (2.35). We now have amplitude vectors $a_t$ and $b_t$ for each time point so the joint distribution in (2.35) can be written as

$$[Y_{1:T}, X_{1:T}, a_{0:T}, b_{0:T}, \theta] = [Y_{1:T}, X_{1:T}|a_{0:T}, b_{0:T}, \theta] [a_{0:T}, b_{0:T}|\theta] [\theta],$$

(2.58)

where, as in Section 2.1.2, we refer to the three parts on the right hand side as the data model, process model and parameter model, respectively.

Making the assumption that the processes $Y_1, \ldots, Y_T, X_1, \ldots, X_T$ are conditionally independent, given the corresponding amplitudes, the data model can be broken down as

$$[Y_{1:T}, X_{1:T}|a_{0:T}, b_{0:T}, \theta] = \prod_{t=1}^{T} [Y_t|a_t, \theta] [X_t|b_t, \theta].$$

(2.59)

The individual distributions can be modeled as a dimension reduction model and we take

$$[Y_t|a_t, \theta] \text{ as } D(U_t a_t, R_t) \quad \text{ and } \quad [X_t|b_t, \theta] \text{ as } D(V_t b_t, S_t), \quad t = 1, \ldots, T,$$

(2.60)

where, as before, $D(\mu, \Sigma)$ stands for some distribution with mean $\mu$ and variance-covariance matrix $\Sigma$. Note that we do not need to include the initial parameter vectors $X_0$ and $Y_0$ in the model, but rather the initial amplitude vectors, $a_0$ and $b_0$. The same comments made about $R$ and $S$ after equation (2.20) are valid for $R_t$ and $S_t$.

Although we allow for time-dependent basis vectors, $U_t$ and $V_t$, we may not want to have different basis vectors at every time point. The reason being that when we use
data-dependent basis vectors (such as those that will be discussed in Section 3) we usually do not have enough data to obtain basis vectors for every time point. In the downscaling example in Chapter 4, for example, we have different basis vectors for each season or month, but they stay the same between years. One potential problem with using the same \( N_Y \) dimensional basis vectors, say \( U \), for more than one time point arises if \( Y_t \) is not observed at the same locations for these time points or there are missing observations. Then the dimensions of \( Y_t \) and \( U \) may not match for some \( t \). A possible way around this is to take
\[
[Y_t|a_t, \theta] \text{ as } D(K_t U a, R_t)
\]
where \( K_t \) is a \( N_Y \times N_Y \) dimensional specified matrix. For example, if there are missing observations in \( Y_t \) then \( K_t \) could be taken as the incidence matrix having the \((i,j)\)th element equal to 1 if the location of the \( i \)th observation in \( Y_t \) corresponds to the \( j \)th location in \( U \) and having 0’s everywhere else (see for example Section 7.1.2 in Cressie and Wikle, 2011). Similar strategies could be applied to the \( X_t \) process.

The process model can be broken down as
\[
[a_{0:T}, b_{0:T}|\theta] = [a_0|\theta] [b_0|\theta] \prod_{t=1}^T [a_t|a_{0:(t-1)}, b_{0:t}, \theta] [b_t|a_{0:(t-1)}, b_{0:(t-1)}, \theta]. \quad (2.61)
\]
A straightforward choice for each part, \([a_t|a_{0:t}, b_{0:t}, \theta]\) and \([b_t|a_{0:(t-1)}, b_{0:(t-1)}, \theta]\) is to model them based on (2.51) and (2.56). As a simple example, we could make the assumption that the process model can be written as
\[
[a_{0:T}, b_{0:T}|\theta] = [a_0|\theta] [b_0|\theta] \prod_{t=1}^T [a_t|b_t, \theta] [b_t|b_{t-1}, \theta], \quad (2.62)
\]
that is, assume that $a_t$ is independent of past amplitudes (given $b_t$) and that $b_t$ only depends on $b_{t-1}$, conditionally. Then, with (2.52) and (2.57) in mind, we can take

$$
[a_t|b_t, \theta] \text{ as } D(H_t b_t, C_t) \text{ for } t = 1, \ldots, T \text{ and } \nonumber
$$

$$
[b_t|b_{t-1}, \theta] \text{ as } D(B_t b_{t-1}, D_t) \text{ for } t = 1, \ldots, T . \tag{2.63}
$$

The distributions for the initial amplitude vectors, $[a_0|\theta]$ and $[b_0|\theta]$, need to be modelled separately.

The parameter model $[\theta]$ is bound to be problem specific and we defer to the statistical downscaling application in Chapter 4 for an example.

### 2.3 Two spatio-temporal processes misaligned in time

Consider the situation where the two processes are observed at different time points or different temporal scales, e.g. monthly versus seasonally. Assume $Y(s, t)$ is observed at time points $t_1, \ldots, t_{T_Y}$ and $X(c, \tau)$ is observed at time points $\tau_1, \ldots, \tau_{T_X}$ where the $t_i$’s and $\tau_j$’s span more or less the same time period. How these processes interact between time scales is bound to be problem specific. But even so, the most natural hierarchical modeling approach would be to model the present given the past.

Data sampled at different frequencies are commonly encountered and hierarchical modeling approaches are known to be a useful strategy in combining such data (see, for example, Wikle et al., 2001; Berliner et al., 2003). Here we briefly explore an example where $X_{\tau}$ is observed more frequently than $Y_t$ (see also Chapter 4) through a hierarchical modeling perspective. Assuming that between each $t_{i-1}$ and $t_i$ there is at least one $\tau_j$ so that $t_{i-1} < \tau_j \leq t_i$ we can break down the joint distribution similar
to (2.36)

\[
[Y_{t_1:t_{T_Y}}, X_{\tau_1:\tau_{T_X}} | \theta] = \prod_{i=1}^{T_Y} [Y_{t_i} | Y_{t_1:t_{i-1}}, X_{\tau_1:\tau_j+}] \prod_{j \in \tau(i)} [X_{\tau_j} | Y_{t_1:t_{i-1}}, X_{\tau_1:\tau_j-1}]
\]

where \( \tau(i) = \{ j : t_{i-1} < \tau_j \leq t_i \} \) and \( j^+ = \max(\tau(i)) \). In other words, the notation \([Y_{t_i} | Y_{t_1:t_{i-1}}, X_{\tau_1:\tau_j+}]\) stands for the conditional distribution of \( Y_{t_i} \) given all past \( Y_t \)’s and both past and (possibly) present \( X_r \)’s. The notation \([X_{\tau_j} | Y_{t_1:t_{i-1}}, X_{\tau_1:\tau_j-1}]\) stands for the conditional distribution of \( X_{\tau_j} \) given all the past \( Y_t \)’s and \( X_r \)’s. Linear models akin to (2.39) and (2.43) along with dimension reduced models \( Y_{t_i} = U_{t_i} a_{t_i} + \eta_{Y,t_i} \) and \( X_{\tau_j} = V_{\tau_j} b_{\tau_j} + \eta_{X,\tau_j} \) will still imply models for the amplitudes \( a_{t_i} \) and \( b_{\tau_j} \) in the same manner as in Section 2.2 and can still be used in a Bayesian hierarchical framework.

The real challenge here is to determine a simpler dependence structure as we did in (2.41) and (2.45). One option is to adopt a similar approach as in Section 2.2, that is to model the \( Y_t \) processes conditional on \( X_r \) processes that are close in time and then model the temporal evolution on only the \( X_r \) processes. For example, we could assume that given the \( X_r \)’s between the last and present time point, \( Y_{t_i} \) is independent of earlier \( X_r \)’s and all past \( Y_t \)’s, i.e.

\[
[Y_{t_i} | Y_{t_1:t_{i-1}}, X_{\tau_1:\tau_{T_X}}] = [Y_{t_i} | X_{\tau_{j-}:\tau_{j+}}] , \tag{2.64}
\]

where \( j^- = \min(\tau(i)) \). Also, we could assume that \( X_{\tau_j} \) is dependent on only the previous \( X_{\tau_j-1} \), i.e.

\[
[X_{\tau_j} | Y_{t_1:t_{i-1}}, X_{\tau_1:\tau_{j-1}}] = [X_{\tau_j} | X_{\tau_{j-1}}] \tag{2.65}
\]

We pursue an application of statistical downscaling in Chapter 4 where the two processes have different temporal scales (seasonally versus monthly) and where we apply similar modeling assumptions.
2.4 Spatio-temporal processes with continuous spatial domains

In foregoing Sections the spatial, or spatio-temporal, processes are assumed to be observed at discrete locations. These observed processes can therefore be collected into random vectors $\mathbf{Y}$ and $\mathbf{X}$ which then are indexed by time where the time steps have also been assumed to be discrete. Indeed, most spatial and spatio-temporal data are collected at discrete locations and times. On the other hand, the processes that are being observed are usually continuous both in space and time. That is, $Y(s,t)$ and $X(c,\tau)$ actually live on continuous spatial domains, $s \in D_Y \subset \mathbb{R}^{d_Y}$ and $c \in D_X \subset \mathbb{R}^{d_X}$, and continuous time intervals, $t \in [t_1, t_T]$ and $\tau \in [\tau_1, \tau_T]$. The boundaries of the spatial domains and time intervals are often only determined by our focus of interest.

In this section we discuss dimension reduced approaches to modeling of two spatio-temporal processes where the spatial domains are continuous. However, we still assume that the temporal domain is discrete, and more over, that $Y(s,t)$ and $X(c,t)$ are observed at the same time points, $t = 1, \ldots, T$. First, we consider one spatio-temporal process, $X(c,t)$ and then expand to two spatio-temporal processes. The case of one spatio-temporal process was developed by Wikle and Cressie (1999). Wikle and Cressie (1999) develop an approach to space-time prediction that involves modeling the spatial process in terms of basis functions. Furthermore, the coefficients of the basis functions are temporally linked and expressed in a state-space formulation. They show how a space-time model defined on a continuous space can be converted into a multivariate temporal model of the coefficients of the basis functions. Here
we present a brief overview of some of the ideas in Wikle and Cressie (1999), but we change the notation in concert with the notation used in Sections 2.1 - 2.3.

Wikle and Cressie (1999) assume that the spatio-temporal process $X(c, t)$ evolves in time according to the state equation

$$X(c, t) = \int_{D_X} g_c(u)X(u, t - 1)du + \varepsilon_X(c, t)$$  \hspace{1cm} (2.66)

where $\varepsilon_X$ is a spatial error process and $g_c(u)$ is a function ($g_c(u) : D_X \rightarrow \mathbb{R}$) representing the interaction between the process $X$ at two locations in $D_X$, $c$ and $u$, at times $t$ and $t - 1$. To assure stationarity over time it is assumed that $|\int_{D_X} g_c(u)du| < 1$.

The model in (2.66) is a first-order Markov model akin to the last model in (2.45).

Wikle and Cressie (1999) then introduce a dimension reduction by assuming that $X(c, t)$ can be decomposed into $K_X$ dominant components

$$X(c, t) = \sum_{k=1}^{K_X} \phi_k(c)b_k(t) \hspace{1cm} \text{for} \hspace{1cm} c \in D_X \hspace{1cm} \text{and} \hspace{1cm} t \in \{1, 2, \ldots\}$$  \hspace{1cm} (2.67)

where $b_k(\cdot)$ are zero-mean time series and $\phi_i(\cdot)$ are deterministic basis functions that are complete and orthonormal. We note two differences between the representation in (2.67) and the dimension reduced model in (2.47). First, the basis functions $\phi_i(\cdot)$ are defined over the whole spatial domain $D_X$ whereas the basis vectors $v_k$ are only defined on the $N_X$ locations where $X$ is observed. Second, the truncation to $K_X$ basis functions in (2.67) is assumed to be exact, but the model in (2.47) allows left over variation, modelled via $\eta_{Xt}$.

Since $\{\phi_i(\cdot); i = 1, \ldots\}$ is a complete basis we can write

$$g_c(u) = \sum_{i=1}^{\infty} q_i(c)\phi_i(u)$$  \hspace{1cm} (2.68)

Substituting (2.67) and (2.68) into (2.66), we get a model in terms of $b_k(\cdot)$’s:

$$\phi(c)'b_t = q(c)'b_{t-1} + \varepsilon(s, t)$$  \hspace{1cm} (2.69)
where
\[ b_t = (b_1(t), \ldots, b_{K_X}(t))', \quad q(c) = (q_1(c), \ldots, q_{K_X}(c))' \quad \text{and} \]
\[ \phi(c) = (\phi_1(c), \ldots, \phi_{K_X}(c))'. \] (2.70)

In order to turn (2.69) into a state-space model for the \( b_t \) vector we need to go to discrete locations. Assuming we have data obtained at \( N_X \) locations, \( c_1, \ldots, c_{N_X} \) then the model in (2.69) in matrix form is
\[ \Phi b_t = Q b_{t-1} + \varepsilon_{X_t} \] (2.71)

where \( \varepsilon_{X_t} = (\varepsilon_{X}(c_1, t), \ldots, \varepsilon_{X}(c_{N_X}, t))' \), \( \Phi \) is an \( N_X \times K_X \) matrix with rows \( \phi(c_i)' \) and \( Q \) is an \( N_X \times K_X \) matrix with rows \( q(c_i)' \) for \( i = 1, \ldots, N_X \). From (2.71) we get
\[ b_t = (\Phi'\Phi)^{-1} \Phi' Q b_{t-1} + (\Phi'\Phi)^{-1} \Phi' \varepsilon_{X_t} \]
\[ = \Phi' Q b_{t-1} + \Phi' \varepsilon_{X_t}, \] (2.72)

where the second line arises if \( \Phi'\Phi = I_{K_X} \). This is similar to the model we got in (2.57). Even though it looks like we ended up in the same place, we note that the advantage of having basis functions \( \phi_i(\cdot) \) defined over the whole spatial domain is that it enables predictions at new spatial locations. The data-dependent basis vectors discussed in Chapter 3 and used in the statistical downscaling examples in Chapter 4 are only defined at the locations where data are observed, but in that example the goal is not to predict at new locations.

2.4.1 Extension to two space-time processes

Now we consider two spatio-temporal processes \( Y(s, t) \) and \( X(c, t) \) where \( s \in D_Y \subset \mathbb{R}^{d_Y} \) and \( c \in D_X \subset \mathbb{R}^{d_X} \) and \( t = 1, \ldots, T \). There are several ways one could
envision extensions to the previously described work in Wikle and Cressie (1999). Here we consider an avenue akin to the models in (2.41) and (2.45), i.e. we model \( Y \) conditional on \( X \) at the same time point but \( X \) is modeled as a first order Markov process just as in (2.66):

\[
Y(s, t) = \int_{D_X} f_s(u)X(u, t)du + \varepsilon_Y(s, t) \quad \text{and} \quad (2.73)
\]

\[
X(c, t) = \int_{D_X} g_c(u)X(u, t - 1)du + \varepsilon_X(c, t) \quad (2.74)
\]

The \( \varepsilon_Y \) and \( \varepsilon_X \) are spatial model error processes, \( g_c(u) \) is a function \((g_c : D_X \to \mathbb{R})\) representing the interaction between the \( X \) process at two locations at times \( t \) and \( t - 1 \) and \( f_s(u) \) is a function \((f_s : D_X \to \mathbb{R})\) representing the interaction between the \( Y \) and \( X \) processes at two locations but at the same time. To assure stationarity over time we assume that \(|\int_{D_X} f_s(u)du| < 1\) and \(|\int_{D_X} g_c(u)du| < 1\) (see Wikle and Cressie, 1999).

We introduce dimension reduction in the same way as Wikle and Cressie (1999) by assuming that the two processes can the decomposed into \( K_Y \) and \( K_X \) dominant components:

\[
Y(s, t) = \sum_{k=1}^{K_Y} \psi_k(s)a_k(t) = \psi(s)'a_t \quad \text{and} \quad (2.75)
\]

\[
X(c, t) = \sum_{k=1}^{K_X} \phi_k(c)b_k(t) = \phi(c)'b_t \quad (2.76)
\]

where

\[
\psi(s) = (\psi_1(s), \ldots, \psi_{K_Y}(s))', \quad \phi(c) = (\phi_1(c), \ldots, \phi_{K_X}(c))',
\]

\[
a_t = (a_1(t), \ldots, a_{K_Y}(t))', \quad b_t = (b_1(t), \ldots, b_{K_X}(t))',
\]

and both \( \{\phi_k(c); k = 1, 2, \ldots\} \) and \( \{\psi_k(s); k = 1, 2, \ldots\} \) are orthonormal and complete basis functions. Note that these are bases for two different spaces; \( \{\phi_k(c); k = 1, 2, \ldots\} \)
is a basis for the space, \( C_X \), of functions with domain \( D_X \), but \( \{ \psi_k(s); k = 1, 2, \ldots \} \) is a basis for the space, \( C_Y \), of functions with domain \( D_Y \).

Since both \( f_s(\cdot) \in C_X \) and \( g_c(\cdot) \in C_X \) and the basis \( \{ \phi_k(c); k = 1, 2, \ldots \} \) is complete we can write both functions in terms of \( \phi \)'s:

\[
f_s(u) = \sum_{i=1}^{\infty} p_i(s) \phi_i(u) \quad \forall s \in D_Y \text{ and } \forall u \in D_X \quad \text{and} \quad (2.78)
\]

\[
g_c(u) = \sum_{i=1}^{\infty} q_i(c) \phi_i(u) \quad \forall c, u \in D_X . \quad (2.79)
\]

Substituting (2.76) and (2.78) into (2.73) we get for all \( s \in D_Y \),

\[
Y(s, t) = \int_{D_X} f_s(u) X(u, t) du + \varepsilon_Y(s, t)
\]

\[
= \int_{D_X} \left( \sum_{i=1}^{K_X} p_i(s) \phi_i(u) \right) \left( \sum_{k=1}^{K_X} \phi_k(u) b_k(t) \right) du + \varepsilon_Y(s, t)
\]

\[
= \int_{D_X} \sum_{k=1}^{K_X} b_k(t) \left[ \sum_{i=1}^{\infty} \phi_k(u) p_i(s) \phi_i(u) \right] du + \varepsilon_Y(s, t)
\]

\[
= \sum_{k=1}^{K_X} b_k(t) \left[ \sum_{i=1}^{\infty} p_i(s) \int_{D_X} \phi_k(u) \phi_i(u) du \right] + \varepsilon_Y(s, t) . \quad (2.80)
\]

The interchange of integral and infinite sum in the last step holds if the basis functions are continuous (see for example Casella and Berger, 2002, Section 2.4), since the series \( \sum_{i=1}^{\infty} \phi_k(u) p_i(s) \phi_i(u) \) converges to \( p_i(s) f_s(u) \) for all \( u \). Since \( \{ \phi_k(c); k = 1, 2, \ldots \} \) is an orthonormal basis we have \( \int_{D_X} \phi_k(u) \phi_i(u) du = 1 \) if \( k = i \) but zero otherwise, so the term in the square brackets is simply \( p_k(s) \) and we get

\[
Y(s, t) = \sum_{k=1}^{K_X} b_k(t) p_k(s) + \varepsilon_Y(s, t) = p(s)' b_t + \varepsilon_Y(s, t) \quad (2.81)
\]

where \( p(s) = (p_1(s), \ldots, p_{K_X}(s))' \). Using (2.75) for the left hand side we can now write

\[
\psi(s)' a_t = p(s)' b_t + \varepsilon_Y(s, t) \quad (2.82)
\]
Similarly, by substituting (2.76) and (2.79) into (2.74) we get for all \( c \in D_X \):

\[
X(c, t) = \int_{D_X} g_c(u) X(u, t - 1) du + \varepsilon_X(c, t)
\]

\[
= \int_{D_X} \left( \sum_{i=1}^{\infty} q_i(c) \phi_i(u) \right) \left( \sum_{k=1}^{K_X} \phi_k(u) b_k(t - 1) \right) du + \varepsilon_X(c, t)
\]

\[
= \int_{D_X} \sum_{k=1}^{K_X} b_k(t - 1) \left[ \sum_{i=1}^{\infty} \phi_k(u) q_i(c) \phi_i(u) \right] du + \varepsilon_X(c, t)
\]

\[
= \sum_{k=1}^{K_X} b_k(t - 1) q_k(c) + \varepsilon_Y(c, t) = q(c)'b_{t-1} + \varepsilon_X(c, t),
\]

where \( q(c) = (q_1(c), \ldots, q_{K_Y}(c))' \). By using (2.76) we can write

\[
\phi(c)'b_t = q(c)'b_{t-1} + \varepsilon_X(c, t).
\]

In summary, by assuming the dimension reduction in (2.75) and (2.76) we have turned the spatio-temporal model of the \( Y \) processes in (2.73) and (2.74) into an equivalent model of the \( K_Y \) and \( K_X \) dimensional vectors \( a_t \) and \( b_t \):

\[
\psi(s)'a_t = p(s)'b_t + \varepsilon_Y(s, t) \quad \forall s \in D_Y \quad \text{and} \quad \forall s \in D_Y
\]

\[
\phi(c)'b_t = q(c)'b_{t-1} + \varepsilon_X(c, t) \quad \forall c \in D_X.
\]

Assume that at each time point \( t \) we have observations of \( Y \) at discrete locations \( s_1, \ldots, s_{N_Y} \) and of \( X \) at locations \( c_1, \ldots, c_{N_X} \). The model in (2.85) and (2.86) for all locations is

\[
\Psi_{N_Y \times K_Y} a_t = P_{N_Y \times K_X} b_t + \varepsilon_Y t \quad \text{and} \quad \Phi_{N_X \times K_X} b_t = Q_{N_X \times K_X} b_{t-1} + \varepsilon_X t
\]
where \( \varepsilon_{Y_t} = (\varepsilon_Y(s_1, t), \ldots, \varepsilon_Y(s_{N_Y}, t))^\prime \) and \( \varepsilon_{X_t} = (\varepsilon_X(c_1, t), \ldots, \varepsilon_X(c_{N_X}, t))^\prime \). Here, \( \Psi \) is an \( N_Y \times K_Y \) dimensional matrix with rows \( \psi(s_i)^\prime \) and \( P \) is an \( N_Y \times K_X \) dimensional matrix with rows \( p(s_i)^\prime \) for \( i = 1, \ldots, N_Y \). The \( N_X \times K_X \) dimensional matrices \( \Phi \) and \( Q \) have rows \( \phi(c_j)^\prime \) and \( q(c_j)^\prime \) respectively for \( j = 1, \ldots, N_X \). From this we obtain a state-space model

\[
a_t = (\Psi'\Psi)^{-1}\Psi'Pb_t + (\Psi'\Psi)^{-1}\Psi'\varepsilon_{Y_t} \\
= \Psi'Pb_t + \Psi'\varepsilon_{Y_t} ; \tag{2.88}
\]

\[
b_t = (\Phi'\Phi)^{-1}\Phi'Qb_{t-1} + (\Phi'\Phi)^{-1}\Phi'\varepsilon_{X_t} \\
= \Phi'Qb_{t-1} + \Phi'\varepsilon_{X_t} , \tag{2.89}
\]

where the second step in each case holds if \( \Psi'\Psi = I_{K_Y} \) and \( \Phi'\Phi = I_{K_X} \).

In most cases a dimension reduced representation as in (2.75) and (2.76) does not hold exactly and there is a left over structure that could be modelled through a random process. More realistic dimension reduced models are

\[
Y(s, t) = \sum_{k=1}^{K_Y} \psi_k(s)a_k(t) = \psi(s)^\prime a_t + \eta_Y(s, t) \quad \text{and} \quad (2.90)
\]

\[
X(c, t) = \sum_{k=1}^{K_X} \phi_k(c)b_k(t) = \phi(c)^\prime b_t + \eta_X(c, t), \tag{2.91}
\]

as in (2.46) and (2.46). The effect of using these models instead of (2.75) and (2.76) in the above derivation is that we would need to add the term

\[
\int_{D_X} \eta_X(u, t)f_s(u)du = \int_{D_X} \eta_X(u) \sum_{i=1}^{\infty} p_i(c)\psi_i(u)du \tag{2.92}
\]

in (2.81) and add the term

\[
\int_{D_X} \eta_X(u, t - 1)g_c(u)du = \int_{D_X} \eta_X(u, t - 1) \sum_{i=1}^{\infty} q_i(c)\phi_i(u)du \tag{2.93}
\]

in (2.83). However, these terms cannot easily be simplified.
2.5 Bayesian Filtering

Bayesian inference for unknown parameters in (2.58) is based on the joint posterior distribution

$$[a_{0:T}, b_{0:T}, \theta | Y_{1:T}, X_{1:T}],$$

(2.94)

where $\theta$ contains all unknown parameters other than the amplitude vectors $a_{0:T}$ and $b_{0:T}$. When we also want predictions of future observations for, say, $\tau$ time steps ahead, we base our predictions on the posterior predictive distribution

$$[Y(T+1):(T+\tau), X(T+1):(T+\tau) | Y_{1:T}, X_{1:T}],$$

(2.95)

which can be obtained as the marginal distribution of

$$[Y(T+1):(T+\tau), X(T+1):(T+\tau), a_{0:(T+\tau)}, b_{0:(T+\tau)}, \theta | Y_{1:T}, X_{1:T}].$$

(2.96)

Analytical expression for these distributions are usually intractable. Instead, we obtain samples via MCMC methods and base our inference and predictions on these samples. We use that approach in Chapter 4. In this section, however, we consider an alternative approach to inference in the case where the data and process models are defined as a linear state space model as in (2.60) and (2.63), namely Bayesian filtering.

Bayesian filtering (see for example Wikle and Berliner, 2007; Doucet et al., 2000) is a procedure that recursively produces the distributions

$$[a_1, b_1, \theta | Y_1, X_1], [a_2, b_2, \theta | Y_{1:2}, X_{1:2}], \ldots, [a_T, b_T, \theta | Y_{1:T}, X_{1:T}].$$

(2.97)

Inference based on these distributions is not equivalent to inference based on the joint posterior in (2.94) since $[a_t, b_t, \theta | Y_{1:t}, X_{1:t}]$ ignores information about $a_t$ and $b_t$ (and
\(\theta\) in “future” data \(Y_{(t+1):T}\) and \(X_{(t+1):T}\). But, when our focus is on inference on the current amplitudes, \(a_T\) and \(b_T\), or prediction of future \(Y\) and \(X\), the filtering approach is sufficient. Note that it is assumed here that the parameter vector \(\theta\) does not depend on \(t\). That means that the covariance matrices \(R\) and \(S\) in (2.60), the covariance matrices \(C\) and \(D\), and the transition matrices \(H\) and \(B\) in (2.63) do not vary with time.

Using the filtering approach we obtain the distribution \([a_t, b_t, \theta|Y_{1:t}, X_{1:t}]\) based on \([a_{t-1}, b_{t-1}, \theta|Y_{1:(t-1)}, X_{1:(t-1)}]\) (for \(t > 1\)) in two steps, a forecasting step and an analysis step. The forecasting step is

\[
[a_t, b_t, \theta|Y_{1:t-1}, X_{1:t-1}] = \int [a_t, b_t, a_{t-1}, b_{t-1}, \theta|Y_{1:t-1}, X_{1:t-1}] da_{t-1} db_{t-1}
\]

\[
= \int [a_t, b_t|a_{t-1}, b_{t-1}, \theta] [a_{t-1}, b_{t-1}, \theta|Y_{1:t-1}, X_{1:t-1}] da_{t-1} db_{t-1}, \tag{2.98}
\]

where the second equality follows from the fact that given \(a_{t-1}\) and \(b_{t-1}\) (and \(\theta\)), \(a_t\) and \(b_t\) are independent of past data (see Lemma A.1.6 in the Appendix). Note that the prior, \([a_t, b_t|a_{t-1}, b_{t-1}, \theta]\), and the current posterior, \([a_{t-1}, b_{t-1}, \theta|Y_{1:t-1}, X_{1:t-1}]\), are known distributions but the integral is usually not analytically tractable.

Next we update the posterior from the forecasting step via the following analysis step:

\[
[a_t, b_t, \theta|Y_{1:t}, X_{1:t}] = [a_t, b_t, \theta|Y_{1:t-1}, Y_t, X_{1:t-1}, X_t]
\]

\[
\propto [Y_t, X_t|a_t, b_t, Y_{1:t-1}, X_{1:t-1}] [a_t, b_t|Z_{1:t-1}]
\]

\[
\propto [Y_t, X_t|a_t, b_t, \theta] [a_t, b_t, \theta|Y_{1:t-1}, X_{1:t-1}], \tag{2.99}
\]

where the last step follows from fact that given \(a_t\) and \(b_t\), \(Y_t\) and \(X_t\) are independent of past data (see Lemma A.1.7 in the Appendix). The data model \([Y_t, X_t|a_t, b_t, \theta]\)
is known and \([a_t, b_t, \theta|Y_{1:t-1}, X_{1:t-1}]\) is the forecasting distribution from (2.98) so again, in principle, the “analysis” distribution in (2.99) can be worked out.

Although in principle we can obtain the forecasting and analysis distributions in (2.98) and (2.99) the reality is usually quite different. The problem lies in that in general \(\theta\) is not known and that the first posterior distribution \([a_1, b_1, \theta|Y_1, X_1]\) is not available in closed form. However, when the distributions in (2.60) and (2.63) are normal and \(\theta\) is known, then both the forecasting distribution, \([a_t, b_t|Y_{1:t-1}, X_{1:t-1}]\), and the updated posterior distribution \([a_t, b_t|Y_{1:t}, X_{1:t}]\), are normal with mean vectors and covariance matrices equal to the classical Kalman filter updates, see for example West and Harrison (1997) and Meinhold and Singpurwalla (1983), and Theorem A.1.5 in the Appendix. The Kalman filter in this setting is given in the next Section 2.5.1.

### 2.5.1 Kalman Filter

Suppose we have the model structure in (2.59) where

\[
[Y_t|a_t] \sim N(U_t a_t, R_t) \quad \text{and} \quad [X_t|b_t] \sim N(V_t b_t, S_t), \quad t = 1, \ldots, T. \tag{2.100}
\]

and \(U_t, V_t, R_t\) and \(S_t\) are known. Also, suppose we have the model structure in (2.62) and that

\[
[a_t|b_t, \theta] \sim N(H_t b_t, C_t) \quad \text{for} \quad t = 1, \ldots, T \quad \text{and} \quad [b_t|b_{t-1}, \theta] \sim N(B_t b_{t-1}, D_t), \quad \text{for} \quad t = 1, \ldots, T \tag{2.101}
\]
where $H_t$, $B_t$, $D_t$ and $D_t$ are known. Then (2.100) and (2.101) can be represented as the following linear state space model

$$\begin{align*}
\begin{pmatrix} Y_t \\ X_t \end{pmatrix} | a_t, b_t &\sim N \left( \begin{pmatrix} a_t \\ b_t \end{pmatrix}, \Sigma_t \right) \quad \text{and} \\
\begin{pmatrix} a_t \\ b_t \end{pmatrix} | b_{t-1}, a_{t-1} &\sim N \left( M_t \begin{pmatrix} a_{t-1} \\ b_{t-1} \end{pmatrix}, \Gamma_t \right)
\end{align*}
$$

(2.102)

(2.103)

where

$$\begin{align*}
\Phi_t &= \begin{pmatrix} U_t & 0 \\ 0 & V_t \end{pmatrix} \quad \text{and} \quad \Sigma_t = \begin{pmatrix} R_t & 0 \\ 0 & S_t \end{pmatrix}, \\
M_t &= \begin{pmatrix} 0 & H_t B_t \\ 0 & B_t \end{pmatrix} \quad \text{and} \quad \Gamma_t = \begin{pmatrix} C_t + H_t D_t H_t' & H_t D_t \\ D_t H_t' & D_t \end{pmatrix}
\end{align*}
$$

(2.104)

(2.105)

(see Theorem A.1.1 in the Appendix). The model presented in (2.102) and (2.103) is also sometimes called a dynamic linear model (West and Harrison, 1997).

Here I will provide the Kalman filter updates (Kalman, 1960; Jazwinski, 1970; Anderson and Moore, 1979) for the model in (2.102) and (2.103). Let $\mu^a_t$, $P^a_t$, $\mu^f_t$ and $P^f_t$ be the mean vectors and covariance matrices of the distributions in the analysis and forecasting steps (2.99) and (2.98):

$$\begin{align*}
\mu^a_{t-1} &= \begin{pmatrix} \mu^a_{1,t-1} \\ \mu^a_{2,t-1} \end{pmatrix} = E \left( \begin{pmatrix} a_{t-1} \\ b_{t-1} \end{pmatrix} | Y_{1:t-1}, X_{1:t-1} \right) \\
P^a_{t-1} &= \begin{pmatrix} P^a_{11,t-1} \\ P^a_{12,t-1} \\ P^a_{21,t-1} \\ P^a_{22,t-1} \end{pmatrix} = Cov \left( \begin{pmatrix} a_{t-1} \\ b_{t-1} \end{pmatrix} | Y_{1:t-1}, X_{1:t-1} \right) \\
\mu^f_t &= \begin{pmatrix} \mu^f_{1,t} \\ \mu^f_{2,t} \end{pmatrix} = E \left( \begin{pmatrix} a_t \\ b_t \end{pmatrix} | Y_{1:t-1}, X_{1:t-1} \right) \\
P^f_t &= \begin{pmatrix} P^f_{11,t} \\ P^f_{12,t} \\ P^f_{21,t} \\ P^f_{22,t} \end{pmatrix} = Cov \left( \begin{pmatrix} a_t \\ b_t \end{pmatrix} | Y_{1:t-1}, X_{1:t-1} \right)
\end{align*}
$$

(2.106)

(2.107)

(2.108)

(2.109)

where the partition of $\mu^a_{t-1}$, $P^a_{t-1}$, $\mu^f_t$ and $P^f_t$ are in accordance to the dimensions of the $a$'s and $b$'s. Kalman filter updates provide a way to obtain $\mu^a_t$ and $P^a_t$ from $\mu^a_{t-1}$
and $P_{t-1}^a$ in two steps:

Forecasting Step: $\mu_f^t = M_t \mu_{t-1}^a$ and $P_f^t = \Gamma_t + M_t P_{t-1}^a M_t'$ (2.110)

Analysis Step: $\mu_a^t = \mu_f^t + K_t \left( (\mu_y^t) - \Phi_t \mu_f^t \right)$ (2.111)

and $P_a^t = (I - K_t \Phi_t) P_f^t$, (2.112)

where $K_t = P_f^t \Phi_t' (\Phi_t P_f^t \Phi_t' + \Sigma_t)^{-1}$ is called the Kalman Gain Matrix.

The forecasting step becomes

$$
\mu_f^t = M_t \mu_{t-1}^a = \begin{pmatrix} 0 & H_t B_t \\ 0 & B_t \end{pmatrix} \begin{pmatrix} \mu_{1,t-1}^a \\ \mu_{2,t-1}^a \end{pmatrix} = \begin{pmatrix} H_t B_t \mu_{2,t-1}^a \\ B_t \mu_{2,t-1}^a \end{pmatrix} (2.113)
$$

and

$$
P_f^t = \Gamma_t + M_t P_{t-1}^a M_t' = \Gamma_t + \begin{pmatrix} 0 & H_t B_t \\ 0 & B_t \end{pmatrix} \begin{pmatrix} P_{11,t-1}^a & P_{12,t-1}^a \\ P_{21,t-1}^a & P_{22,t-1}^a \end{pmatrix} \begin{pmatrix} 0 & 0 \\ B_t' H_t' B_t' \end{pmatrix} = \Gamma_t + \begin{pmatrix} H_t B_t P_{22,t-1}^a B_t' H_t' & H_t B_t P_{22,t-1}^a B_t' \\ B_t P_{22,t-1}^a B_t' & B_t P_{22,t-1}^a \end{pmatrix} (2.114)
$$

From this we can see that we do not need the full $K_Y + K_X$ dimensional vector $\mu_a^t$ and the $(K_Y + K_X) \times (K_Y + K_X)$ dimensional matrix $P_{t-1}^a$ to get the forecasting distribution, only the $K_X$ dimensional vector $\mu_{2,t-1}^a$ and the $K_X \times K_X$ dimensional matrix $P_{22,t-1}^a$. This, of course, is due to the fact that only the $b$ amplitudes model the temporal evolution.

For the analysis step we need the Kalman gain matrix

$$
K_t = P_f^t \Phi_t' (\Phi_t P_f^t \Phi_t' + \Sigma_t)^{-1} (2.115)
$$

$$
= P_f^t \Phi_t' \left( \Sigma_t^{-1} - \Sigma_t^{-1} \Phi_t (\Phi_t' \Sigma_t^{-1} \Phi_t + P_f^t)^{-1} \Phi_t' \Sigma_t^{-1} \right) (2.116)
$$

The Kalman Gain matrix is of dimensions $(K_Y + K_X) \times (N_Y + N_X)$ and involves an inverse of an $(N_Y + N_X) \times (N_Y + N_X)$ dimensional matrix. In the expression in (2.115)
an inverse of this dimension needs to be calculated in each Kalman Filter update. The equivalent expression in (2.116) also involves an inverse of an \((N_Y + N_X) \times (N_Y + N_X)\) dimensional matrix, \(\Sigma_t\). However, if \(\Sigma_t^{-1}\) is known and constant over time then the Kalman Filter update only requires an inverse of a \((K_Y + K_X) \times (K_Y + K_X)\) dimensional matrix which can save a considerable amount of computing time. Furthermore, the full \(\Sigma_t^{-1}\) matrix does not need to be stored, it is sufficient to only keep the \((K_Y + K_X) \times (N_Y + N_X)\) dimensional matrix \(\Phi_t'\Sigma^{-1}\) and the \((K_Y + K_X) \times (K_Y + K_X)\) dimensional matrix \(\Phi_t'\Sigma^{-1}\Phi_t\).
Chapter 3: Data-Dependent Basis Vectors

In this Chapter I demonstrate how the multivariate analysis techniques, Maximum Covariance Analysis (MCA, Section 3.2) and Canonical Correlation Analysis (CCA, Section 3.3) can be used to obtain data-dependent basis vectors for the joint modeling of two space-time processes within the hierarchical framework of Chapter 2. Furthermore, I develop a new approach to obtain data-dependent basis vectors that I call Maximum Covariance Patterns (MCP), in Section 3.4. An orthogonalized version is developed in Section 3.5. First, I briefly review the use of Empirical Orthogonal Functions (EOFs) in dimension-reduced modeling of one space-time process.

3.1 Empirical Orthogonal Functions

In EOF analysis, or Principal Components Analysis (PCA), we aim to explain the variance-covariance structure of a set of random variables through a few linear combinations. Let $\mathbf{Y}$ be an $N_Y$-dimensional random vector with mean $\mathbf{0}$ and covariance matrix $\Sigma_Y$. For example, $\mathbf{Y}$ could be a spatial process observed at $N_Y$ locations. The first principal component is defined as the linear combination $A_1 = \mathbf{u}_1' \mathbf{Y}$ that has the largest variance. The $i^{th}$ principal component is a linear combination $A_i = \mathbf{u}_i' \mathbf{Y}$ that has the largest variance among all those that are uncorrelated with the first $i - 1$ principal components. This maximization of the variance is performed subject to the
constraint that the principal component vectors are of unit length, i.e. $u_i^t u_i = 1$ for all $i = 1, \ldots, N_Y$. The solution is obtained by the spectral decomposition of $\Sigma_Y$ (see for example Johnson and Wickern (2002), Ch. 8):

$$\Sigma_Y = \mathbf{U} \Lambda \mathbf{U}' .$$

(3.1)

The eigenvectors (columns of $\mathbf{U}$) are the principal component vectors and the eigenvalues (diagonal of $\Lambda$) are the variances of the principal components, $\lambda_i = Var(A_i)$. Since $\text{tr}(\Sigma_Y) = \text{tr}(\Lambda)$ the proportion of the total variance due to (explained by) the $i$th principal component is

$$q_i = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_N} .$$

(3.2)

In the Earth Sciences principal component vectors are usually called EOFs or EOF patterns and the principal components, $A_i$, are called EOF amplitudes. From here on out I will only use the name EOF analysis. EOF analysis is used in many disciplines and has been popular in climatic sciences to identify patterns of so-called simultaneous variations, or modes of variability, in climatic fields (see, for example, Zhang et al., 1997; Esteban-Parra et al., 1998).

The eigenvectors in $\mathbf{U}$ form an orthonormal basis of $\mathbb{R}^{N_Y}$. Hence, they can be used in a dimension reduced modeling approach for a spatio-temporal process similar to the approach in Section 2.2 (see also Section 1.1). That is, we can define the data model as $Y_t = U_t a_t + \eta_t$ where $U_t$ is the first $K_Y$ eigenvectors of $\Sigma_{Y_t}$ and then we can define the process model as a first-order Markov model, $a_t = H_t a_{t-1} + \gamma_t$. See also Chapter 7.1.3 in Cressie and Wikle (2011).

In practice, the covariance matrix $\Sigma_{Y_t}$ is unknown and estimating a covariance matrix for each time point is typically infeasible. The approach that is usually taken
is to treat the data, $Y_{t}^{\text{obs}}$, observed at different time points as repeated measurements. Then the EOFs are obtained from the sample covariance matrix. Suppose we have $T$ observations of $Y$ and let $Y$ be the $N_Y \times T$ centered data matrix. The centering is temporal (not spatial) in that for each location $s_i$ the $i^{\text{th}}$ row of $Y$ sums to zero ($\sum_{t=1}^{T} y_{it} = 0$). Then the sample covariance matrix is

$$S_Y = \frac{1}{T-1} YY'. \quad (3.3)$$

The sample EOFs are an orthonormal basis for $\mathbb{R}^{N_Y}$ and can be used in dimension reduced modeling of $Y_t$.

Note that even if we have fewer time points than locations ($T < N_Y$), implying that $S_Y$ will be singular, we can obtain EOF patterns. In that case only the first $T$ eigenvalues can be positive and the rest are zero; and the first $T$ EOFs will explain all the total sample variance. Therefore the upper limit for choosing $K_Y$ will be $T$ but not $N_Y$. Note also that we do not have to actually calculate the $N_Y \times N_Y$ dimensional matrix $S_Y$ to obtain the EOFs, it suffices to perform SVD analysis of $Y$ (or $\frac{1}{\sqrt{T-1}} Y$) which is computationally efficient if $N_Y$ is much larger than $T$ (see Section A.1.3 in the Appendix).

The way we estimate the covariance matrix in (3.3) assumes that the locations are the same at every time point. The same assumption will apply in following sections when we estimate cross-covariance matrices. If each location is not observed at all $T$ time points we must estimate the covariance matrix a different way, but the EOF analysis continues as before by eigen decomposition of the estimated covariance matrix. How to obtain an estimate of the covariance matrix would need to be addressed.

---

Note that the observations (time points) are the columns of $Y$ and the locations are the rows. Usually in multivariate statistics data matrices have observations in rows and variables (i.e. locations) in columns.
on a case-by-case basis. For example, if a few observations are missing at random we could input missing data by interpolating data from neighboring locations and then use the estimator in (3.3). Note that these interpolated data values would not be included in the Bayesian model but rather we would use an indicator matrix as discussed in Section 2.2.1.

One appealing aspect of using EOFs as basis vectors is the ordering of them in terms of total (sample) variance. That is, the first EOF amplitude $A_1$ has the highest variance and therefore $\mathbf{u}_1 A_1$ captures the largest possible part of the total variance in $\mathbf{Y}$. Then adding the next EOF, $\mathbf{u}_2$, we have captured more of the total variance and so on. So by using the first $K_Y$ EOFs we have (hopefully) captured the most essential spatial structure in $\mathbf{Y}$ and the proportion of total (sample) variance explained by the first $K_Y$ EOFs (see equation 3.2) can be used as a guideline for how many EOFs to keep. It is important to keep in mind that the proportion of sample variance explained is an average over the $T$ time points. We can expect the chosen $K_Y$ EOFs to be a good representation of the spatial structure at some time points and not so good at other time points. Also, at each time point different EOFs may be important, i.e. at each time point only some of the chosen EOFs may be active in explaining the spatial structure. Furthermore, since EOFs are constructed to capture spatial structure, a critical concern is whether the EOFs, or rather their amplitudes, provide a good representation of the dynamics of the spatio-temporal process.

EOFs have been successfully used in modeling of space-time processes. Wikle and Cressie (1999) apply this approach to Near-Surface winds over the tropical Pacific ocean (see also Section 2.4). Berliner et al. (2000) also apply this technique using a fully Bayesian approach and predict sea-surface temperatures in the tropical Pacific
ocean with a 7 month lead time. For further discussion of the use of EOFs in modeling of a spatio-temporal process see Chapter 7.1.3 in Cressie and Wikle (2011).

3.2 Maximum Covariance Analysis

Let $Y$ and $X$ be $N_Y$ and $N_X$ dimensional random vectors. Recall from Section 1.2.1.2 that the objective in Maximum Covariance Analysis (MCA) is to maximize covariances of linear combinations of the random vectors under constraints that the vectors have to be mutually orthonormal i.e.,

$$\text{maximize } Cov(u'_i Y, v'_i X) = u'_i \Sigma_{YX} v_i$$
subject to $u'_i u_i = 1$, $v'_i v_i = 1$, $u'_i u'_j = 0$, $v'_i v'_j = 0$, for $i \neq j$. \hspace{0.5cm} (3.4)

As stated in Section 1.2.1.2 the MC vectors $u_i$ and $v_i$ are obtained via SVD of the cross-covariance matrix,

$$\Sigma_{YX} = \tilde{U} \tilde{D} \tilde{V}'. \hspace{0.5cm} (3.5)$$

The left and right singular vectors, the *MC patterns*, in $\tilde{U}$ and $\tilde{V}$ form orthonormal bases that span $\mathbb{R}^{N_Y}$ and $\mathbb{R}^{N_X}$ and can be used in a dimension reduced model as described in Section 2.2.

The sample MC patterns are obtained by SVD analysis on the sample cross covariance matrix. Suppose we have $T$ observations of both $Y_t$ and $X_t$ and let $Y$ and $X$ be $N_Y \times T$ and $N_X \times T$ dimensional (temporally) centered data matrices. Then the SVD of the sample cross covariance matrix can be written as

$$S_{YX} = \frac{1}{T-1} XY' = UDV'; \hspace{0.5cm} (3.6)$$
where the first singular vectors in $U$ and $V$ correspond to the largest singular value. We can use the first few column vectors in $U$ and $V$ as basis vectors for our model.
Note that here we assume that $Y_t$ and $X_t$ are observed at an equal number of time points ($T$). Hence, this method of using MC patterns for basis vectors is not directly applicable when we have different temporal resolutions in the $Y$ and $X$ data. The new approach, MCP (see Section 3.4), is more flexible in this regard. Furthermore, note that the number of non-zero singular values (the diagonal of $D$) is no more than the minimum of $T$, $N_Y$ and $N_X$. Therefore, when the number of time points is less than the number of locations, as is the case in the statistical downscaling example in Chapter 4, the upper limit for choosing the number of basis vectors to use (i.e. $K_Y$ and $K_X$) is $T$.

Since the MC patterns maximize the covariance between the amplitudes they are an appealing choice of basis vectors for a hierarchical model that models the $Y_t$ process given the $X_t$ process, as in Section 2.2. When the basis vectors are set up so that the amplitudes have a strong statistical relationship (e.g., high correlation or high covariance), we have set the stage for a good predictive model; i.e. the model for the $a_t$ and $b_t$ amplitudes. This appealing aspect could potentially also be a downside for the following reason. Even though $A_{t,i} = u_i Y_t$ and $B_{t,i} = v_i X_t$, for say $i = 1, \ldots, K$, have high covariance it is not clear whether the dimension reduction representations of each field,

$$Y_t = \sum_{k=1}^{K} u_i a_{t,k} + \eta_{Yt} \quad \text{and} \quad X_t = \sum_{k=1}^{K} v_i b_{t,k} + \eta_{Xt},$$

are “good” models. In other words, it is not clear that the most essential spatial structures in each field are necessarily captured by the MC patterns which are constructed to capture a common structure. One way to partially address this issue is to look at the proportion of total variance of $Y$ and $X$ (separately) that is explained by the first $K_Y$ and $K_X$ MC patterns. We can look at the fitted values $\hat{Y}_t = U \hat{a}_t = UU' Y_t^{obs}$
and $\hat{X}_t = V \hat{b}_t = V V' X_t^{obs}$ where $U$ and $V$ contain the first $k$ vectors in $U$ and $V$. Let $\hat{Y}_k$ and $\hat{X}_k$ be $N_Y \times T$ and $N_X \times T$ matrices containing $\hat{Y}_t$ and $\hat{X}_t$ calculated with $k$ MC patterns. Consider the total variance of the sample covariance matrix of the dimension reduced representation as a proportion of the total variance of the original sample covariance matrix for each $k$:

$$\frac{\text{tr}(S_{\hat{Y}_k \hat{Y}_k})}{\text{tr}(S_{YY})} = \frac{\text{tr}(\hat{Y}_k \hat{Y}_k')}{\text{tr}(YY')}, \quad \text{and} \quad \frac{\text{tr}(S_{\hat{X}_k \hat{X}_k})}{\text{tr}(S_{XX})} = \frac{\text{tr}(\hat{X}_k \hat{X}_k')}{\text{tr}(XX')}.$$  \hspace{1cm} (3.8)

Note that for any matrix $M$ we have that $\text{tr}(MM')$ is the sum of the squared singular values of $M$ which makes the calculation of the ratios in (3.8) easy. We also note that even though the ratios in (3.8) increase for every $k$, the increments are not necessarily decreasing with $k$ as in (3.2). Finally, we note that using the ratios in (3.8) are only guidelines and that choosing the number of basis vectors is ultimately a model selection problem.

### 3.3 Canonical Correlation Analysis

I will briefly show that a cousin of MCA, Canonical Correlation Analysis (CCA), can be used to obtain basis vectors in a similar way. Recall from Section 1.2.1.1 that the objective in CCA is to maximize the correlation of linear combinations of random vectors $Y$ and $X$. The constraints are that the linear combinations are uncorrelated:

$$\text{maximize } \text{Cor}(A_i, B_i) = \text{Cor}(u_i'Y, v_i'X) = u_i'YX v_i$$

subject to $u_i'u_i = 1, \ v_i'v_i = 1, \ \text{Cor}(A_i, A_j) = 0, \ \text{Cor}(B_i, B_j) = 0$ for $i \neq j$. \hspace{1cm} (3.9)

As in Section 3.2 let $Y$ and $X$ be $N_Y \times T$ and $N_X \times T$ dimensional (temporally) centered data matrices containing $T$ observations of $Y_t$ and $X_t$. The sample canonical vectors,
\( \mathbf{u}_i \) and \( \mathbf{v}_i \) are obtained by first performing SVD of the following matrix

\[
S_{YY}^{-1/2} S_{YX} S_{XX}^{-1/2} = \hat{U} D \hat{V}'
\]  

(3.10)

where

\[
S_{YY} = \frac{1}{T - 1} Y Y', \quad S_{YX} = \frac{1}{T - 1} Y X', \quad S_{YY} = \frac{1}{T - 1} X X'.
\]  

(3.11)

The sample canonical vectors are the columns of

\[
\mathbf{U} = S_{YY}^{-1/2} \hat{U} \quad \text{and} \quad \mathbf{V} = S_{XX}^{-1/2} \hat{V}.
\]  

(3.12)

The sample CC patterns are the columns of

\[
S_{YY} \mathbf{U} = S_{YY}^{1/2} \hat{U} \quad \text{and} \quad S_{XX} \mathbf{V} = S_{XX}^{1/2} \hat{V}.
\]  

(3.13)

The CC patterns form normalized (but not orthogonal) bases for \( \mathbb{R}^{N_Y} \) and \( \mathbb{R}^{N_X} \) and the first \( K_Y \) and \( K_X \) patterns can be used as basis vectors within the modeling framework of Chapter 2. Note however that to obtain the CC patterns we need the inverse square-root of a matrices \( S_{YY}^{-1/2} \) and \( S_{XX}^{-1/2} \). The inverse square-root matrix \( S_{YY}^{-1/2} \) is defined as \( S_{YY}^{-1/2} = PE^{-1/2}P' \) where \( PEP' \) is the spectral decomposition of \( S \) and \( E^{-1/2} \) is a diagonal matrix with diagonal elements equal to the inverses of the square-roots of the diagonal elements in \( E \). This matrix is only defined when \( S_{YY} \) is of full rank which causes problems when \( T < N_Y \). Similarly, \( S_{XX}^{-1/2} \) is not defined if \( T < N_X \). Therefore the CCA approach is not as widely applicable as MCA.

As for MCA the appealing aspect of this approach is the maximization of correlations between amplitudes. Furthermore, the amplitudes are uncorrelated between pairs. But, as for MCA, the question becomes whether these vectors that are designed to capture the common structure are any good at capturing the spatial structure of
the two individual fields. Again we can look at the total variance of the sample covariance matrix of the dimension reduced representation as a proportion of the total variance of the original sample covariance matrix for each $k$:

$$\frac{\text{tr}(S_{\hat{Y}_k \hat{Y}_k})}{\text{tr}(S_{YY})} = \frac{\text{tr}(\hat{Y}_k \hat{Y}_k')}{\text{tr}(YY')} \quad \text{and} \quad \frac{\text{tr}(S_{\hat{X}_k \hat{X}_k})}{\text{tr}(S_{XX})} = \frac{\text{tr}(\hat{X}_k \hat{X}_k')}{\text{tr}(XX')} .$$ (3.14)

where $\hat{Y}_k$ and $\hat{X}_k$ are $N_Y \times T$ and $N_X \times T$ matrices containing $\hat{Y}_t$ and $\hat{X}_t$ calculated with $k$ CC patterns, i.e. $\hat{Y}_t = U \hat{a}_t = UU'Y_t^{obs}$ and $\hat{X}_t = V \hat{b}_t = VV'X_t^{obs}$ where $U$ and $V$ contain the first $K_Y$ and $K_X$ column vectors of the $U$ and $V$ in (3.13).

### 3.4 Maximum Covariance Patterns

As discussed in Sections 3.2 and 3.3, one of the concerns with using MCA (or CCA) to obtain basis vectors for the modeling framework in Chapter 2 is that even if the first few vectors in $U$ and $V$ capture most of the common spatial structure of the $Y$ and $X$ processes (i.e. the amplitudes have high covariance or correlation), there is no guaranty that the $U$ and $V$ vectors are a good representation of the individual fields. If our primary objective is to predict the $Y$ process from observations of the $X$ process, we may only be concerned by whether we capture the spatial structure of $Y$. In other words, we just want the patterns in $X$ that are good at predicting the patterns that make up the spatial structure in $Y$. Suppose we have patterns $u_1, \ldots, u_{K_Y}$ that capture the most essential structure in $Y$, for example EOFs. The question then becomes whether we can find $N_X$-dimensional vectors, $v_1, \ldots, v_{K_X}$, so that the amplitudes $b_k = v'_kX$ are good at predicting the amplitudes $a_k = u'_kY$. The approach developed here is to find for each vector $u_k$, a vector $v_k$ that maximizes the covariance of $\text{Cov}(u'_kY, v'_kX)$.

We begin with a general theorem and a corollary.
Theorem 3.4.1. Let \( \mathbf{c} \) be a known vector in \( \mathbb{R}^N \). The solution to the maximization problem

\[
\begin{align*}
\text{maximize } f(\mathbf{v}) &= \mathbf{c}'\mathbf{v} \\
\text{subject to } \mathbf{v}'\mathbf{v} &= 1
\end{align*}
\] (3.15)

is

\[
\mathbf{v}^* = \frac{\mathbf{c}}{||\mathbf{c}||},
\] (3.16)

where \( ||\mathbf{c}|| = \sqrt{\mathbf{c}'\mathbf{c}} \) is the Euclidean length of the vector \( \mathbf{c} \). The vector that minimizes \( f(\mathbf{v}) \) subject to the same constraint is \( -\frac{\mathbf{c}}{||\mathbf{c}||} \).

Proof. The objective function \( f(\mathbf{v}) \) is the inner product of the vectors \( \mathbf{v} \) and \( \mathbf{c} \). Let \( \theta \) be the angle between the vectors \( \mathbf{v} \) and \( \mathbf{c} \). A well-known geometry fact is that

\[
f(\mathbf{v}) = \mathbf{c}'\mathbf{v} = \mathbf{c} \cdot \mathbf{v} = ||\mathbf{c}|| ||\mathbf{v}|| \cos(\theta).
\] (3.17)

Since \( \mathbf{c} \) is fixed and we have the constraint that \( ||\mathbf{v}|| = 1 \) it is obvious that \( f(\mathbf{v}) \) is maximized when \( \cos(\theta) = 1 \), that is when \( \theta = 0 \). That in turn means that the optimal unit-length vector has to be in the same direction as is \( \mathbf{c} \) and so \( \mathbf{v}^* = \mathbf{c}/||\mathbf{c}|| \) and \( f(\mathbf{v}^*) = ||\mathbf{c}|| \). Also, it is obvious that \( f(\mathbf{v}) \) is minimized when \( \cos(\theta) = -1 \), or \( \theta = \pi \), and so the unit-length vector that minimizes \( f(\mathbf{v}) \) has to be in the opposite direction from \( \mathbf{c} \), that is \( -\mathbf{c}/||\mathbf{c}|| = -\mathbf{v}^* \) and \( f(-\mathbf{v}^*) = -||\mathbf{c}|| \). An alternative (and longer) proof is offered in Appendix A.2. \( \square \)

Corollary 3.4.2. Let \( \mathbf{Y} \) and \( \mathbf{X} \) be \( N_Y \) and \( N_X \) dimensional random vectors with cross-covariance matrix \( \text{Cov}(\mathbf{Y}, \mathbf{X}) = \Sigma_{YX} \). Let \( \mathbf{u} \) be a fixed \( N_Y \)-dimensional vector. Let then the linear combination \( \mathbf{v}'\mathbf{X} \) that maximizes the covariance \( \text{Cov}(\mathbf{u}'\mathbf{Y}, \mathbf{v}'\mathbf{X}) \), with
the constraint that \( \mathbf{v} \) is of length 1, is
\[
\mathbf{v}^* = \frac{\Sigma_{XY} \mathbf{u}}{||\Sigma_{XY} \mathbf{u}||}.
\] (3.18)

Proof. We have that \( \text{Cov}(\mathbf{u}' \mathbf{Y}, \mathbf{v}' \mathbf{X}) = (\mathbf{u}' \Sigma_{YX}) \mathbf{v} \). The setup here is therefore the same as in Theorem 3.4.1 with \( \mathbf{c}' = \mathbf{u}' \Sigma_{YX} \). It follows that the vector that maximizes the covariance, subject to being of length 1, is
\[
\mathbf{v}^* = \frac{\mathbf{c}}{||\mathbf{c}||} = \frac{(\mathbf{u}' \Sigma_{YX})'}{||((\mathbf{u}' \Sigma_{YX})')||} = \frac{\Sigma_{YX}' \mathbf{u}}{||\Sigma_{YX}' \mathbf{u}||} = \frac{\Sigma_{XY} \mathbf{u}}{||\Sigma_{XY} \mathbf{u}||}.
\] (3.19)

This corollary provides a method for obtaining basis vectors for the \( \mathbf{X} \) process based on given basis vectors, or patterns, for the \( \mathbf{Y} \) process. Say we have patterns \( \mathbf{u}_1, \ldots, \mathbf{u}_K \). For example they could be the first \( K \) sample EOFs of the \( \mathbf{Y} \) process. The vectors
\[
\mathbf{v}_k = \frac{\Sigma_{XY} \mathbf{u}_k}{||\Sigma_{XY} \mathbf{u}_k||} \quad k = 1, \ldots, K
\] (3.20)
are a set of patterns (or basis vectors, but see the last paragraph in this Section) which we can use for dimension reduction for \( \mathbf{X} \). They have the property that the amplitudes \( A_k = \mathbf{u}_k' \mathbf{Y} \) and \( B_k = \mathbf{v}_k' \mathbf{X} \) have highest possible covariance. That is an appealing property for use in the hierarchical modeling framework in Chapter 2. We have the added bonus of being able to pick \( \mathbf{u}_k \) vectors that capture well the spatial structure in the target process \( \mathbf{Y} \). We call the \( \mathbf{v}_k \) vectors in (3.20) Maximum Covariance Patterns (MCPs).

The sample MCPs can be obtained by replacing \( \Sigma_{XY} \) with the sample covariance matrix. Let \( \mathbf{Y} \) and \( \mathbf{X} \) be the \( N_Y \times T \) and \( N_X \times T \) dimensional temporally centered
data matrices. Then the sample MCPs are
\[
v_k = \frac{S_{XY} u_k}{||S_{XY} u_k||} = \frac{\frac{1}{T-1} Y' u_k}{\frac{1}{T-1} Y' Y'} = \frac{X Y' u_k}{||X Y' u_k||}
\]
\[k = 1, \ldots, K\]  \hspace{1cm} (3.21)
Note that by first multiplying $Y' u_k$ we do not need to store the large $N_X \times N_Y$ matrix $X Y'$.

Using the sample covariance matrix $\frac{1}{T-1} X Y'$ to obtain the MCPs in (3.21) requires that $Y$ and $X$ are observed at the same number of time points ($T$). At first glance this seems to be the same restriction as for MCA and CCA but there is an important distinction in that the $u$ vectors for the $Y$ process are fixed. If $X$ is observed at a higher frequency (e.g. monthly versus seasonally), but so that each sub-frequency is observed as often as the $Y$ process, then we can obtain the vectors in (3.21) for each sub-frequency separately. See Chapter 4 for an example of this.

We need to issue some words of warning at this stage. Namely, the vectors in (3.20) are not guaranteed to be linearly independent. The same is true for the vectors in (3.21). The problem lies with the cross covariance matrix $\Sigma_{XY}$ (or $X Y'$). First note that $v_k$ is not defined if $\Sigma_{XY} u_k = 0$, which can happen if $u_k$ is in the null space of $\Sigma_{XY}$ or if $\Sigma_{XY}$ is a matrix with all zero elements. Of course, if the latter case is true then all elements of $X$ and $Y$ have zero covariance and there is not much hope of (linearly) predicting $Y$ from $X$ in the first place! Assume that $\Sigma_{XY} u_k \neq 0$ for all columns $u_k$ from an $N_Y \times N_Y$ basis matrix $U$. The full set of MCPs can then be written as
\[
V = \Sigma_{XY} \text{diag} \left( \frac{1}{||\Sigma_{XY} u_1||}, \ldots, \frac{1}{||\Sigma_{XY} u_{NY}||} \right) U.
\]
Now, $V$ is $N_X \times N_Y$ dimensional and cannot be of full rank if $N_X < N_Y$. Furthermore, since $\text{diag} \left( \frac{1}{||\Sigma_{XY} u_1||}, \ldots, \frac{1}{||\Sigma_{XY} u_{NY}||} \right) U$ is a non-singular matrix (since $\Sigma_{XY} u_k \neq 0$)
then we have

\[
\text{rank}(\mathcal{V}) = \text{rank}(\Sigma_{XY})
\]  \hspace{1cm} (3.23)

(see, for example, Ravishanker and Dey, 2002, page 24). This means that the rank of \( \mathcal{V} \) may be less than \( N_X \) which disqualifies the column vectors of \( \mathcal{V} \) from spanning the \( \mathbb{R}^{N_X} \) space. In the case of sample MCPs, we know that \( S_{XY} \) is of maximum rank \( T \) (assuming \( T < N_X \) and \( T < N_Y \)) which puts an upper limit for \( K \). Of course, if the \( u_k \)'s are sample EOFs, \( T \) is an upper limit for the number of \( u_k \) vectors. We would never need all \( N_X \) vectors in \( \mathcal{V} \), it suffices that the first \( K \) vectors are linearly independent but there is no guarantee for that. In practice one could check whether the obtained \( K \) MCPs are linearly independent, that is if the matrix \( V \) that contains the \( K \) \( v_k \) vectors has full column rank. That could be done by checking whether \( V'V \) is singular or nearly singular by examining the condition number. Another option would be to add the constraint that the \( v \) vectors should be orthogonal. That is the subject of Section 3.5

### 3.4.1 Proportion of total variance explained

As in Sections 3.2 and 3.3 we can consider the proportion of total (sample) variance of \( Y \) and \( X \) that is explained by the first few \( u_k \) and \( v_k \) vectors, see equations (3.8) and (3.14). Of course, if the \( u_k \) vectors are EOFs, we know that the proportion of total variance of \( Y \) explained by the first \( K \) EOFs is the sum \( q_1 + \cdots + q_K \) where

\[
q_k = \frac{\text{Var}(u_k'Y)}{\text{tr}(\Sigma_{YY})} = \frac{\text{Var}(A_k)}{\text{tr}(\Sigma_{YY})};
\]  \hspace{1cm} (3.24)

see also equation (3.2).

The motivation behind MCPs was our primary interest in explaining the structure of \( Y \). Furthermore, the modeling of the structure of \( Y \) is in two stages. First,
through the basis vectors for \( Y (u_1, \ldots, u_K) \) and second, through the basis vectors for \( X (v_1, \ldots, v_K) \) whose amplitudes are used to predict the amplitudes of \( u_1, \ldots, u_K \). Therefore, both the choice of \( u_k \) vectors and the choice of \( v_k \) vectors is relevant to how much of the structure in \( Y \) is captured. With that in mind we ask the question “What proportion of the total variance in \( Y \) is explained by the first \( v_k \) vectors through the first \( u_k \) vectors?” We show that this proportion is higher if we use the first MCP basis vector for \( X \) than if we use the first EOF of \( X \).

In general, the squared correlation between two random variables, say \( A \) and \( B \), can be viewed as the proportion of the variance of one variable explained by the other. This stems from the simple fact that

\[
Cor(A, B)^2 = \frac{Cov(A, B)^2}{Var(B)} / Var(A) .
\]

(3.25)

In fact, the squared correlation is the same as the coefficient of determination, \( R^2 \), of the simple regression of \( A \) on \( B \). Combining this with (3.24) we present the following definition:

**Definition 3.4.3.** Let \( Y \) and \( X \) be \( N_Y \) and \( N_X \) dimensional random vectors and let \( u \) and \( v \) be \( N_Y \) and \( N_X \) dimensional known vectors. Let \( q \) be the proportion of the total variance in \( Y \) that is explained by \( u \), i.e.

\[
q = \frac{Var(u'Y)}{\text{tr}(\Sigma_{YY})} .
\]

(3.26)

where \( \Sigma_{YY} \) is the covariance matrix of \( Y \). Let \( R^2 \) be the coefficient of determination of the simple regression of \( A = u'Y \) on \( B = v'X \). We define the *proportion of the*
total variance in $Y$ that is explained by $v$ through $u$ by

$$p(u, v) = qR^2 = \frac{\text{Var}(A)}{\text{tr}(\Sigma_{YY})} \frac{\text{Cov}(A, B)^2}{\text{Var}(A)\text{Var}(B)} = \frac{\text{Cov}(A, B)^2/\text{Var}(B)}{\text{tr}(\Sigma_{YY})}.$$  

(3.27)

The proportion in (3.27) can be used as an indicator of how much of the structure in $Y$ is captured by $X$ through the basis vectors $u$ and $v$. The next theorem shows that for a given vector $u$, $p$ is higher when $v$ is the MCP with respect to $u$ than if $v$ is chosen as the first EOF of $X$.

Theorem 3.4.4. Let $Y$ and $X$ be $N_Y$ and $N_X$ dimensional random vectors and let $u$ be an $N_Y$ dimensional known vector. Let $v$ be the MCP of $X$ with respect to $u$ and let $f$ the first EOF of $X$. Then

$$p(u, v) \geq p(u, f).$$  

(3.28)

Proof. By construction we know that $f$ maximizes the variance $\text{Var}(f'X)$ under the constraint that $f$ is of unit length. Since $v$ is (by construction) of length 1 we get

$$\text{Var}(v'X) \leq \text{Var}(f'X).$$  

(3.29)

Also we know that $v$ maximizes $\text{Cov}(u'Y, v'X)$ and so we get that

$$\text{Cov}(u'Y, v'X) \geq \text{Cov}(u'Y, f'X).$$  

(3.30)

Then we get that

$$p(u, v) = \frac{\text{Cov}(u'Y, v'X)^2/\text{Var}(v'X)}{\text{tr}(\Sigma_{YY})} \geq \frac{\text{Cov}(u'Y, v'X)^2/\text{Var}(v'X)}{\text{tr}(\Sigma_{YY})} \geq \frac{\text{Cov}(u'Y, f'X)^2/\text{Var}(f'X)}{\text{tr}(\Sigma_{YY})} = p(u, f)$$  

(3.31)
The first inequality follows from (3.29) and the second inequality follows from (3.30).

The consequence of Theorem 3.4.4 is that we have shown that in terms of total variance in $Y$ explained, it is better to use the MCPs as basis vectors for $X$ than EOFs. That is, the proportion of total variance of $Y$ explained by the MCP through $u$ is no less than (and possibly greater than) the proportion of total variance of $Y$ explained by the first EOF of $X$ through $u$.

3.4.2 Covariance of EOF and MCP amplitudes across pairs

The MCPs are constructed so that for a given basis vectors $u_i$ the amplitudes $A_i = u_i Y$ and $B_i = v_i X$, $i = 1, \ldots, K$, have maximum covariance. Here we investigate what can be said about the covariance between $A_i$ and $B_j$ amplitudes for $i \neq j$. The motivation comes from the data example in Chapter 4 where these amplitudes were found to have weak or no relationships across pairs, see Section 4.4 and Appendix B.3.1. So the question is whether the $B_j$ amplitude that is designed to be a good (linear) predictor of $A_j$ is of value for predicting another amplitude $A_i$.

Let $\angle(\Sigma_{XY} u_i, \Sigma_{XY} u_j)$ denote the angle between vectors $\Sigma_{XY} u_i$ and $\Sigma_{XY} u_j$. Recalling that $v_j = \frac{\Sigma_{XY} u_j}{||\Sigma_{XY} u_j||}$ we can write the covariance between $A_i$ and $B_j$, $i \neq j$ as

$$
\text{Cov}(A_i, B_j) = \text{Cov}(u_i' Y, v_j' X) = u_i' \Sigma_{YX} v_j
$$

$$
= u_i' \Sigma_{YX} \Sigma_{XY} u_j ||\Sigma_{XY} u_j||^{-1} = (\Sigma_{XY} u_i)' \Sigma_{XY} u_j ||\Sigma_{XY} u_j||^{-1}
$$

$$
= ||\Sigma_{XY} u_i|| ||\Sigma_{XY} u_j|| \cos (\angle(\Sigma_{XY} u_i, \Sigma_{XY} u_j)) \frac{1}{||\Sigma_{XY} u_j||}
$$

$$
= ||\Sigma_{XY} u_i|| \cos (\angle(\Sigma_{XY} u_i, \Sigma_{XY} u_j)) .
$$

(3.32)
We note that by construction \( \text{Cov}(A_i, B_j) \leq \text{Cov}(A_i, B_i) \) so the covariance between \( A_i \) and \( B_j \) cannot be larger than the covariance between \( A_i \) and \( B_i \). Furthermore, since \( \text{Cov}(A_i, B_i) = ||\Sigma_{XY}u_i|| \) we have

\[
\text{Cov}(A_i, B_j) = \text{Cov}(A_i, B_i) \cos (\angle(\Sigma_{XY}u_i, \Sigma_{XY}u_j)) \quad (3.33)
\]

So the difference between \( \text{Cov}(A_i, B_j) \) and \( \text{Cov}(A_i, B_i) \) depends on the angle between \( \Sigma_{XY}u_i \) and \( \Sigma_{XY}u_j \). If \( \Sigma_{XY}u_i \) and \( \Sigma_{XY}u_j \) are orthogonal then \( \text{Cov}(A_i, B_j) = 0 \). That could happen if the cross-covariance matrix \( \Sigma_{XY} \) is a linear transform that preserves angles and, \( u_i \) and \( u_j \) are orthogonal; e.g. EOFs. If the angle between \( \Sigma_{XY}u_i \) and \( \Sigma_{XY}u_j \) is zero then \( \text{Cov}(A_i, B_j) = \text{Cov}(A_i, B_i) \), but in that case the \( i \)th and \( j \)th MCP \((v_i \text{ and } v_j) \) would be identical.

In light of (3.32) and (3.33) it seems that the findings we demonstrate in Section 4.4 are specific to that data set but not a general property of MCPs.

### 3.4.3 MCP for more than two spatio-temporal processes

Suppose we have more than one spatio-temporal process, \( X_i \), that we wish to use to predict the spatio-temporal process \( Y \). In the statistical downscaling example in Chapter 4 we have three course-resolution temperature fields (ERA-40 data) from three months that we want to use to predict one seasonal high-resolution temperature field (Polar MM5). For illustration purposes I will assume that we have three processes \( X_1, X_2 \) and \( X_3 \) but the considerations here apply for any number of \( X_i \) processes.

Let \( u \) be a basis vector for \( \mathbb{R}^{N_Y} \), for example the first EOF of \( Y \). Let \( X_i \) be an \( N_{X_i} \) dimensional random vector containing the \( X_i \) process at \( N_{X_i} \) locations, \( i = 1, 2, 3 \). To incorporate all \( X_i \) processes in prediction of the amplitude \( A = u'Y \) we consider
finding a combined $\sum_{i=1}^{3} N_{X_i}$ dimensional vector $\mathbf{w}$ that maximizes

$$Cov(\mathbf{u}'Y, \mathbf{w}'\mathbf{X}) \quad \text{where} \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{pmatrix}. \quad (3.34)$$

From Theorem 3.4.2 we know that the vector that maximizes the covariance above is

$$\mathbf{w} = \frac{\Sigma_{XY} \mathbf{u}}{\|\Sigma_{XY} \mathbf{u}\|} \quad \text{where} \quad \Sigma_{XY} = \begin{pmatrix} \Sigma_{X_1Y} \\ \Sigma_{X_2Y} \\ \Sigma_{X_3Y} \end{pmatrix}. \quad (3.35)$$

Alternatively, we could also consider finding the MCPs for each $\mathbf{X}_i$, i.e.

$$\mathbf{v}_i = \frac{\Sigma_{X_iY} \mathbf{u}}{\|\Sigma_{X_iY} \mathbf{u}\|} \quad \text{for} \quad i = 1, 2, 3. \quad (3.36)$$

The $\mathbf{w}$ vector can be written in terms of the $\mathbf{v}_i$ vectors:

$$\mathbf{w} = \frac{1}{\|\Sigma_{XY} \mathbf{u}\|} \begin{pmatrix} \Sigma_{X_1Y} \\ \Sigma_{X_2Y} \\ \Sigma_{X_3Y} \end{pmatrix} \mathbf{u} = \frac{1}{\|\Sigma_{XY} \mathbf{u}\|} \begin{pmatrix} \Sigma_{X_1Y} \mathbf{u} \\ \Sigma_{X_2Y} \mathbf{u} \\ \Sigma_{X_3Y} \mathbf{u} \end{pmatrix}$$

$$= \frac{1}{\|\Sigma_{XY} \mathbf{u}\|} \begin{pmatrix} \|\Sigma_{X_1Y} \mathbf{u}\| \mathbf{v}_1 \\ \|\Sigma_{X_2Y} \mathbf{u}\| \mathbf{v}_2 \\ \|\Sigma_{X_3Y} \mathbf{u}\| \mathbf{v}_3 \end{pmatrix} \quad (3.37)$$

The vectors $\mathbf{w}$ and $\left( \begin{array}{c} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{array} \right)$ are not the same in the space $\mathbb{R}^{N_{X_1}+N_{X_2}+N_{X_3}}$. But the first $N_{X_1}$ elements of $\mathbf{w}$ are the same as $\mathbf{v}_1$ save for a multiplicative scalar. That is, they have the same direction but may not be of the same length. The same is true for the next $N_{X_2}$ elements of $\mathbf{w}$ and $\mathbf{v}_2$ and also for the last $N_{X_3}$ elements of $\mathbf{w}$ and $\mathbf{v}_3$. This means that $\mathbf{w}$ picks up the same patterns from $\mathbf{X}_i$ as $\mathbf{v}_i$ does.

### 3.5 Orthogonal Maximum Covariance Patterns

In this Section we consider adding the constraint on the Maximum Covariance Patterns that they be orthogonal. We start with a general theorem and a corollary.
Theorem 3.5.1. Let $c$ be a known (constant) vector in $\mathbb{R}^N$ and let $v_1, \ldots, v_k$ be orthonormal vectors in $\mathbb{R}^N$ where $1 \leq k < N$. The solution to the maximization problem

$$\begin{align*}
\text{maximize } & f(v_{k+1}) = c'v_{k+1} \\
\text{subject to } & v'_{k+1}v_{k+1} = 1 \text{ and } v'_{k+1}v_j = 0 \text{ for } j = 1, \ldots, k \quad (3.38)
\end{align*}$$

is

$$v^*_{k+1} = \frac{1}{L} \left( c - \sum_{j=1}^{k} (c'v_j)v_j \right) \quad (3.39)$$

where $L = ||c - \sum_{j=1}^{k} (c'v_j)v_j||$. Furthermore, the vector that minimizes $f(v)$ subject to the same constraints is $-v^*_{k+1}$.

Proof. The objective function is the same as in Theorem 3.4.1 and we can write it as

$$f(v_{k+1}) = c'v_{k+1} = c \cdot v_{k+1} = ||c|| ||v_{k+1}|| \cos(\theta_{k+1}) \quad (3.40)$$

where $\theta_{k+1}$ is the angle between the vectors $v_{k+1}$ and $c$. Since we have the added constraint that $v_{k+1}$ has to be orthogonal to $v_j, j = 1, \ldots, k$ we cannot simply take $\theta_{k+1} = 0$. Note that $\cos(\theta_{k+1})$ decreases as $\theta_{k+1}$ increases (for $\theta_{k+1} \in [0, \pi]$). Therefore the optimal vector is both orthogonal to $v_j, j = 1, \ldots, k$ and has the smallest $\theta_{k+1}$, or in other words, it is the residual vector from the orthogonal projection of $c$ onto the space spanned by $v_1, \ldots, v_k$. Furthermore, the vector that gives the largest $\theta_{k+1}$ while still orthogonal to $v_j, j = 1, \ldots, k$ is $-v^*_{k+1}$. An alternative (and longer) proof is provided in Appendix A.2. \qed
Corollary 3.5.2. Let $Y$ and $X$ be $N_Y$ and $N_X$ dimensional random vectors with cross-covariance matrix $\text{Cov}(Y, X) = \Sigma_{YX}$. Let $u_1, \ldots, u_K$ be fixed, linearly independent, $N_Y$ dimensional vectors. Then the unit-length vectors $v_1, \ldots, v_K$ that maximize the covariances $\text{Cov}(u'_1 Y, v'_1 X), \ldots, \text{Cov}(u'_K Y, v'_K X)$ under the constraint that $v_1, \ldots, v_K$ are orthogonal are the following:

\begin{align*}
v_1^* &= \frac{\Sigma_{XY} u_1}{||\Sigma_{XY} u_1||} \\
v_2^* &= \frac{1}{L_2} (\Sigma_{XY} u_2 - ((\Sigma_{XY} u_2)' v_1^*) v_1^*) \\
& \vdots \\
v_K^* &= \frac{1}{L_K} \left( \Sigma_{XY} u_K - \sum_{j=1}^{K-1} ((\Sigma_{XY} u_K)' v_j^*) v_j^* \right) \quad (3.41)
\end{align*}

where $\Sigma_{XY} = \Sigma_{YX}'$ and $L_2, \ldots, L_K$ are normalizing constants.

Proof. We know from Corollary 3.4.2 that we can take

\begin{equation}
v_1^* = \frac{\Sigma_{XY} u_1}{||\Sigma_{XY} u_1||} \tag{3.42}
\end{equation}

since it maximizes $\text{Cov}(u'_1 Y, v'_1 X)$. Next, consider $v_2$. We want to maximize the covariance,

\begin{equation}
\text{Cov}(u'_2 Y, v'_2 X) = u'_2 \Sigma_{YX} v_2 = (\Sigma_{XY} u_2)' v_2, \tag{3.43}
\end{equation}

under the constraint that $||v_2|| = 1$ and that $v_2$ is orthogonal to $v_1^*$. By Theorem 3.5.1 we get

\begin{equation}
v_2^* = \frac{1}{L_2} (\Sigma_{XY} u_2 - ((\Sigma_{XY} u_2)' v_1^*) v_1^*) \tag{3.44}
\end{equation}

where $L_2$ is the normalizing constant. In general for $k = 2, \ldots, K$ we want to maximize $\text{Cov}(u'_k Y, v'_k X) = (\Sigma_{XY} u_k)' v_k$ under the constraint that $||v_k|| = 1$ and that $v_k$ is orthogonal to $v_1^*, \ldots, v_{k-1}^*$. By Theorem 3.5.1 we get that

\begin{equation}
v_k^* = \frac{1}{L_k} \left( \Sigma_{XY} u_k - \sum_{j=1}^{k-1} ((\Sigma_{XY} u_k)' v_j^*) v_j^* \right) \tag{3.45}
\end{equation}
where $L_k$ is a normalizing constant.

Corollary (3.5.2) gives a recipe for obtaining MCPs as in Section 3.4 with the added constraint that they are orthogonal. We call the optimal vectors $\mathbf{v}_1^*, \ldots, \mathbf{v}_K^*$ in corollary 3.5.2 Orthogonal Maximum Covariance Patterns (OMCPs). We note that the OMCPs are in fact the Gram-Schmidt orthogonalization the MCPs.

The orthogonality of the vectors in (3.41) eliminates concerns raised in Section 3.4 about MCPs possibly being linearly dependent and furthermore, using orthonormal basis vectors in dimension reduced modeling is computationally convenient. However, the constraint of orthogonality will of course lower the covariance of the amplitudes compared to the covariance of MCP amplitudes. This issue is explored in the next section.

### 3.5.1 Cost of orthogonality of OMCPs

In this section we examine the difference in the covariance $\text{Cov}(\mathbf{u}_{k+1}', \mathbf{v}_{k+1}')$ when $\mathbf{v}_{k+1}$ is an MCP versus an OMCP. We start by considering, how much the optimum value of the objective function in Theorem 3.5.1 changed by demanding that the $\mathbf{v}_{k+1}$ are orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_k$. The answer depends on the angle between $\mathbf{c} (= \Sigma_{XY} \mathbf{u}_{k+1})$ and the optimal vector $\mathbf{v}_{k+1}^*$. Recall that, assuming $\mathbf{v}_{k+1}^*$ has length 1,

$$f(\mathbf{v}_{k+1}^*) = \mathbf{c}' \mathbf{v}_{k+1}^* = ||\mathbf{c}|| ||\mathbf{v}_{k+1}^*|| \cos(\theta_{k+1}) = ||\mathbf{c}|| \cos(\theta_{k+1})$$  \hspace{1cm} (3.46)

where $\theta_{k+1}$ is the angle between $\mathbf{c}$ and $\mathbf{v}_{k+1}^*$. When $\theta_{k+1} = 0$ then $f(\mathbf{v}_{k+1}^*) = ||\mathbf{c}||$ as in Theorem 3.5.1, i.e. without the orthogonality constraint. When $\theta > 0$ we can measure the change in $f(\mathbf{v}_{k+1}^*)$ in terms of

$$\cos(\theta) = \frac{\mathbf{c}' \mathbf{v}_{k+1}^*}{||\mathbf{c}||}.$$  \hspace{1cm} (3.47)
The optimum value of the objective function goes from $||c||$ to $||c||\cos(\theta_{k+1})$ where $\cos(\theta_{k+1})$ can be calculated as in (3.47). Furthermore, we note that $\cos(\theta_{k+1})$ is between 0 and 1 (when we minimize $f$ we get $\cos(\theta_{k+1})$ between -1 and 0) so we can view

$$1 - \cos(\theta_{k+1})$$

as the proportion of the optimum value that is lost when demanding that $v_{k+1}$ is orthogonal to $v_1, \ldots, v_k$.

These considerations lead to the following corollary.

**Corollary 3.5.3.** Let $Y$, $X$, $u_k$, $v_k$ and $v^*_k$, $k = 1, \ldots, K$ be as in Corollary 3.5.2. Then for each $k$ the optimum covariance $\text{Cov}(u_k' Y, v^*_k' X)$ is $100(1 - \cos(\theta_k))\%$ lower than it would be without the constraint that $v_1, \ldots, v_k$ need to be orthogonal, where

$$\cos(\theta_k) = \frac{u_k' \Sigma_{YX} v^*_k}{||u_k' \Sigma_{YX}||}. \quad (3.49)$$

In practice we can estimate $\cos(\theta_k)$ by calculating

$$\hat{\cos}(\theta_k) = \frac{u_k' Y' X' v^*_k}{||u_k' Y' X'||}, \quad (3.50)$$

where $Y$ and $X$ are data matrices.
Chapter 4: Statistical Downscaling of Temperatures over the Antarctic

This Chapter contains a case study of statistical downscaling of climate data using methods introduced in Chapters 2 and 3. The climate data we work with here are Polar MM5 temperature data and ECMWF 40 Year Re-analysis (ERA-40) temperature data where ECMWF stands for European Centre for Medium-Range Weather Forecasts (see Section 4.1). The Polar MM5 data are on a high-resolution grid that covers the Antarctic continent while the ERA-40 data are on a low-resolution grid. The goal of this Chapter is to downscale ERA-40 temperatures onto the high-resolution Polar MM5 grid; i.e. to predict the Polar MM5 temperature data using ERA-40 temperature data. Both these datasets are high-dimensional and our approach is to build a dimension reduced Bayesian hierarchical model as discussed in Section 2.2. We take into account that the two datasets are not on the same temporal scales, as was briefly discussed in Section 2.3. Since the goal is to predict Polar MM5 temperatures from ERA-40 temperatures we use EOFs as basis vectors for the Polar MM5 data and OMCPs as basis vectors for ERA-40 data (see Sections 3.1 and 3.5). We choose to use the orthogonal MCPs due to their computational appeal.

We start this Chapter with a brief description of the Polar MM5 and ERA-40 datasets, see Section 4.1. In Section 4.2 we explore the two datasets mostly via
graphical displays. In Section 4.3 we obtain and explore the EOFs and OMCPs for the two datasets and in Section 4.4 we perform exploratory analysis on estimated amplitudes of the EOFs and OMCPs. In Section 4.5 we present a Bayesian hierarchical model of the Polar MM5 and ERA-40 data that uses EOFs and OMCPs for dimension reduction. In Section 4.6 we discuss methods to predict future Polar MM5 temperature data both with new ERA-40 data (i.e. downscaling) and without new ERA-40 data (i.e. prediction). We discuss the results in Section 4.7 which include posterior inferences of unknown parameters, prediction of new Polar MM5 data and downscaling of ERA-40 data onto the Polar MM5 grid.

4.1 The Polar MM5 and ERA-40 projects

The ERA-40 project is a global atmospheric analysis of both conventional observations and satellite data for the time period September 1957 - July 2002 (Uppala et al., 2006). Analyses were produced daily at 6-hour intervals. To obtain surface fields, including 2-meter surface temperatures, an atmospheric model was run on a reduced Gaussian grid with approximately uniform 125km spacing, see Figure 4.1 (left). Detailed descriptions of the project and the data assimilation system are available in the ERA-40 Archive Plan documents from ECMWF (ECMWF, 2002). ECMWF has derived a set of many variables on a regular 2.5° latitude-longitude grid (see Figure 4.1, right). The ERA-40 data used in this study are monthly 2-meter temperatures (monthly means of daily means) on the regular 2.5° latitude-longitude grid. The data were obtained from the ECMWF Data Server (http://data-portal.ecmwf.int/data/d/era40_moda/).
Figure 4.1: The reduced Gaussian grid (left) used in the ERA-40 project and the regular 2.5° latitude-longitude grid (right) on which the data are provided. These Figures were obtained from the ERA-40 archive at NCAR, http://dss.ucar.edu/pub/era40/

The Polar MM5 regional climate model is a Fifth-generation Mesoscale Model (MM5) modified for use in polar regions by scientists at the Byrd Polar Research Center at The Ohio State University (Monaghan et al., 2006b). The MM5 was developed at the Pennsylvania State University (PSU) and the National Center for Atmospheric Research (NCAR). The Polar MM5 model was, among other things, used for the Antarctic Hindcast Project at the Byrd Polar Research Center. The goal of the Antarctic Hindcast Project was to model variability and change in the net accumulation of moisture over Antarctica in recent decades. The motivation was to better understand the contribution of Antarctica, which holds about 90% of the fresh water on earth, to global sea level change (Monaghan et al., 2006a; Monaghan and Bromwich, 2008). The project also provided a valuable dataset of temperatures, pressure, wind and other climate variables for the period January 1979 through August 2002 that are
available for further studies (http://polarmet.osu.edu/PolarMet/ant_hindcast.html). An overview map of the Antarctic continent is given in Figure B.1 in Appendix B.

The Polar MM5 simulations were performed on a $121 \times 121$ polar stereographic grid covering the Antarctic and centered over the South Pole, see Figures 4.2 and 4.3. The model resolution is 60-km in each horizontal direction. Atmospheric data and sea surface temperatures provided by the ERA-40 data were used for initial and boundary conditions. Model topography was interpolated from the 1-km resolution, digital elevation model of Liu et al. (2001). The data we use in this Chapter are seasonally averaged 2-meter surface temperatures. For each Polar MM5 grid point we also have a variable which classifies the location as either ocean (72.75% of the grid points) or land/permanent ice.

4.2 Exploratory data analysis of Polar MM5 and ERA-40 data

The Polar MM5 and ER-40 data cover 93 seasons, from fall 1979 through fall 2002. Being on the South Pole, the fall months are March, April and May (MAM), winter is June, July and August (JJA), summer is December, January and February (DJF) and spring is September, October and November (SON). For the purposes of this project we consider only ERA-40 data south of and including 45°S, which covers the whole Polar MM5 grid; see Figure 4.3. To get a sense of the dimensions of the datasets, we note that the Polar MM5 grid has $121 \times 121 = 14,641$ grid points (locations) and with the 93 seasons, we have 1,361,613 data points. The ERA-40

\footnote{A stereographic projection is a mapping that projects the earth onto a plane and preserves angles but not areas. The projection point here is North Pole, so the center of the map is the South Pole.}
grid, south of 45°S, has 19 latitudes and 144 longitudes, leading to 2,736 grid points. With 93 seasons each containing 3 months, we have 763,344 ERA-40 data points.

We start with visualizing the two temperature fields by making maps of Antarctica that show 2-meter temperature values for each grid point. Figure 4.4 shows both the ERA-40 and Polar MM5 temperature fields from 1986, data from other years are shown in Appendix B.1 (Figures B.2 - B.7). There is strong spatial dependence in the temperature data and, not surprisingly, temperatures are warmer the further we go from the South Pole and in general the ocean is warmer than the land/permanent ice. Comparing temperature fields to the terrain (bottom of Figure 4.4) we see that temperatures are highly correlated with elevation, the higher up we go the colder it is. We see this correlation even more clearly in Figure 4.5 which shows Polar MM5
Figure 4.3: Another look at the ERA-40 grid, a regular $2.5^\circ \times 2.5^\circ$ latitude-longitude grid (grey points), and the stereographic Polar MM5 grid (blue points) using a stereographic projection. The Polar MM5 grid covers the South Georgia Island (the upper left corner of the blue grid) but just misses the Falkland Islands.
Figure 4.4: Temperature fields, ERA-40 (circles) and Polar MM5 (squares) in Kelvin, year 1986. Note that $273.15K = 0^\circ C$. Bottom panel: Elevation values (in meters) for each Polar MM5 grid point.
temperatures for all grid points not over ocean and from all seasons plotted against elevation.

Inspecting temperatures fields for the 24 years (Figures B.2 - B.7 in Appendix B.1), we see that the fields vary from year to year but it is not clear whether there is a temporal trend over this period. Figure 4.6 shows temporal plots of average Polar MM5 2-meter temperatures for land/permanent ice (left of Figure 4.6) and ocean locations (right of Figure 4.6). We clearly see the seasonal differences in average temperatures but there is no clear trend over time in these aggregated data. Figure 4.7 shows temporal plots of average ERA-40 2 meter temperatures, averaged over all locations (we do not have terrain or land use variables for the ERA-40 grid). Here too we see the seasonal differences in average temperatures but there is no clear trend in these data over the 24 years, except perhaps in the three winter months where there seems to be a slight increase in temperatures over the two decades.

4.3 Centered data, OMCPs and EOF patterns

In this Section we obtain the sample EOF patterns of the Polar MM5 data and the sample OMCPs of the ERA-40 data. Note that we need to center the data before applying the methods in Section 3.1 and 3.5.

We start by establishing some notation. Let $Y_{l,t,i}$ be the centered 2-meter temperature model output of the Polar MM5 from season $l$, year $t$ and location $i$ and let $X_{m,t,j}$ be the centered ERA-40 2-meter temperatures for month $m$, year $t$ and location $j$. The number of spatial locations are $N_Y = 14641$ for the Polar MM5 data and $N_X = 2736$ for the ERA-40 data.
Figure 4.5: Polar MM5 2-meter temperatures from all 24 years versus elevation for land locations only. The land use variable (i.e. ocean or land/permanent ice) was provided with the Polar MM5 data. We also show boxplots for every 200 meter interval where the width of the box is proportional to the amount of data in the interval.

Figure 4.6: Average Polar MM5 2-meter temperatures over land/permanent ice (left) and ocean (right) locations, for each season.
Figure 4.7: Average ERA-40 2-meter temperatures over all locations, for each month. The colors denote the four seasons and the line type is used to denote the first, second and third month within a season.

We have data from the fall season (March) of 1979 through the fall season (May) of 2002; a total of 93 seasons (279 months). We leave out the last year of data so that we can compare them to our predictions for that period (see Sections 4.7.1 - 4.7.3). We therefore obtain the basis vectors and fit the model using only the first 89 seasons (267 months), i.e. through fall (May) of 2001.

Note that seasons and months start and end at different years. Also, we have no summer data for the first year (1979) and no winter or spring data the 23\textsuperscript{rd} year (2001). Let \( \tau_l \) and \( T_l \) be the the first and last year we have available data for season \( l, l = s, f, w, p \). In particular, we have summer data for years \( t = 1980, \ldots, 2001 \) so \( \tau_s = 1980 \) and \( T_s = 2001 \). We have data for all other seasons from the first year so \( \tau_f = \tau_w = \tau_p = 1979 \). We have fall data from the last year so \( T_f = 2002 \) but no winter and spring data the last year so \( T_s = T_f = 2001 \). Similarly, let \( \tau_m \) and \( T_m \) be the
first and last year we have data for month $m$, $m = 1, \ldots, 12$. Finally, note that the number of observations (years) for each season is $T_l - \tau_l + 1$ and similarly the number of observations (years) for each month is $T_m - \tau_m + 1$.

We define data vectors and data matrices as follows:

$Y_{l,t}$ is an $N_Y$-dimensional vector of centered seasonal Polar MM5 2-meter surface temperatures, indexed by year $t = \tau_l, \ldots, T_l$ and season $l = s, f, w, p$.

$Y_l$ is a $N_Y \times (T_l - \tau_l + 1)$ dimensional data matrix with columns $Y_{l,\tau_l}, \ldots, Y_{l,T_l}$.

$X_{m,t}$ is an $N_X$-dimensional vector of centered monthly ERA 2-meter surface temperatures, indexed by year $t = \tau_m, \ldots, T_m$ and month $m = 1, \ldots, 12$.

$X_m$ is a $N_X \times (T_m - \tau_m + 1)$ dimensional data matrix with columns $X_{m,\tau_m}, \ldots, X_{m,T_m}$.

Centered data are obtained by subtracting the seasonal (Polar MM5) or monthly (ERA-40) means at each location. This means that

$$
\sum_{t=\tau_l}^{T_l} Y_{l,t,i} = 0 \quad \text{and} \quad \sum_{t=\tau_m}^{T_m} X_{m,t,j} = 0 \quad \text{for all } l, m, i, j.
$$

(4.1)

Figure 4.8 shows the centered temperature fields for 1986, other years can be seen in Figures B.8 - B.13 in Appendix B.2. The centered-data maps look very different from the maps in Figures B.2 - B.7. Most centered temperatures are close to 0K, especially over the ocean. This indicates that temperatures at these locations do not vary much over time. Locations on land/permanent ice vary more, especially those close to the coast of Antarctica. The differences from the averages can be as much as 15K for the ERA-40 data and as much as 10K for the Polar MM5 data. Furthermore, the spatial structures in the centered data seem to be more alike within seasons and within months (i.e. across years) than across seasons and months.
Figure 4.8: Centered temperature fields, ERA-40 (circles) and Polar MM5 (squares) in Kelvin, year 1986. The color palettes are separate for the ERA-40 and Polar MM5 data.
To allow for different spatial structures between seasons in the centered Polar MM5 data, we construct EOF basis vectors for the centered Polar MM5 data separately for each season. This means that for each season the 22 years (23 for fall season) are treated as repeated measurements as opposed to treating all 89 seasons as repetitions of one single process. We obtain matrices of four EOFs, $\mathbf{U}_t$, by performing singular value decomposition (SVD) on the data matrices $Y_s$, $Y_f$, $Y_w$ and $Y_p$.

We have three months of centered ERA-40 data for every season. We obtain OMCPs separately for each month, using EOFs from the corresponding season. For example, the first OMCPs for March, April and May are

$$v_{3,1} = \frac{X_3 Y'_f u_{f,1}}{||X_3 Y'_f u_{f,1}||}, \quad v_{4,1} = \frac{X_4 Y'_f u_{f,1}}{||X_4 Y'_f u_{f,1}||} \quad \text{and} \quad v_{5,1} = \frac{X_5 Y'_f u_{f,1}}{||X_5 Y'_f u_{f,1}||}$$

where $u_{f,1}$ is the first EOF for the fall season. The patterns $v_{3,k}$, $v_{4,k}$, $v_{5,k}$ and $u_{f,k}$ are referred to as group $k$ for fall season.

The first five groups of EOFs and OMCPs are shown in Figures 4.9 - 4.12. These patterns show strong spatial dependence. We see that most of the first vectors (see the first row of Figures 4.9 - 4.12) have a strong land versus ocean component. Other patterns often emphasize smaller regions, e.g. along the coast of Antarctica or parts of the Pacific Ocean.

The proportion of total sample variance of each field explained by increasing number on EOFs and OMCPs is shown in Figure 4.13. For how these are calculated see equations (3.2) and (3.8). The first four EOFs capture about 70%-80% of the total variances, see also Table 4.1. With this in mind we use the first four EOFs and OMCPs in the Bayesian model in Section 4.5.

Figure 4.14 shows the proportion of the sample covariance between the amplitudes is lost by orthogonalizing the MCPs. For most of the first three patterns we lose less
Figure 4.9: First five groups of OMCPs (circles) and EOF patterns (squares) for summer season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$).
Figure 4.10: First five groups of OMCPs (circles) and EOF patterns (squares) for fall season \((v_{3,k}, v_{4,k}, v_{5,k} \text{ and } u_{f,k} \text{ for } k = 1, \ldots, 5)\).
Figure 4.11: First five groups of OMCPs (circles) and EOF patterns (squares) for winter season ($v_{6,k}$, $v_{7,k}$, $v_{8,k}$ and $u_{w,k}$ for $k = 1, \ldots, 5$).
Figure 4.12: First five groups of OMCPs (circles) and EOF patterns (squares) for spring season ($v_{9,k}$, $v_{10,k}$, $v_{11,k}$ and $u_{p,k}$ for $k = 1, \ldots, 5$).
Figure 4.13: Proportion of total sample variance explained by the EOFs for each season (left) and OMCPs for each month (right). The colors denote the four seasons and the line type is used to denote the first (solid), second (dashed) and third (dotted) month within a season.

than 20% of the sample covariance while some of the 4th and 5th patterns lose as much as 60%.

4.4 Exploratory analysis on estimated amplitudes

The hierarchical model that we discuss in Section 4.5 is built by modeling the amplitudes of the basis vectors. Some preliminary analysis on the amplitudes is therefore in order. We estimate amplitude vectors for each season, month and year using regular least squares estimates. Let $U_l$ be a matrix that contains the first few EOFs in $U_l$ and let $V_m$ be a matrix that contains the first few OMCPs in $V_m$. Note that since the column vectors in $U_l$ and $V_m$ are orthogonal, we have $U_l'U_l = I$ and
EOFs of Polar MM5 data

<table>
<thead>
<tr>
<th>Season</th>
<th>Summer</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. expl.</td>
<td>72.38%</td>
<td>68.31%</td>
<td>81.28%</td>
<td>76.58%</td>
</tr>
</tbody>
</table>

OMCPs of ERA-40 data

<table>
<thead>
<tr>
<th>Month</th>
<th>December</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. expl.</td>
<td>47.76%</td>
<td>72.81%</td>
<td>84.77%</td>
<td>64.77%</td>
<td>53.72%</td>
<td>53.57%</td>
</tr>
<tr>
<td>Month</td>
<td>June</td>
<td>July</td>
<td>August</td>
<td>September</td>
<td>October</td>
<td>November</td>
</tr>
<tr>
<td>Prop. expl.</td>
<td>50.34%</td>
<td>35.59%</td>
<td>56.48%</td>
<td>65.67%</td>
<td>55.07%</td>
<td>62.65%</td>
</tr>
</tbody>
</table>

Table 4.1: Proportion of total sample variance explained by the first four EOFs for each season and the first four OMCPs for each month.

Figure 4.14: Proportion of the sample covariance between amplitudes lost by orthogonalizing the MCPs. The colors denote the four seasons and the line type is used to denote the first, second and third month within a season.
\[ V'_m V_m = I. \] The estimated amplitude vectors are calculated as follows:

\[
\hat{a}_{l,t} = (U'_l U_l)^{-1} U'_l Y_{l,t} \quad \text{and} \quad \hat{b}_{m,t} = (V'_m V_m)^{-1} V'_m X_{m,t} = V'_m X_{m,t} \quad (4.3)
\]

This means that we have \(T_l - \tau_l + 1\) estimates of each amplitude vector and we can explore the relationships between them.

First, we consider the relationship between an amplitude vector \(\hat{a}_{l,t}\) and each of the three \(\hat{b}_{m,t}\) amplitude vectors from the same season. The goal here is to suggest models for \([a_{l,t}|b_{m_{1,t}}, b_{m_{2,t}}, b_{m_{3,t}}, \theta]\), where \(m_{l1}, m_{l2}\) and \(m_{l3}\) are the three months in season \(l\). Figure 4.15 shows the first five estimated summer EOF amplitudes, \(\hat{a}_{s,t,k_Y}\) for \(k_Y = 1, \ldots , 5\), plotted against the first five December OMCP amplitudes, \(\hat{b}_{12,t,k_X}\) for \(k_X = 1, \ldots , 5\). Note that each point on the graphs represents one year \(t\).

Similar graphs for other seasons and months are shown in Appendix B.3.1, figures B.14 - B.24. The lines (red or blue) show the fitted regression line in each case. We see that amplitudes from the same group i.e., when \(k_Y = k_X\), have strong linear relationships (see the diagonal graphs in Figure 4.15 with red regression lines). This is not surprising since MCP basis vectors are designed so that these amplitudes would have high covariance. The graphs in the upper triangle, i.e. higher order EOFs versus lower order OMCPs \((k_Y > k_X)\), show virtually no relationships. Some of the graphs in the lower triangle, i.e. lower order EOFs versus higher order OMCPs \((k_Y < k_X)\) show linear relationships but most are weaker than the ones on the diagonal. The same is true for all other combinations of seasons and months, see Figures B.14 - B.24 in Appendix B.3.1. These exploratory analyses indicate that when we model the amplitude \(a_{l,t,k_Y}\) conditional on the three amplitude vectors \(b_{m_{1,t}}, b_{m_{2,t}}\) and \(b_{m_{3,t}}\) we can make do with only the amplitudes within the same group, i.e. \(k_X = k_Y\).
Figure 4.15: First five estimated Polar MM5 EOF amplitudes for summer (all 22 years), $a_{s,t,k_Y}$, plotted against the first five estimated ERA-40 OMCP amplitudes for the first summer month (December, all 22 years), $b_{12,t,k_X}$, for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 

93
The considerations above are only concerned with pairwise relationships of \( a \) and \( b \) amplitudes. We next consider multiple linear regression of the estimated amplitudes EOF \( \hat{a}_{l,t,k} \) on estimated OMCP amplitudes from the same group, \( \hat{b}_{m_1,t,k} \), \( \hat{b}_{m_2,t,k} \) and \( \hat{b}_{m_3,t,k} \), see Table 4.2. The \( \hat{b}_{m,t,k} \) amplitudes do indeed explain a big part of the variation in \( \hat{a}_{l,t,k} \), \( R^2 \) of these regressions range from about 56% to 86%. Furthermore, in most cases it is sufficient to include only the last month of the season, see last column of Table 4.2.

Our next investigation concerns the covariance structure of

\[
[a_{l,t}|b_{m_1,t}, b_{m_2,t}, b_{m_3,t}, \theta],
\]

i.e. the correlation between EOF amplitudes, given the OMCP amplitudes, within the same season but across groups. We do that by calculating the correlations between the residuals from the multiple linear regressions of the estimated amplitudes \( \hat{a}_{l,t,k} \) on \( \hat{b}_{m_1,t,k} \), \( \hat{b}_{m_2,t,k} \) and \( \hat{b}_{m_3,t,k} \) (as in Table 4.2). These correlations, for all combination of groups, are shown in Table 4.3. Some of the amplitudes have a moderate correlation (e.g. Spring \( k = 2 \) and \( k = 3 \)) but others have very low correlation.

Next we consider the relationship between the OMCP amplitudes and the OMCP amplitude vectors from the previous month. The goal here is to explore the temporal aspect, i.e. the distribution \( [b_{m,t}|b_{m-1,t}, \theta] \). Figure 4.16 shows the first five estimated December OMCP amplitudes, \( \hat{b}_{12,t,k_1} \) for \( k_1 = 1, \ldots, 5 \), plotted against the first five November OMCP amplitudes, \( \hat{b}_{11,t,k_2} \) for \( k_2 = 1, \ldots, 5 \). Again, each point on the graphs represents one year \( t \) and the lines (red or blue) show the fitted regression line. Other seasons and months are shown in Figures B.25 - B.35 in Appendix B.3.2. A few graphs in Figure 4.16 show a moderately strong linear relationship between \( \hat{b}_{12,t,k_1} \) and \( \hat{b}_{11,t,k_2} \), but we do not see a pattern of the diagonal graphs (\( k_1 = k_2 \),
### Table 4.2: Results from regressing estimated amplitudes \( \hat{a}_{l,t,k} \) on \( \hat{b}_{m_1,t,k} \), \( \hat{b}_{m_2,t,k} \) and \( \hat{b}_{m_3,t,k} \) for each season \( l \) and group \( k \). Columns 2 through 5 show the estimated regression coefficients and stars that indicate whether the coefficient is significantly different from zero. The last column shows the \( R^2 \) when only the last month is included in the regression.

<table>
<thead>
<tr>
<th>Season, ( k )</th>
<th>Int.</th>
<th>Month 1</th>
<th>Month 2</th>
<th>Month 3</th>
<th>( R^2 )</th>
<th>( \hat{\sigma} )</th>
<th>( R^2(3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer, ( k = 1 )</td>
<td>0</td>
<td>0.0779</td>
<td>0.0506</td>
<td>0.7584</td>
<td>**</td>
<td>80.8</td>
<td>28.6</td>
</tr>
<tr>
<td>Summer, ( k = 2 )</td>
<td>0</td>
<td>0.1856</td>
<td>0.2883</td>
<td>0.9055</td>
<td></td>
<td>55.6</td>
<td>21.8</td>
</tr>
<tr>
<td>Summer, ( k = 3 )</td>
<td>-0.0132</td>
<td>0.6684</td>
<td>**</td>
<td>1.2796</td>
<td>***</td>
<td>89.0</td>
<td>10.7</td>
</tr>
<tr>
<td>Summer, ( k = 4 )</td>
<td>0</td>
<td>0.3412</td>
<td>**</td>
<td>0.1676</td>
<td></td>
<td>85.7</td>
<td>9.1</td>
</tr>
<tr>
<td>Fall, ( k = 1 )</td>
<td>0</td>
<td>0.7764</td>
<td>0.2577</td>
<td>1.1009</td>
<td>***</td>
<td>67.3</td>
<td>44.3</td>
</tr>
<tr>
<td>Fall, ( k = 2 )</td>
<td>0</td>
<td>0.0356</td>
<td>0.2805</td>
<td>0.9442</td>
<td>***</td>
<td>65.1</td>
<td>24.1</td>
</tr>
<tr>
<td>Fall, ( k = 3 )</td>
<td>0</td>
<td>0.2615</td>
<td>0.9521</td>
<td>**</td>
<td>0.5682</td>
<td>75.2</td>
<td>18.7</td>
</tr>
<tr>
<td>Fall, ( k = 4 )</td>
<td>0</td>
<td>0.3757</td>
<td>0.7984</td>
<td>*</td>
<td>0.5894</td>
<td>**</td>
<td>65.3</td>
</tr>
<tr>
<td>Winter, ( k = 1 )</td>
<td>0</td>
<td>0.6927</td>
<td>*</td>
<td>0.8948</td>
<td></td>
<td>69.4</td>
<td>57.8</td>
</tr>
<tr>
<td>Winter, ( k = 2 )</td>
<td>0</td>
<td>0.2861</td>
<td>0.4881</td>
<td>0.7964</td>
<td>**</td>
<td>55.8</td>
<td>48.8</td>
</tr>
<tr>
<td>Winter, ( k = 3 )</td>
<td>0</td>
<td>0.6295</td>
<td>**</td>
<td>0.2318</td>
<td></td>
<td>77.2</td>
<td>27.3</td>
</tr>
<tr>
<td>Winter, ( k = 4 )</td>
<td>0</td>
<td>0.5894</td>
<td>*</td>
<td>0.6139</td>
<td></td>
<td>72.0</td>
<td>23.4</td>
</tr>
<tr>
<td>Spring, ( k = 1 )</td>
<td>0</td>
<td>0.2246</td>
<td>*</td>
<td>0.0985</td>
<td></td>
<td>64.9</td>
<td>38.3</td>
</tr>
<tr>
<td>Spring, ( k = 2 )</td>
<td>0</td>
<td>0.1588</td>
<td>-0.0294</td>
<td>0.6909</td>
<td>***</td>
<td>76.7</td>
<td>24.1</td>
</tr>
<tr>
<td>Spring, ( k = 3 )</td>
<td>0</td>
<td>0.2289</td>
<td>0.1152</td>
<td>0.4171</td>
<td></td>
<td>61.6</td>
<td>25.7</td>
</tr>
<tr>
<td>Spring, ( k = 4 )</td>
<td>0</td>
<td>0.0413</td>
<td>0.0875</td>
<td>0.7829</td>
<td>***</td>
<td>77.9</td>
<td>14.9</td>
</tr>
</tbody>
</table>

**code:** '***': \( P < 0.001 \), '**': \( P < 0.01 \), '*': \( P < 0.05 \), '.': \( P < 0.1 \)

### Table 4.3: Correlation coefficients between residuals from the regression of \( \hat{a}_{l,t,k} \) on \( \hat{b}_{m_1,t,k} \), \( \hat{b}_{m_2,t,k} \) and \( \hat{b}_{m_3,t,k} \) (see also Table 4.2) for each season.

<table>
<thead>
<tr>
<th>Season, ( k )</th>
<th>Int.</th>
<th>Month 1</th>
<th>Month 2</th>
<th>Month 3</th>
<th>( R^2 )</th>
<th>( \hat{\sigma} )</th>
<th>( R^2(3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 1</td>
<td>1</td>
<td>0.075</td>
<td>-0.204</td>
<td>0.423</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 2</td>
<td>0.075</td>
<td>1</td>
<td>-0.009</td>
<td>0.143</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 3</td>
<td>-0.204</td>
<td>-0.009</td>
<td>1</td>
<td>-0.348</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 4</td>
<td>0.423</td>
<td>0.143</td>
<td>-0.348</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 1</td>
<td>1</td>
<td>-0.077</td>
<td>-0.171</td>
<td>-0.023</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 2</td>
<td>-0.077</td>
<td>1</td>
<td>-0.150</td>
<td>0.114</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 3</td>
<td>-0.171</td>
<td>-0.150</td>
<td>1</td>
<td>-0.290</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 4</td>
<td>-0.023</td>
<td>0.114</td>
<td>-0.290</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 1</td>
<td>1</td>
<td>0.094</td>
<td>0.045</td>
<td>0.053</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 2</td>
<td>0.094</td>
<td>1</td>
<td>0.198</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 3</td>
<td>0.045</td>
<td>0.198</td>
<td>1</td>
<td>0.098</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 4</td>
<td>0.053</td>
<td>0</td>
<td>0.098</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 1</td>
<td>1</td>
<td>0.041</td>
<td>0.076</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 2</td>
<td>0.041</td>
<td>1</td>
<td>0.535</td>
<td>-0.015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 3</td>
<td>0.076</td>
<td>0.535</td>
<td>1</td>
<td>-0.075</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 4</td>
<td>0.004</td>
<td>-0.015</td>
<td>-0.075</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
red regression lines) consistently showing stronger linear relationships than the off-diagonal graphs. The same is true, more or less, for other months (see Figures B.25 - B.35). It therefore seems when modeling amplitudes $\hat{b}_{m,t,k}$ given amplitudes the month before that there is reason to include all the $\hat{b}_{m-1,t,k}$ amplitudes.

We next consider multiple linear regression of the estimated OMCP amplitudes $\hat{b}_{m,t,k}$ on the first four OMCP amplitudes from the month before, i.e. $\hat{b}_{m-1,t,1}$, $\hat{b}_{m-1,t,2}$, $\hat{b}_{m-1,t,3}$ and $\hat{b}_{m-1,t,4}$, see Table 4.4. There seem to be some predictive potential in the previous month’s amplitudes, although the range of $R^2$ is quite wide from 4% to 73% and most of the $R^2$ values are under 50%.

To investigate the auto-covariance structure of $b_{m,t}$ we calculate the correlations between residuals of the multiple regressions within each month in Table 4.4. These correlations are shown in Table 4.5. Some OMCP amplitudes have strong correlations between groups, for example September groups 2 and 3 (correlation 0.73) and April groups 1 and 2 (correlation 0.93).

Finally, in an effort to improve prediction of OMCP amplitudes, we investigate whether there could be an autoregressive relationship within the same month but between years. We regressed estimated OMCP amplitudes $\hat{b}_{m,t,k}$ on both the whole OMCP amplitude vector from the month before ($\hat{b}_{m-1,t,1}$, $\hat{b}_{m-1,t,2}$, $\hat{b}_{m-1,t,3}$ and $\hat{b}_{m-1,t,4}$) and the same amplitude one year earlier ($\hat{b}_{m,t-1,k}$), see Table 4.6. Only in few cases is the added (previous year) amplitude significant, most notably for March amplitudes. Note that March is a beginning of a new season and so this suggests that prediction of March amplitudes from February amplitudes could be improved by also including March amplitudes form the year before.
Figure 4.16: Estimated December OMCP amplitudes plotted against estimated November OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
<table>
<thead>
<tr>
<th>Month, k</th>
<th>Intercept</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
<th>$R^2$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan, $k = 1$</td>
<td>0</td>
<td>1.014</td>
<td>0.065</td>
<td>-3.50</td>
<td>-0.258</td>
<td>31.3</td>
<td>61.9</td>
</tr>
<tr>
<td>Jan, $k = 2$</td>
<td>0</td>
<td>-0.945</td>
<td>0.255</td>
<td>0.081</td>
<td>-0.092</td>
<td>60.9</td>
<td>9.4</td>
</tr>
<tr>
<td>Jan, $k = 3$</td>
<td>0</td>
<td>-0.108</td>
<td>-0.125</td>
<td>0.339</td>
<td><strong>0.088</strong></td>
<td>38.3</td>
<td>17.3</td>
</tr>
<tr>
<td>Jan, $k = 4$</td>
<td>0</td>
<td>-0.033</td>
<td>0.061</td>
<td>-0.057</td>
<td>0.247</td>
<td>22.2</td>
<td>17.2</td>
</tr>
<tr>
<td>Feb, $k = 1$</td>
<td>0</td>
<td>0.874</td>
<td>***1.031</td>
<td>-0.325</td>
<td>0.502</td>
<td>67.3</td>
<td>41.8</td>
</tr>
<tr>
<td>Feb, $k = 2$</td>
<td>0</td>
<td>0.039</td>
<td>0.979</td>
<td><strong>-0.313</strong></td>
<td>-0.115</td>
<td>73.0</td>
<td>8.2</td>
</tr>
<tr>
<td>Feb, $k = 3$</td>
<td>0</td>
<td>0.007</td>
<td>-0.078</td>
<td>0.491</td>
<td><strong>0.234</strong></td>
<td>40.8</td>
<td>12.5</td>
</tr>
<tr>
<td>Feb, $k = 4$</td>
<td>0</td>
<td>0.036</td>
<td>0.106</td>
<td>0.005</td>
<td><strong>0.287</strong></td>
<td>32.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Mar, $k = 1$</td>
<td>0.482</td>
<td>0.162</td>
<td>0.097</td>
<td><strong>0.217</strong></td>
<td>-0.525</td>
<td>46.7</td>
<td>17.7</td>
</tr>
<tr>
<td>Mar, $k = 2$</td>
<td>0.723</td>
<td>0.327</td>
<td><strong>0.252</strong></td>
<td>-0.336</td>
<td>0.053</td>
<td>40.8</td>
<td>29.4</td>
</tr>
<tr>
<td>Mar, $k = 3$</td>
<td>-0.604</td>
<td>0.007</td>
<td>-0.078</td>
<td>0.491</td>
<td><strong>0.234</strong></td>
<td>49.3</td>
<td>16.4</td>
</tr>
<tr>
<td>Mar, $k = 4$</td>
<td>0.352</td>
<td>0.013</td>
<td>-0.085</td>
<td>0.032</td>
<td>-0.826</td>
<td><strong>32.4</strong></td>
<td>13.8</td>
</tr>
<tr>
<td>Apr, $k = 1$</td>
<td>0</td>
<td>0.461</td>
<td>0.212</td>
<td>-0.277</td>
<td>-0.458</td>
<td>25.5</td>
<td>29.2</td>
</tr>
<tr>
<td>Apr, $k = 2$</td>
<td>0</td>
<td>-0.693</td>
<td>0.984</td>
<td>-0.748</td>
<td>-0.340</td>
<td>30.2</td>
<td>32.5</td>
</tr>
<tr>
<td>Apr, $k = 3$</td>
<td>0</td>
<td>0.163</td>
<td>-0.317</td>
<td>0.630</td>
<td>-0.253</td>
<td>23.3</td>
<td>17.5</td>
</tr>
<tr>
<td>Apr, $k = 4$</td>
<td>0</td>
<td>-0.042</td>
<td>-0.090</td>
<td>0.138</td>
<td>-0.472</td>
<td>25.8</td>
<td>15.0</td>
</tr>
<tr>
<td>May, $k = 1$</td>
<td>1.618</td>
<td><strong>-1.308</strong></td>
<td><strong>-0.447</strong></td>
<td>-0.657</td>
<td>41.9</td>
<td>36.0</td>
<td></td>
</tr>
<tr>
<td>May, $k = 2$</td>
<td>0.085</td>
<td>0.201</td>
<td>0.146</td>
<td>0.311</td>
<td>13.2</td>
<td>28.2</td>
<td></td>
</tr>
<tr>
<td>May, $k = 3$</td>
<td>0.125</td>
<td>-0.093</td>
<td>0.693</td>
<td><strong>0.016</strong></td>
<td>33.1</td>
<td>21.7</td>
<td></td>
</tr>
<tr>
<td>May, $k = 4$</td>
<td>0</td>
<td>-0.205</td>
<td>0.756</td>
<td>19.3</td>
<td>26.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jun, $k = 1$</td>
<td>0.471</td>
<td>-0.373</td>
<td>0.257</td>
<td>0.033</td>
<td>-0.663</td>
<td>26.1</td>
<td>43.1</td>
</tr>
<tr>
<td>Jun, $k = 2$</td>
<td>1.324</td>
<td>-0.555</td>
<td>-0.248</td>
<td>-0.298</td>
<td>-0.121</td>
<td>14.7</td>
<td>47.4</td>
</tr>
<tr>
<td>Jun, $k = 3$</td>
<td>-2.416</td>
<td>0.129</td>
<td>0.596</td>
<td>0.853</td>
<td><strong>-0.194</strong></td>
<td>42.6</td>
<td>33.3</td>
</tr>
<tr>
<td>Jun, $k = 4$</td>
<td>0.780</td>
<td>0.134</td>
<td>-0.228</td>
<td>-0.173</td>
<td>0.053</td>
<td>18.4</td>
<td>22.6</td>
</tr>
<tr>
<td>Jul, $k = 1$</td>
<td>0</td>
<td>0.472</td>
<td>0.015</td>
<td>-0.341</td>
<td>0.068</td>
<td>29.5</td>
<td>31.4</td>
</tr>
<tr>
<td>Jul, $k = 2$</td>
<td>0</td>
<td>-0.078</td>
<td>0.278</td>
<td>-0.030</td>
<td>-0.019</td>
<td>19.9</td>
<td>29.3</td>
</tr>
<tr>
<td>Jul, $k = 3$</td>
<td>0</td>
<td>-0.199</td>
<td>-0.083</td>
<td>0.598</td>
<td>0.430</td>
<td>24.1</td>
<td>43.3</td>
</tr>
<tr>
<td>Jul, $k = 4$</td>
<td>0</td>
<td>-0.022</td>
<td>0.059</td>
<td>0.004</td>
<td>0.362</td>
<td>13.7</td>
<td>26.1</td>
</tr>
<tr>
<td>Aug, $k = 1$</td>
<td>1.432</td>
<td><strong>-1.240</strong></td>
<td>0.231</td>
<td>1.785</td>
<td><strong>49.0</strong></td>
<td>49.2</td>
<td></td>
</tr>
<tr>
<td>Aug, $k = 2$</td>
<td>0</td>
<td>-0.069</td>
<td>0.710</td>
<td>-0.030</td>
<td>-0.341</td>
<td>19.7</td>
<td>44.6</td>
</tr>
<tr>
<td>Aug, $k = 3$</td>
<td>0</td>
<td>-0.066</td>
<td>-0.288</td>
<td>0.723</td>
<td><strong>0.668</strong></td>
<td>48.1</td>
<td>32.9</td>
</tr>
<tr>
<td>Aug, $k = 4$</td>
<td>0</td>
<td>0.528</td>
<td>-0.006</td>
<td>0.179</td>
<td>0.872</td>
<td>28.3</td>
<td>41.9</td>
</tr>
<tr>
<td>Sep, $k = 1$</td>
<td>0</td>
<td>-0.967</td>
<td><strong>-0.791</strong></td>
<td>0.588</td>
<td>-0.021</td>
<td>38.9</td>
<td>82.9</td>
</tr>
<tr>
<td>Sep, $k = 2$</td>
<td>0</td>
<td>0.188</td>
<td>-0.409</td>
<td>0.224</td>
<td>0.150</td>
<td>20.2</td>
<td>60.3</td>
</tr>
<tr>
<td>Sep, $k = 3$</td>
<td>0</td>
<td>0.227</td>
<td>-0.148</td>
<td>-0.466</td>
<td>0.020</td>
<td>21.8</td>
<td>49.8</td>
</tr>
<tr>
<td>Sep, $k = 4$</td>
<td>0</td>
<td>0.418</td>
<td><strong>0.191</strong></td>
<td>-0.417</td>
<td>-0.084</td>
<td>40.0</td>
<td>35.8</td>
</tr>
<tr>
<td>Oct, $k = 1$</td>
<td>0</td>
<td>0.383</td>
<td><strong>-0.480</strong></td>
<td>0.580</td>
<td>0.104</td>
<td>37.2</td>
<td>64.0</td>
</tr>
<tr>
<td>Oct, $k = 2$</td>
<td>0</td>
<td>0.051</td>
<td>0.380</td>
<td>-0.097</td>
<td>0.229</td>
<td>19.1</td>
<td>49.4</td>
</tr>
<tr>
<td>Oct, $k = 3$</td>
<td>0</td>
<td>-0.091</td>
<td>-0.003</td>
<td>0.355</td>
<td>0.144</td>
<td>32.7</td>
<td>35.0</td>
</tr>
<tr>
<td>Oct, $k = 4$</td>
<td>0</td>
<td>-0.214</td>
<td>0.057</td>
<td>0.503</td>
<td>0.132</td>
<td>38.9</td>
<td>47.7</td>
</tr>
<tr>
<td>Nov, $k = 1$</td>
<td>0</td>
<td>0.302</td>
<td><strong>-0.227</strong></td>
<td>0.151</td>
<td>0.428</td>
<td>42.0</td>
<td>42.9</td>
</tr>
<tr>
<td>Nov, $k = 2$</td>
<td>0</td>
<td>-0.109</td>
<td>0.421</td>
<td>-0.116</td>
<td>-0.015</td>
<td>16.3</td>
<td>53.0</td>
</tr>
<tr>
<td>Nov, $k = 3$</td>
<td>0</td>
<td>-0.092</td>
<td>-0.112</td>
<td>0.869</td>
<td><strong>0.090</strong></td>
<td>59.5</td>
<td>31.6</td>
</tr>
<tr>
<td>Nov, $k = 4$</td>
<td>0</td>
<td>0.092</td>
<td>-0.061</td>
<td>0.214</td>
<td>0.235</td>
<td>40.9</td>
<td>24.1</td>
</tr>
<tr>
<td>Dec, $k = 1$</td>
<td>0</td>
<td>0.439</td>
<td><strong>0.152</strong></td>
<td>0.306</td>
<td><strong>-0.951</strong></td>
<td>56.7</td>
<td>25.8</td>
</tr>
<tr>
<td>Dec, $k = 2$</td>
<td>0</td>
<td>-0.495</td>
<td>0.142</td>
<td>-0.031</td>
<td>0.527</td>
<td>31.0</td>
<td>38.3</td>
</tr>
<tr>
<td>Dec, $k = 3$</td>
<td>0</td>
<td>-0.068</td>
<td>0.030</td>
<td>-0.108</td>
<td>-0.019</td>
<td>4.4</td>
<td>35.6</td>
</tr>
<tr>
<td>Dec, $k = 4$</td>
<td>0</td>
<td>-0.075</td>
<td>0.231</td>
<td><strong>-0.061</strong></td>
<td>0.087</td>
<td>30.0</td>
<td>22.4</td>
</tr>
</tbody>
</table>

Table 4.4: Results from regressing estimated OMCP amplitudes, $\hat{b}_{m,t,k}$, on the OMCP amplitudes the month before, $\hat{b}_{m-1,t,1}$, $\hat{b}_{m-1,t,2}$, $\hat{b}_{m-1,t,3}$ and $\hat{b}_{m-1,t,4}$, for July through December. Columns 2 through 6 show the estimated regression coefficients and stars that indicate whether the coefficient is significantly different from zero.
<table>
<thead>
<tr>
<th>January</th>
<th>February</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 1$</td>
<td>$k = 1$</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>$k = 2$</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>$k = 3$</td>
</tr>
<tr>
<td>$k = 4$</td>
<td>$k = 4$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-0.263</td>
<td>-0.224</td>
</tr>
<tr>
<td>0.410</td>
<td>0.176</td>
</tr>
<tr>
<td>-0.196</td>
<td>-0.195</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.240</td>
<td>1</td>
</tr>
<tr>
<td>0.448</td>
<td>0.105</td>
</tr>
<tr>
<td>-0.237</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-0.196</td>
<td>-0.224</td>
</tr>
<tr>
<td>0.448</td>
<td>1</td>
</tr>
<tr>
<td>-0.237</td>
<td>0.283</td>
</tr>
<tr>
<td>1</td>
<td>-0.008</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.725</td>
<td>0.932</td>
</tr>
<tr>
<td>0.735</td>
<td>0.259</td>
</tr>
<tr>
<td>-0.341</td>
<td>-0.004</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.725</td>
<td>1</td>
</tr>
<tr>
<td>0.886</td>
<td>0.063</td>
</tr>
<tr>
<td>-0.632</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.71</td>
<td>0.089</td>
</tr>
<tr>
<td>0.446</td>
<td>0.072</td>
</tr>
<tr>
<td>-0.328</td>
<td>-0.135</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.299</td>
<td>0.299</td>
</tr>
<tr>
<td>0.469</td>
<td>0.055</td>
</tr>
<tr>
<td>-0.172</td>
<td>-0.083</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.299</td>
<td>0.089</td>
</tr>
<tr>
<td>0.733</td>
<td>0.055</td>
</tr>
<tr>
<td>-0.172</td>
<td>-0.177</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.312</td>
<td>0.344</td>
</tr>
<tr>
<td>0.078</td>
<td>0.462</td>
</tr>
<tr>
<td>-0.205</td>
<td>-0.110</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.5: Correlation coefficients between residuals from the regression of $\hat{b}_{m,t,k}$ on $\hat{b}_{m-1,t,k_2}$ for $k_2 = 1, 2, 3, 4$, (see also Table 4.4) for each month.
<table>
<thead>
<tr>
<th>Month, k</th>
<th>Int.</th>
<th>k = 1</th>
<th>k = 2</th>
<th>k = 3</th>
<th>k = 4</th>
<th>Last year</th>
<th>$R^2$</th>
<th>$\hat{\phi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan, k = 1</td>
<td>0.17</td>
<td>0.92</td>
<td>0.03</td>
<td>-0.35</td>
<td>-0.28</td>
<td>0.13</td>
<td>31.9</td>
<td>65.2</td>
</tr>
<tr>
<td>Jan, k = 2</td>
<td>-0.87</td>
<td>-0.04</td>
<td>0.19 *</td>
<td>0.03</td>
<td>-0.13</td>
<td>0.24</td>
<td>60.9</td>
<td>8.9</td>
</tr>
<tr>
<td>Jan, k = 3</td>
<td>-1.99</td>
<td>-0.13</td>
<td>-0.16</td>
<td>0.30 *</td>
<td>-0.06</td>
<td>-0.16</td>
<td>48.8</td>
<td>15.5</td>
</tr>
<tr>
<td>Jan, k = 4</td>
<td>-0.35</td>
<td>-0.04</td>
<td>0.05</td>
<td>-0.07</td>
<td>0.23</td>
<td>0.02</td>
<td>16.9</td>
<td>18.3</td>
</tr>
<tr>
<td>Feb, k = 1</td>
<td>-0.67</td>
<td>0.83 ***</td>
<td>1.12</td>
<td>-0.46</td>
<td>0.26</td>
<td>0.18</td>
<td>70.4</td>
<td>42.3</td>
</tr>
<tr>
<td>Feb, k = 2</td>
<td>0.26</td>
<td>0.04</td>
<td>0.91 ***</td>
<td>-0.28 *</td>
<td>-0.12</td>
<td>0.19</td>
<td>75.7</td>
<td>8.2</td>
</tr>
<tr>
<td>Feb, k = 3</td>
<td>-1.27</td>
<td>0.01</td>
<td>-0.22</td>
<td>0.37 *</td>
<td>0.16</td>
<td>-0.01</td>
<td>30.4</td>
<td>12.0</td>
</tr>
<tr>
<td>Feb, k = 4</td>
<td>-0.84</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.25</td>
<td>0.08</td>
<td>30.5</td>
<td>8.9</td>
</tr>
<tr>
<td>Mar, k = 1</td>
<td>-2.34</td>
<td>0.18 **</td>
<td>0.67 *</td>
<td>-0.23</td>
<td>-0.80 *</td>
<td>0.43 **</td>
<td>68.6</td>
<td>14.0</td>
</tr>
<tr>
<td>Mar, k = 2</td>
<td>-4.16</td>
<td>0.34 ***</td>
<td>0.20</td>
<td>-1.09 *</td>
<td>-0.19</td>
<td>0.54 *</td>
<td>65.4</td>
<td>23.9</td>
</tr>
<tr>
<td>Mar, k = 3</td>
<td>-2.60</td>
<td>0.22 ***</td>
<td>-0.18</td>
<td>-0.13</td>
<td>-0.38</td>
<td>0.35 *</td>
<td>63.6</td>
<td>14.3</td>
</tr>
<tr>
<td>Mar, k = 4</td>
<td>0.72</td>
<td>0.00</td>
<td>0.09</td>
<td>0.10</td>
<td>-0.87 *</td>
<td>0.31</td>
<td>35.4</td>
<td>14.0</td>
</tr>
<tr>
<td>Apr, k = 1</td>
<td>1.17</td>
<td>0.63</td>
<td>0.16</td>
<td>-0.08</td>
<td>-0.87</td>
<td>-0.62</td>
<td>37.2</td>
<td>28.4</td>
</tr>
<tr>
<td>Apr, k = 2</td>
<td>2.21</td>
<td>0.04</td>
<td>0.91</td>
<td>0.07</td>
<td>-0.39</td>
<td>-0.92</td>
<td>48.1</td>
<td>29.7</td>
</tr>
<tr>
<td>Apr, k = 3</td>
<td>-0.09</td>
<td>0.14</td>
<td>-0.31</td>
<td>0.65</td>
<td>-0.30</td>
<td>-0.09</td>
<td>22.5</td>
<td>18.4</td>
</tr>
<tr>
<td>Apr, k = 4</td>
<td>0.29</td>
<td>0.02</td>
<td>-0.12</td>
<td>0.10</td>
<td>0.37</td>
<td>0.27</td>
<td>30.6</td>
<td>15.4</td>
</tr>
<tr>
<td>May, k = 1</td>
<td>-0.23</td>
<td>1.65 **</td>
<td>-1.36 **</td>
<td>-0.43</td>
<td>-0.78</td>
<td>-0.13</td>
<td>42.6</td>
<td>37.9</td>
</tr>
<tr>
<td>May, k = 2</td>
<td>0.24</td>
<td>0.08</td>
<td>0.24</td>
<td>0.12</td>
<td>0.42</td>
<td>0.15</td>
<td>16.2</td>
<td>29.3</td>
</tr>
<tr>
<td>May, k = 3</td>
<td>0.32</td>
<td>0.13</td>
<td>-0.09</td>
<td>0.66</td>
<td>0.07</td>
<td>0.22</td>
<td>36.0</td>
<td>22.5</td>
</tr>
<tr>
<td>May, k = 4</td>
<td>1.07</td>
<td>0.21</td>
<td>0.18</td>
<td>0.08</td>
<td>0.65</td>
<td>0.04</td>
<td>31.8</td>
<td>25.6</td>
</tr>
<tr>
<td>Jun, k = 1</td>
<td>-6.83</td>
<td>-0.32</td>
<td>0.38</td>
<td>0.05</td>
<td>-0.45</td>
<td>-0.13</td>
<td>12.6</td>
<td>29.0</td>
</tr>
<tr>
<td>Jun, k = 2</td>
<td>4.14</td>
<td>-0.43</td>
<td>-0.35</td>
<td>-0.31</td>
<td>-0.19</td>
<td>-0.12</td>
<td>18.1</td>
<td>48.2</td>
</tr>
<tr>
<td>Jun, k = 3</td>
<td>-6.85</td>
<td>0.16</td>
<td>0.68</td>
<td>0.92 **</td>
<td>-0.09</td>
<td>-0.12</td>
<td>58.4</td>
<td>26.8</td>
</tr>
<tr>
<td>Jun, k = 4</td>
<td>0.70</td>
<td>0.20</td>
<td>-0.16</td>
<td>-0.15</td>
<td>0.05</td>
<td>0.41</td>
<td>33.4</td>
<td>21.8</td>
</tr>
<tr>
<td>Jul, k = 1</td>
<td>0.58</td>
<td>0.53</td>
<td>0.05</td>
<td>-0.31</td>
<td>0.09</td>
<td>0.22</td>
<td>33.5</td>
<td>31.4</td>
</tr>
<tr>
<td>Jul, k = 2</td>
<td>2.10</td>
<td>0.06</td>
<td>0.20</td>
<td>0.06</td>
<td>0.17</td>
<td>0.10</td>
<td>17.2</td>
<td>30.3</td>
</tr>
<tr>
<td>Jul, k = 3</td>
<td>-2.35</td>
<td>-0.28</td>
<td>0.00</td>
<td>0.50</td>
<td>0.30</td>
<td>0.28</td>
<td>28.4</td>
<td>43.9</td>
</tr>
<tr>
<td>Jul, k = 4</td>
<td>0.17</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.00</td>
<td>0.38</td>
<td>-0.01</td>
<td>13.3</td>
<td>27.8</td>
</tr>
<tr>
<td>Aug, k = 1</td>
<td>0.45</td>
<td>1.49 * 1.44</td>
<td>0.27</td>
<td>1.81 **</td>
<td>0.29</td>
<td>45.0</td>
<td>51.3</td>
<td></td>
</tr>
<tr>
<td>Aug, k = 2</td>
<td>-1.01</td>
<td>-0.05</td>
<td>-0.36</td>
<td>-0.11</td>
<td>-0.30</td>
<td>-0.51</td>
<td>22.0</td>
<td>46.5</td>
</tr>
<tr>
<td>Aug, k = 3</td>
<td>-1.63</td>
<td>-0.20</td>
<td>-0.13</td>
<td>0.61 * 0.46</td>
<td>0.15</td>
<td>48.2</td>
<td>33.7</td>
<td></td>
</tr>
<tr>
<td>Aug, k = 4</td>
<td>-0.82</td>
<td>0.30</td>
<td>0.19</td>
<td>0.17</td>
<td>0.89</td>
<td>0.40</td>
<td>31.2</td>
<td>43.3</td>
</tr>
<tr>
<td>Sep, k = 1</td>
<td>11.71</td>
<td>-0.75 * -0.97</td>
<td>0.94 * 0.26</td>
<td>-0.28</td>
<td>50.5</td>
<td>66.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep, k = 2</td>
<td>1.36</td>
<td>0.25</td>
<td>-0.35</td>
<td>-0.19</td>
<td>0.06</td>
<td>0.27</td>
<td>27.3</td>
<td>61.3</td>
</tr>
<tr>
<td>Sep, k = 3</td>
<td>5.89</td>
<td>0.32</td>
<td>-0.10</td>
<td>0.17</td>
<td>0.26</td>
<td>-0.33</td>
<td>31.2</td>
<td>47.1</td>
</tr>
<tr>
<td>Sep, k = 4</td>
<td>1.11</td>
<td>0.51 ** 0.16</td>
<td>-0.38</td>
<td>-0.08</td>
<td>-0.30</td>
<td>50.9</td>
<td>34.4</td>
<td></td>
</tr>
<tr>
<td>Oct, k = 1</td>
<td>3.27</td>
<td>0.29</td>
<td>-0.37</td>
<td>0.43</td>
<td>0.08</td>
<td>0.21</td>
<td>26.0</td>
<td>62.8</td>
</tr>
<tr>
<td>Oct, k = 2</td>
<td>-2.00</td>
<td>0.11</td>
<td>0.32</td>
<td>-0.02</td>
<td>0.30</td>
<td>0.07</td>
<td>20.5</td>
<td>52.0</td>
</tr>
<tr>
<td>Oct, k = 3</td>
<td>-2.05</td>
<td>-0.05</td>
<td>-0.12</td>
<td>0.52</td>
<td>0.06</td>
<td>-0.10</td>
<td>36.8</td>
<td>35.9</td>
</tr>
<tr>
<td>Oct, k = 4</td>
<td>2.75</td>
<td>-0.30</td>
<td>0.16</td>
<td>0.38</td>
<td>0.11</td>
<td>-0.09</td>
<td>42.6</td>
<td>49.0</td>
</tr>
<tr>
<td>Nov, k = 1</td>
<td>0.40</td>
<td>0.29</td>
<td>-0.22</td>
<td>0.15</td>
<td>0.42</td>
<td>0.02</td>
<td>36.6</td>
<td>45.0</td>
</tr>
<tr>
<td>Nov, k = 2</td>
<td>2.82</td>
<td>-0.20</td>
<td>0.47</td>
<td>-0.07</td>
<td>-0.05</td>
<td>-0.01</td>
<td>20.3</td>
<td>54.9</td>
</tr>
<tr>
<td>Nov, k = 3</td>
<td>-1.08</td>
<td>-0.02</td>
<td>-0.11</td>
<td>0.84 *** 0.06</td>
<td>0.30</td>
<td>64.5</td>
<td>31.1</td>
<td></td>
</tr>
<tr>
<td>Nov, k = 4</td>
<td>-1.08</td>
<td>0.10</td>
<td>-0.09</td>
<td>0.17</td>
<td>0.23</td>
<td>-0.14</td>
<td>50.6</td>
<td>23.5</td>
</tr>
<tr>
<td>Dec, k = 1</td>
<td>-0.77</td>
<td>0.47 ** 0.18</td>
<td>0.23</td>
<td>-0.92 ** 0.19</td>
<td>64.1</td>
<td>24.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dec, k = 2</td>
<td>-2.42</td>
<td>-0.38</td>
<td>0.17</td>
<td>-0.08</td>
<td>0.39</td>
<td>-0.03</td>
<td>25.8</td>
<td>38.9</td>
</tr>
<tr>
<td>Dec, k = 3</td>
<td>-1.02</td>
<td>-0.01</td>
<td>0.05</td>
<td>-0.11</td>
<td>-0.09</td>
<td>-0.06</td>
<td>5.3</td>
<td>37.3</td>
</tr>
<tr>
<td>Dec, k = 4</td>
<td>-2.90</td>
<td>0.06</td>
<td>0.27 ** -0.13</td>
<td>-0.05</td>
<td>0.05</td>
<td>45.2</td>
<td>18.8</td>
<td></td>
</tr>
</tbody>
</table>

*code: ***: $P < 0.001$, **: $P < 0.01$, *: $P < 0.05$, .: $P < 0.1$*

Table 4.6: Results from regressing estimated OMCP amplitudes, $b_{m,t,k}$, on the OMCP amplitudes the month before, $b_{m-1,t,k}$, and the amplitude for same month but the year before, $b_{m, t-1,k}$. Columns 2 through 6 show the estimated regression coefficients and stars that indicate whether the coefficient is significantly different from zero.
4.5 Bayesian hierarchical model

Here we describe a hierarchical Bayesian model in the spirit of Section 2.2 (and Section 2.3) for the Polar MM5 and ERA-40 temperature data using the EOFs and OMCPs as basis vectors. The primary goal is statistical downscaling of the coarse-resolution ERA-40 data onto the high-resolution Polar MM5 grid, or in other words to predict Polar MM5 data based on the ERA-40 data. The model is defined hierarchically via a data model, process model and parameter model, introduced in Sections 4.5.1, 4.5.2 and 4.5.3, respectively.

4.5.1 Data model

In addition to the notation introduced in Section 4.3, we define the collection of all $Y_{l,t}$ and $X_{m,t}$ data vectors as follows:

$Y_{1:T}$ is the collection of all $N_Y$-dimensional seasonal Polar MM5 data vectors, and $X_{1:T}$ is the collection of all $N_X$-dimensional seasonal ERA data vectors,

that is,

$$Y_{1:T} = \{ Y_{f,1}, Y_{w,1}, Y_{p,1}, Y_{s,2}, Y_{f,2}, Y_{w,2}, Y_{p,2}, \ldots, Y_{p,22}, Y_{s,23}, Y_{f,23} \} \quad (4.5)$$

$$X_{1:T} = \{ X_{3,1}, X_{4,1}, \ldots, X_{12,1}, X_{1,2}, X_{2,2}, X_{3,2}, \ldots, X_{4,23}, X_{5,23} \} \quad (4.6)$$

where the first year is 1979 and the 23rd year is 2001, and the indexes $f$, $w$, $p$ and $s$ stand for fall, winter, spring and summer, respectively.

The data model defines the distribution of the centered data vectors given the basis vectors and their amplitudes. We assume that the observations are conditionally
independent given the amplitude vectors $a_{l,t}$ and $b_{m,t}$,

$$
\left[ Y_{1:T}, X_{1:T} \mid a_{1:T}, b_{1:T}, R_{1:T}, S_{1:T} \right] = \prod_{t=s,f,w,p} \prod_{t=t_1}^{T_t} [Y_{t,t} \mid a_{t,t}, R_{t,t}] \times \prod_{m=1}^{12} \prod_{t=t_m}^{T_m} [X_{m,t} \mid b_{m,t}, S_{m,t}] .
$$

(4.7)

The notation $[X]$ stands for “the distribution of $X$” and the indexes “$1 : T$” indicate all time points where we have data, just as in (4.5) and (4.6). For each time point (month $m$ or season $l$) we assume a normal error structure,

$$
[Y_{t,t} \mid a_{t,t}, R_{t,t}] = N(U_l a_{t,t}, R_{t,t}) \quad \text{and} \quad [X_{m,t} \mid b_{m,t}, S_{m,t}] = N(V_m b_{m,t}, S_{m,t})
$$

(4.8)

where $R_{t,t}$ and $S_{m,t}$ are unknown variance-covariance matrices and $U_l$ and $V_m$ are the first four EOFs and OMCPs for season $l$ and month $m$; that is, $K_Y = 4$ and $K_X = 4$. The selection of only four patterns is based on the preliminary analysis in Section 4.3.

There we found that the first four EOFs explained 70%-80% of the total variance of the data, see Figure 4.13 and Table 4.1.

4.5.2 Process model

The process model defines the joint distribution of the amplitude vectors $a_{l,t}$ and $b_{m,t}$. We assume that the Polar MM5 amplitude vectors, $a_{l,t}$, are conditionally independent given the ERA amplitude vectors, $b_{m_1,t}$, $b_{m_1,t}$ and $b_{m_1,t}$, and that the
\( b_{m,t} \) vectors have a first-order Markovian structure:

\[
\begin{align*}
&\left[ a_{1:T}, b_{1:T} | H_{1:4}, B_{1:12}, C_{1:4}, D_{1:12} \right] = \\
&\left[ b_{2,1} \right] \prod_{t = \tau_s}^{T_s} \left[ a_{s,t} | b_{12,t-1}, b_{1,t}, b_{2,t}, H_s, C_s \right] \prod_{t = \tau_f}^{T_f} \left[ a_{f,t} | b_{3,t}, b_{4,t}, b_{5,t}, H_f, C_f \right] \\
&\times \prod_{t = \tau_w}^{T_w} \left[ a_{w,t} | b_{6,t}, b_{7,t}, b_{8,t}, H_w, C_w \right] \prod_{t = \tau_p}^{T_p} \left[ a_{p,t} | b_{9,t}, b_{10,t}, b_{11,t}, H_p, C_p \right] \\
&\times \prod_{m=1}^{12} \prod_{t = \tau_m}^{T_m} \left[ b_{m,t} | b_{m-1,t}, B_m, D_m \right].
\end{align*}
\]

(4.9)

The dependence structure in (4.9) is easily described in the following illustration (or Bayesian network), where each amplitude vector, given the vectors that point to it, is conditionally independent of all other amplitude vectors:

Note that in the exploratory analysis of estimated amplitudes in Section 4.4, there was some indication that, at least for some months, it might be beneficial to model each \( b_{m,t} \) amplitude vector dependent on both the month before and also the same month one year earlier, see Table 4.6. However, this only applied for a few months and amplitudes, so for the sake of parsimony, we will not include such dependence structure here.
For all seasons we assume a normal distribution for the Polar MM5 amplitudes given the ERA amplitudes:

\[
\begin{align*}
\{a_s,t | b_{12,t-1}, b_{1,t}, b_{2,t} H_s, C_s\} &= N\left( H_s \begin{bmatrix} b_{12,t-1} \\ b_{1,t} \\ b_{2,t} \end{bmatrix}, C_s \right) \\
\{a_f,t | b_{3,t}, b_{4,t}, b_{5,t} H_f, C_f\} &= N\left( H_f \begin{bmatrix} b_{3,t} \\ b_{4,t} \\ b_{5,t} \end{bmatrix}, C_f \right) \\
\{a_w,t | b_{6,t}, b_{7,t}, b_{8,t} H_w, C_w\} &= N\left( H_w \begin{bmatrix} b_{6,t} \\ b_{7,t} \\ b_{8,t} \end{bmatrix}, C_w \right) \\
\{a_p,t | b_{9,t}, b_{10,t}, b_{11,t} H_p, C_p\} &= N\left( H_p \begin{bmatrix} b_{9,t} \\ b_{10,t} \\ b_{11,t} \end{bmatrix}, C_p \right)
\end{align*}
\]

(4.10)

where \(C_s, C_f, C_w\) and \(C_p\) are unknown \(K_Y \times K_Y\) dimensional covariance matrices and

\[
H_s = \begin{pmatrix} H_{12} & H_1 & H_2 \end{pmatrix}, \quad H_f = \begin{pmatrix} H_3 & H_4 & H_5 \end{pmatrix}, \\
H_w = \begin{pmatrix} H_6 & H_7 & H_8 \end{pmatrix} \quad \text{and} \quad H_p = \begin{pmatrix} H_9 & H_{10} & H_{11} \end{pmatrix}.
\]

(4.11)

Each of the \(H_m, m = 1,\ldots, 12\), matrices are \(K_Y \times K_X\) dimensional, so each \(H_l\) matrix is \(K_Y \times 3K_X\) dimensional, \(l = s, f, w, p\). Based on the preliminary data analysis in Section 4.4, we make the \(H_m\) matrices diagonal.

The \(b_{m,t}\) amplitude vectors are modeled by a first order Markov process with normally distributed errors,

\[
\begin{align*}
\{b_{m,t} | b_{m-1,t}, B_m, D_m\} &= N\left( B_m b_{m-1,t}, D_m \right)
\end{align*}
\]

(4.12)

where \(D_m\) is an unknown \(K_X \times K_X\) dimensional covariance matrix. Note that here we use the notation \(b_{0,t}\) for \(b_{12,t-1}\). We let the \(B_m\) matrices have full structure, that is we have \(K_X^2\) unknowns for each \(B_m\) matrix.

Finally we use a normal prior on the first ERA amplitude vector,

\[
\{b_{2,1}\} = N\left( \mu_b, \Sigma_b \right).
\]

(4.13)
4.5.3 Parameter model

The parameter model defines prior distributions on unknown parameters introduced in the data and process models. For ease of computation, we use conjugate prior distributions. There are three groups of parameters we need to consider: (I) the transition matrices $H_m$ and $B_m$; (II) the process model covariance matrices $C_l$ and $D_m$; and (III) the data model covariance matrices $R_{l,t}$ and $S_{m,t}$.

(I) Transition matrices $H_m$ and $B_m$: Since the $H_m$ matrices are assumed diagonal, there are $K_Y = 4$ parameters for each $H_m$ matrix, but the $B_m$ include $K_X^2 = 16$ parameters each. Let $h_m = (H_{m,1}, \ldots, H_{m,K_Y})'$ be a vector that contains the diagonal elements of $H_m$. Let $\text{vec}(B_m)$ be the vectorization of $B_m$, i.e. $\text{vec}(B_m)$ is a column vector with the columns of $B_m$ stacked on top of each other. We assume that the $H_m$ and $B_m$ matrices are apriori independent and we use the following prior distributions for the elements of the $H_m$ and $B_m$ matrices:

$$[h_m] = N(\mu_{1,m}, \Sigma_{1,m}), \quad m = 1, \ldots, 12 \quad \text{and}$$

$$[\text{vec}(B_m)] = N(\mu_{2,m}, \Sigma_{2,m}), \quad m = 1, \ldots, 12.$$  

(4.14)

The parameters $\mu_{1,m}$, $\Sigma_{1,m}$, $\mu_{2,m}$ and $\Sigma_{2,m}$ are constant hyperparameters. We discuss how they are selected in Section 4.5.3.1.

(II) Process model covariance matrices $C_l$ and $D_m$: The covariance matrices $C_l$ and $D_m$ describe the (conditional) covariance between elements of the amplitude vectors $a_{l,t}$ and $b_{m,t}$. It is not straightforward to visualize a natural covariance model for the amplitudes of basis vectors. One option is to assume that these $K_Y \times K_Y$ and $K_X \times K_X$ covariance matrices are diagonal, but with different variances along
the diagonal. Here we choose to allow for dependence between the amplitudes, in light of the correlations shown in Tables 4.3 and 4.5 in Section 4.4. A conjugate prior for the covariance matrices $C_l$ and $D_m$ is the inverse-Wishart (IW) distribution. We assume that all $C_l$ and $D_m$ are apriori independent and we use the following prior distributions:

$$C_l \sim IW_{\nu_l}(W_l), \quad l = s, f, w, p, \quad \text{and} \quad D_m \sim IW_{\nu_m}(W_m), \quad m = 1, \ldots, 12 \quad (4.15)$$

The hyperparameters $W_l$ and $W_m$ are positive definite matrices and $\nu_l$ and $\nu_m$ are degrees of freedom, see Section 4.5.3.1.

(III) Data model covariances $R_{l,t}$ and $S_{m,t}$: The covariance matrices in the data model are very large; $R_{l,t}$ are $(N_Y \times N_Y)$-dimensional and $S_{m,t}$ are $(N_X \times N_X)$-dimensional. $R_{l,t}$ and $S_{m,t}$ account for many sources of variation that arise when modeling the observations given the coefficient vectors $a_{l,t}$ and $b_{m,t}$. These include measurement error, small non-time-varying variation in the processes and error due to the dimension reduction itself. One way to avoid the high dimensional covariance matrices is to assume (conditional) independence and take $R_{l,t}$ and $S_{m,t}$ as diagonal matrices. But in light of the many complex source of variation in $R_{l,t}$ and $S_{m,t}$ that seems too simplistic. Instead we apply an approach used in Berliner et al. (2000) which is both computationally appealing and attempts to account for some of the structure left over after the dimension reduction. The idea is to form $R_{l,t}$ and $S_{m,t}$ by using a few of the next basis vectors to capture some of what is left of the spatial structure in $Y_{l,t}$ and $X_{m,t}$. In principle $R_{l,t}$ and $S_{m,t}$ could be time-varying (different for each year $t$), but here we assume that they only vary between seasons or months and can therefore be denoted as $R_l$ and $S_m$. 

106
Let $\tilde{U}_l$ be a matrix that contains columns $K_Y + 1$ through $L_Y$ of $U_l$ and let $\tilde{D}_l$ be a diagonal matrix containing the corresponding eigenvalues, $d_{l,j}$, i.e. the variances associated with these EOFs. As in Berliner et al. (2000) we set

$$R_l = r_l \left( c_l I_{N_Y} + \sum_{j=K_Y+1}^{L_Y} d_{l,j} u_{l,j} u_{l,j}' \right) = r_l \left( c_l I_{N_Y} + \tilde{U}_l \tilde{D}_l \tilde{U}_l' \right) \equiv r_l \tilde{R}_w .$$

(4.16)

Here, $c_l$ is a constant set to be the total sample variance left after accounting for the first $L_Y$ EOFs. That is,

$$c_l = \sum_{k=K_Y+1}^{N_Y} d_{l,k} \quad \text{for } l = s, f, w, p$$

(4.17)

where $d_{l,k}$ are the eigenvalues of $S_Y = \frac{1}{T_l-1} Y_l Y_l'$. The values of $d_{l,k}$ and $c_l$ are listed in Table 4.7. In practice we obtain $\sqrt{d_{l,k}}$ as the singular values of $\frac{1}{\sqrt{T_l-1}} Y_l$ since the SVD of $\frac{1}{\sqrt{T_l-1}} Y_l$ is computationally feasible but even storing $\frac{1}{T_l-1} Y_l Y_l'$ is problematic. Note that, although $\tilde{U}_l \tilde{D}_l \tilde{U}_l'$ is not of full rank, by adding the diagonal matrix $c_l I_{N_Y}$ the covariance matrix $R_l$ is guaranteed to be non-singular. Using this approach we have reduced the unknown $N_Y \times N_Y$ covariance matrix $R_l$ down to a single unknown scalar $r_l$ while still capturing extra spatial structure through the extra basis vectors.

Furthermore, as we will see shortly, computations never require one to actually store or do calculations with $R_l$, we only need to store $\tilde{U}$ and a few scalars.

We use a similar approach for $S_m$ and let $\tilde{V}_m$ be a matrix that contains columns $K_X + 1$ through $L_Y$ of $V_m$. Now, the $V_m$ are not the singular vectors of $\frac{1}{\sqrt{T_m-1}} X_m$ and so we do not have corresponding singular values. However, we can still obtain the variance explained by each basis vector $v_{m,k}$. The matrix $v_{m,k} v_{m,k}' X_m$ represents fitted vectors resulting from regressing each $X_{m,t}$ onto $v_{m,k}$. Therefore, the squared singular values, $d_{m,k}$, of the matrix $v_{m,k} v_{m,k}' X_m / \sqrt{T_m-1}$ can be viewed as the part of the total sample variance of $X$ explained by the basis vector $v_{m,k}$. With that in
### Table 4.7: The $d_{t,k}$ and $c_l$ constant used for the $R_t$ matrices, see equations (4.16) and (4.17)

<table>
<thead>
<tr>
<th></th>
<th>Summer</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{t,5}$</td>
<td>442</td>
<td>910</td>
<td>969</td>
<td>459</td>
</tr>
<tr>
<td>$d_{t,6}$</td>
<td>339</td>
<td>581</td>
<td>652</td>
<td>366</td>
</tr>
<tr>
<td>$d_{t,7}$</td>
<td>268</td>
<td>454</td>
<td>493</td>
<td>326</td>
</tr>
<tr>
<td>$d_{t,8}$</td>
<td>221</td>
<td>377</td>
<td>363</td>
<td>205</td>
</tr>
<tr>
<td>$d_{t,9}$</td>
<td>192</td>
<td>342</td>
<td>289</td>
<td>191</td>
</tr>
<tr>
<td>$d_{t,10}$</td>
<td>138</td>
<td>238</td>
<td>263</td>
<td>145</td>
</tr>
<tr>
<td>$c_l$</td>
<td>670</td>
<td>1279</td>
<td>1220</td>
<td>778</td>
</tr>
</tbody>
</table>

### Table 4.8: The $d_{m,k}$ and $c_m$ constant used for the $S_m$ matrices, see equations (4.19) and (4.18)

<table>
<thead>
<tr>
<th></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>$d_{m,5}$</td>
<td>208</td>
<td>122</td>
<td>189</td>
<td>492</td>
<td>353</td>
<td>1141</td>
<td>586</td>
<td>853</td>
<td>700</td>
<td>2762</td>
<td>1286</td>
<td>1447</td>
</tr>
<tr>
<td>$d_{m,6}$</td>
<td>202</td>
<td>91</td>
<td>135</td>
<td>237</td>
<td>715</td>
<td>489</td>
<td>624</td>
<td>730</td>
<td>1368</td>
<td>1317</td>
<td>365</td>
<td>362</td>
</tr>
<tr>
<td>$d_{m,7}$</td>
<td>264</td>
<td>96</td>
<td>118</td>
<td>348</td>
<td>315</td>
<td>725</td>
<td>843</td>
<td>1054</td>
<td>872</td>
<td>484</td>
<td>323</td>
<td>207</td>
</tr>
<tr>
<td>$d_{m,8}$</td>
<td>241</td>
<td>81</td>
<td>59</td>
<td>162</td>
<td>164</td>
<td>295</td>
<td>1434</td>
<td>645</td>
<td>770</td>
<td>765</td>
<td>311</td>
<td>332</td>
</tr>
<tr>
<td>$d_{m,9}$</td>
<td>109</td>
<td>66</td>
<td>84</td>
<td>140</td>
<td>197</td>
<td>269</td>
<td>1211</td>
<td>337</td>
<td>948</td>
<td>507</td>
<td>332</td>
<td>761</td>
</tr>
<tr>
<td>$d_{m,10}$</td>
<td>184</td>
<td>60</td>
<td>81</td>
<td>120</td>
<td>150</td>
<td>447</td>
<td>876</td>
<td>552</td>
<td>542</td>
<td>566</td>
<td>266</td>
<td>397</td>
</tr>
<tr>
<td>$c_m$</td>
<td>808</td>
<td>353</td>
<td>551</td>
<td>867</td>
<td>1468</td>
<td>2790</td>
<td>2810</td>
<td>3157</td>
<td>3740</td>
<td>3553</td>
<td>1934</td>
<td>1546</td>
</tr>
</tbody>
</table>
mind we set $\tilde{D}_m$ to be a diagonal matrix with elements $d_{m,K_{LX}+1}, \ldots, d_{m,LX}$ and we set $c_m$ to be the total variance left in $X$ after accounting for the first $LX$ basis vectors, i.e.

$$c_m = \sum_{k=LX+1}^{NX} d_{m,k} \quad \text{for } m = 1, \ldots, 12. \tag{4.18}$$

Then we set

$$S_m = s_m \left( c_m I_{NX} + \sum_{j=K_{LX}+1}^{LX} d_{m,j} v_{m,j} v_{m,j}' \right)$$

$$= s_m \left( c_m I_{NX} + \tilde{V}_m \tilde{D}_m \tilde{V}_m' \right) \equiv s_m \tilde{S}_m. \tag{4.19}$$

The values of $d_{m,k}$ and $c_m$ are listed in Table 4.8. This approach reduces the $NX \times NX$ dimensional matrix $S_m$ down to one scalar, $s_m$. Note that though $LX$ and $LY$ are the same for all seasons, we can easily relax that assumption.

The unknown scalars $r_l$ and $s_m$ are assumed to be independent and are assigned conjugate inverse gamma priors:

$$[r_l] = IG(\alpha_1, \beta_1) \text{ for } l = s, f, w, p \quad \text{and}$$

$$[s_m] = IG(\alpha_2, \beta_2) \text{ for } m = 1, \ldots, 12. \tag{4.20}$$

The selection of the hyperparameters $\alpha_1, \beta_1, \alpha_2$ and $\beta_2$ is discussed in Section 4.5.3.1.

Turning to the computational advantages of this approach, we start by examining the inverses of $R_l$ and $S_m$. Using the matrix identity

$$(\Sigma + A' \Gamma A)^{-1} = \Sigma^{-1} - \Sigma^{-1} A'(A \Sigma^{-1} A' + \Gamma^{-1})^{-1} A \Sigma^{-1} \tag{4.21}$$
and noting that \( \tilde{U}' \tilde{U} \) and \( \tilde{V}' \tilde{V} \) are identity matrices (since the basis vectors are orthonormal) we obtain

\[
R_l^{-1} = \frac{1}{r_l} \left( \frac{1}{c_l} I_{N_Y} - \frac{1}{c_l^2} \tilde{U} \left( \frac{1}{c_l} \tilde{U}' \tilde{U} + \tilde{D}^{-1}_l \right)^{-1} \tilde{U} \right) = \frac{1}{r_l} \left( \frac{1}{c_l} I_{N_Y} - \tilde{U} \Lambda_l \tilde{U}' \right) \equiv \frac{1}{r_l} E_l \text{ and }
\]

\[
S_m^{-1} = \frac{1}{s_m} \left( \frac{1}{c_m} I_{N_X} - \frac{1}{c_m^2} \tilde{V} \left( \frac{1}{c_m} \tilde{V}' \tilde{V} + \tilde{D}^{-1}_m \right)^{-1} \tilde{V}' \right) = \frac{1}{s_m} \left( \frac{1}{c_m} I_{N_X} - \tilde{V} \Lambda_m \tilde{V}' \right) \equiv \frac{1}{s_m} E_m
\]

(4.22)

where \( \Lambda_l \) and \( \Lambda_m \) are diagonal matrices with diagonal elements

\[
\frac{1}{c_l^2} \frac{1}{c_l} + \frac{1}{d_{k,l}} = \frac{d_{k,l}}{c_l(c_l + d_{k,l})} \text{ for } k = K_Y + 1, \ldots, L_Y \quad \text{and} \quad \frac{1}{c_m^2} \frac{1}{c_m} + \frac{1}{d_{m,k}} = \frac{d_{m,k}}{c_m(c_m + d_{m,k})} \text{ for } k = K_X + 1, \ldots, L_X .
\]

(4.24)

Furthermore, we note that \( R_l \) and \( S_m \) appear in the Gibbs sampler (see Appendix C) as \( U'_l R_l^{-1} Y_{l,t} \), \( U'_l R_l^{-1} U_l \), \( V'_m S_m^{-1} X_{m,t} \) and \( V'_m S_m^{-1} V_m \) which can be written as

\[
U'_l R_l^{-1} Y_{l,t} = U'_l \frac{1}{r_l} \left( \frac{1}{c_l} I_{N_Y} - \tilde{U} \Lambda_l \tilde{U}' \right) Y_{l,t} = \frac{1}{r_l} \left( \frac{1}{c_l} U'_l Y_{l,t} - U'_l \tilde{U} \Lambda_l \tilde{U}' Y_{l,t} \right)
\]

\[
= \frac{1}{r_l c_l} U'_l Y_{l,t},
\]

(4.25)

\[
U'_l R_l^{-1} U_l = U'_l \frac{1}{r_l} \left( \frac{1}{c_l} I_{N_Y} - \tilde{U} \Lambda_l \tilde{U}' \right) U_l = \frac{1}{r_l} \left( \frac{1}{c_l} U'_l U_l - U'_l \tilde{U} \Lambda_l \tilde{U}' U_l \right)
\]

\[
= \frac{1}{r_l c_l} I_{K_Y},
\]

(4.26)

and

\[
V'_m S_m^{-1} X_{m,t} = \frac{1}{s_m c_m} V'_m X_{m,t} \quad \text{and} \quad V'_m S_m^{-1} V_m = \frac{1}{s_m c_m} I_{K_X} .
\]

(4.27)

In the calculations above we used the fact that \( \tilde{U}' \tilde{U} \) and \( \tilde{V}' \tilde{V} \) are matrices with all elements equal to zero due to the orthogonality of the basis vectors.
To summarize, the parameter model is the joint distribution of all the unknowns discussed in (I)-(III). We assume apriori independence between most parameters as is indicated in the following:

\[
[h_1, \text{vec}(B_1), D_1, s_1, \ldots, h_{12}, \text{vec}(B_{12}), D_{12}, s_{12}, C_s, C_f, C_w, C_p, r_s, r_f, r_w, r_p]
= \prod_{m=1}^{12} [h_m] \ [\text{vec}(B_m)] \ [D_m] \ [s_m] \ \prod_{l=s,f,w,p} [C_l] \ [r_l].
\] (4.28)

The individual distributions on the left hand side in (4.28) are defined in (4.14), (4.15) and (4.20)

4.5.3.1 Hyperparameters

The prior distributions in the parameter model have hyperparameters that need to be selected. The hyperparameters are the following:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I) Transition matrices</td>
<td>$H_m$ and $B_m$</td>
</tr>
<tr>
<td>(II) Initial amplitudes</td>
<td>$b_{2,1}$</td>
</tr>
<tr>
<td>(III) Covariance matrices</td>
<td>$C_l$ and $D_m$</td>
</tr>
<tr>
<td>(IV) Covariance matrices</td>
<td>$R_l$ and $S_l$</td>
</tr>
</tbody>
</table>

(I) Transition matrices ($\mu_{1,m}, \mu_{2,m}, \Sigma_{1,m}, \Sigma_{2,m}$): The hyperparameters $\mu_{1,m}, \mu_{2,m}, \Sigma_{1,m}$ and $\Sigma_{2,m}$ are the means and covariances of the diagonal of the $H_m$ matrices, $h_m$, and the vectorized transition matrices, $\text{vec}(B_m)$. Note that the process model for the $k^{th}$ Polar MM5 amplitudes from one season, say $l = f$, can be written as:

\[
a_{f,k,t} = H_{3,k}b_{3,k,t} + H_{4,k}b_{4,k,t} + H_{5,k}b_{5,k,t} + \epsilon_t
\] (4.29)

where $H_{m,k}$ is the $k$th diagonal element of the $H_m$ matrix. There is no prior information available on the $H_{m,k}$ parameters and without seeing any examples of the $a_{f,k,t}$ and $b_{m,k,t}$ amplitudes, it is difficult to even postulate about the scale of $H_{m,k}$. We therefore allow ourselves a peak at the data to guide the selection of $\mu_{1,m}$ and
Viewing $H_{m,k}$ as unknown parameters and using estimated amplitudes $\hat{a}_{l,k,t}$ and $\hat{b}_{m,k,t}$ (see equation (4.3)) as observations, we obtain estimates, $\hat{H}_{m,k}$, of $H_{m,k}$ and the residual standard error, $\hat{\sigma}_{m,k}$, via least squares estimation. This was done in section 4.4; see Table 4.2. We see that the $\hat{H}_{m,k}$ in Table 4.2 are almost all positive and the largest ones are about 1.5. With that in mind we set the prior mean equal to 0.5 for all $m$ and $k$ and the prior standard deviation equal to 4, i.e.

$$\mu_{1,m} = 0.5\mathbf{1}_{KY} \quad \text{and} \quad \Sigma_{1,m} = 4^2 I_{KY}, \quad \forall m . \tag{4.30}$$

Note that these hyperparameters put approximately 95% of the prior density mass of $H_{m,k}$ on the interval $(-7.5, 8.5)$ which is considerably wider than the range of $\hat{H}_{m,k}$ values in Table 4.2.

Similarly, the process model for the $k^{th}$ ERA amplitude for month $m$ can be written as

$$b_{m,k,t} = B_{m,k,1}b_{m-1,1,t} + \cdots + B_{m,k,K_X}b_{m-1,K_X,t} + \epsilon_t, \tag{4.31}$$

where $B_{m,i,j}$ is the $(i,j)^{th}$ element of the matrix $B_m$. As before, coming up with a sensible prior on $B_{m,i,j}$ is difficult, so we again take a peak at the data. We obtained estimates, $\hat{B}_{m,i,j}$, of each element $B_{m,i,j}$ and the corresponding standard error $\hat{\sigma}_{m,i,j}$ using the estimated amplitudes $\hat{b}_{m,k,t}$ and $\hat{b}_{m-1,k,t}$, see Table 4.4. We see that the $B_{m,i,j}$ values in Table 4.4 are all within -1 and 2. With that in mind we set the prior mean equal to 0 for all $m$, $i$ and $j$ and the prior standard deviation equal to 4, i.e.

$$\mu_{2,m} = \mathbf{0}_{K_X} \quad \text{and} \quad \Sigma_{2,m} = 4^2 I_{K_X} , \quad \forall m . \tag{4.32}$$

(II) Initial amplitudes $(\mu_b, \Sigma_b)$: The hyperparameters $\mu_b$ and $\Sigma_b$ are the mean and covariance of the initial ERA amplitude vector, $b_{2,1}$, for February 1979. In our
case we actually have ERA data from this period, $X_{2,1}$ (see Figures B.2 and B.8 in Appendices B.1 and B.2), so we set the prior mean equal to the estimated amplitudes,

$$ \mu_b = V_2 X_{2,1} = \hat{b}_{2,1} = (-26.784, 12.094, -5.186, 5.562)^\prime. $$

(4.33)

In comparison, February amplitudes for other years range from about -150 to 40 (see Figure B.15 in Appendix B.3.1). The estimated standard error of the above estimates of $b_{2,1}$ is 0.72 (same for all $k$). In order to avoid being too concentrated in our prior selection, we set the prior standard deviation to 10, i.e.

$$ \Sigma_b = 10^2 I_{K_X}. $$

(4.34)

(III) Covariance matrices of the process model ($W_l, W_m, \nu_l, \nu_m$): The expected value of an $IW_\nu(W)$ distributed, $k \times k$ matrix is $\frac{1}{\nu-k-1} W$. Furthermore, the larger $\nu$ is the tighter the distribution is around its mean. To select a vague prior, we select $\nu_m = K_X + 2$ and $\nu_l = K_Y + 2$ which means that the mean of the inverse-Wishart distributions will be $W_l$ and $W_m$. Looking back at Table 4.2 we see that the residual errors are no larger than 60 and with that in mind we select $W_l = 60^2 I_{K_Y}$ for all seasons $l$. Similarly, looking at Table 4.4 we see that the residual errors are no larger than 70 and so we select $W_m = 70^2 I_{K_Y}$ for all seasons $m$.

(IV) Covariance matrices of the data model ($\alpha_1, \beta_1, \alpha_2, \beta_2$): The hyperparameters $\alpha_1$, $\beta_1$, $\alpha_2$ and $\beta_2$ define the prior distribution of $r_l$ and $s_m$ which are the only unknown parameters of the data-model covariance matrices $R_l$ and $S_m$. We set proper, but vague priors on $r_l$ and $s_m$. In particular we select $\alpha_1 = \alpha_2 = 2$ and $\beta_1 = \beta_2 = 1$ which means that the prior mean is 1 but the prior distribution has an infinite variance. The reason for picking the mean as 1 is that we have included the
total variance of the data within the data model (through the \( c_l \) and \( c_m \) in \( R_l \) and \( S_m \)). The density function for this prior distribution is shown in Figure 4.17 (left). We note that \( r_l \) and \( s_m \) can be interpreted as scaling parameters of the total variance of the data and so the square root of \( r_l \) and \( s_m \) are easier to interpret. The implied densities (see Appendix A.1.4) for \( \sqrt{r_l} \) and \( \sqrt{s_m} \) are shown in Figure 4.17 (right).

We note that in (I) - (III) we used the data to guide our selection of hyperparameters. The situation where hyperparameters are estimated from data is referred to as \textit{empirical} Bayes (Berger (1985); Section 4.5) and is not strictly a fully Bayesian approach. Applying Bayes theorem with hyperparameters that are estimated from data ignores errors introduced in the estimation. These errors will therefore not be reflected in the posterior distribution. To partially mitigate this problem we selected prior densities wider than what estimated hyperparameters would have suggested. Ideally we should perform sensitivity analysis on the selected hyperparameters, i.e. examine how sensitive our conclusions are to the values of the hyperparameters.
4.6 Predicting Polar MM5 temperatures with and without new ERA-40 data

In this Section we outline ways to obtain predictions of Polar MM5 temperatures. The data included in the Bayesian model in Section 4.5 extended through the fall season of year 23\textsuperscript{rd} year (calendar year 2001). The data from winter 2001 through fall 2002 was set aside. We can therefore compare our predictions to this last year of data. First, we consider the situation where we have no new data and we wish to predict the next four seasons of both ERA-40 and Polar MM5 temperatures, see Section 4.6.1. We refer to this situation as forecasting. Next we consider the situation where we have obtained ERA-40 data for the next season and wish to use that to predict the Polar MM5 temperatures for that same season, see Section 4.6.2. We refer to this latter situation as downscaling.

4.6.1 Forecasting without new ERA-40 data

We start by looking at how we forecast both Polar MM5 and ERA-40 temperatures for the next season, winter of the 23\textsuperscript{rd} year (calendar year 2001), without having access to any new ERA-40 data. Hence we use the posterior predictive distribution of $Y_{w,23}$ and $X_{6,23}, X_{7,23}, X_{8,23}$ given all the previous data $Y_{1:T}$ and $X_{1:T}$. Recall that the index notation “1 : T” refers only to data through fall of the 23\textsuperscript{rd} year, see (4.5) and (4.6). We assume the same data model for future data as we did in (4.7), that is, we assume that the data vectors $Y_{l,t}$ and $X_{m,t}$ are independent given the corresponding amplitude vectors and the parameters $n_l$ and $s_m$. The posterior predictive distribution
can therefore be written as

\[
\begin{align*}
[Y_{w,23}, X_{6,23}, X_{7,23}, X_{8,23}|Y_{1:T}, X_{1:T}] &= \int \left[ Y_{w,23}, X_{6,23}, X_{7,23}, X_{8,23}, a_{w,23}, b_{6,23}, b_{7,23}, b_{8,23}, r_w, s_6, s_7, s_8 | Y_{1:T}, X_{1:T} \right] \\
& \quad \quad da_{w,23} \, db_{6,23} \, db_{7,23} \, db_{8,23} \, dr_w \, ds_6 \, ds_7 \, ds_8 \\
&= \int \left[ Y_{w,23} | a_{w,23}, r_w \right] \left[ X_{6,23} | b_{6,23}, s_6 \right] \left[ X_{7,23} | b_{7,23}, s_7 \right] \left[ X_{8,23} | b_{8,23}, s_8 \right] \\
& \quad \quad \times \left[ a_{w,23}, b_{6,23}, b_{7,23}, b_{8,23}, r_w, s_6, s_7, s_8 | Y_{1:T}, X_{1:T} \right] \\
& \quad \quad \quad \quad da_{w,23} \, db_{6,23} \, db_{7,23} \, db_{8,23} \, dr_w \, ds_6 \, ds_7 \, ds_8 . \quad (4.35)
\end{align*}
\]

The distribution in (4.35) is not tractable analytically but we can obtain samples from it using the following well-known Lemma:

**Lemma 4.6.1.** Let \( \mathbf{Y} \) and \( \mathbf{X} \) be random vectors and let

\[
[\mathbf{Y}] = \int [\mathbf{Y} | \mathbf{X}] [\mathbf{X}] d\mathbf{X} \quad (4.36)
\]

where the distributions \([\mathbf{Y} | \mathbf{X}]\) and \([\mathbf{X}]\) are known. Then we can obtain samples \( \mathbf{y}^{(i)} \), \( i = 1, \ldots, M \), from the marginal distribution \([\mathbf{Y}]\) in two steps:

(i) Sample \( \mathbf{x}^{(i)} \) from \([\mathbf{X}]\), \( i = 1, \ldots, M \).

(ii) Sample \( \mathbf{y}^{(i)} \) from \([\mathbf{Y} | \mathbf{x}^{(i)}]\), \( i = 1, \ldots, M \).

Let \( a_{w,23}^{(i)}, b_{6,23}^{(i)}, b_{7,23}^{(i)}, b_{8,23}^{(i)}, r_w^{(i)}, s_6^{(i)}, s_7^{(i)}, s_8^{(i)} \), \( i = 1, \ldots, M \) be samples from the distribution

\[
[a_{w,23}, b_{6,23}, b_{7,23}, b_{8,23}, r_w, s_6, s_7, s_8 | Y_{1:T}, X_{1:T}] . \quad (4.37)
\]

How to obtain these samples is discussed below, see equations (4.40) and (4.41). Then, by Lemma 4.6.1, we can obtain samples from the posterior predictive distribution in
If we are only interested in prediction of Polar MM5 temperatures, we only need the \( y^{(i)}_{w,23} \) samples. These samples can be viewed as an ensemble of predictions of \( Y_{w,23} \).

We can obtain an average predicted vector and uncertainty bounds in the same way as we use MCMC samples for inference about parameters in a Bayesian model. However, sampling the distributions in (4.38) is computationally very expensive due to the high dimensions of \( \tilde{R}_w \) and \( \tilde{S}_m \). When obtaining the samples in (4.38) is too difficult, we consider basing our predictions of \( Y_{w,23} \) and \( X_{6,23}, X_{7,23}, X_{8,23} \) on the ensembles

\[
V_6 b^{(i)}_{6,23}, V_7 b^{(i)}_{7,23}, V_8 b^{(i)}_{8,23}, U_w a^{(i)}_{w,23}, \quad i = 1, \ldots, M
\]

(4.39) instead. We note that \( U_w a^{(i)}_{w,23} \) is a sample of the mean, \( E(Y_{w,23} | a_{w,23}, Y_{1:T}, X_{1:T}) \), and does not account for the error structure in \( r^{(i)}_w \tilde{R}_w \). Predictions based on (4.39) can therefore be expected to be too smooth relative to the actual processes.

Now, we focus on the posterior predictive distribution of the amplitudes in (4.37). First note that \( a_{w,23} \) is independent of past data, \( Y_{1:T} \) and \( X_{1:T} \); given the amplitude vectors, \( b_{6,23}, b_{7,23} \) and \( b_{8,23} \), and other parameters. Similarly, the \( b_{m,t} \) amplitude vectors are independent of past data given the amplitude vector a month before. This is a consequence of the conditional independence assumptions in (4.7) and (4.9), see also Lemma A.1.6. Let \( \theta = \{C_w, H_w, D_6, D_7, D_8, B_6, B_7, B_7\} \). The distribution in

(4.35) by sampling for \( i = 1, \ldots, M \),

\[
\begin{align*}
\mathbf{x}^{(i)}_{6,23} & \text{ from } N(V_6 b^{(i)}_{6,23}, s^{(i)}_6 \tilde{S}_6), \\
\mathbf{x}^{(i)}_{7,23} & \text{ from } N(V_7 b^{(i)}_{7,23}, s^{(i)}_7 \tilde{S}_7), \\
\mathbf{x}^{(i)}_{8,23} & \text{ from } N(V_8 b^{(i)}_{8,23}, s^{(i)}_8 \tilde{S}_8) \quad \text{ and } \\
y^{(i)}_{w,23} & \text{ from } N(U_w a^{(i)}_{w,23}, r^{(i)}_w \tilde{R}_w) .
\end{align*}
\]
(4.37) can be written as

\[
\begin{align*}
[a_{w,23}, b_{6,23}, b_{7,23}, b_{8,23}, r_w, s_6, s_7, s_8 | Y_{1:T}, X_{1:T}] \\
= \int \left[ a_{w,23}, b_{6,23}, b_{7,23}, b_{8,23}, r_w, s_6, s_7, s_8, b_{5,23}, \theta | Y_{1:T}, X_{1:T} \right] db_{5,23} d\theta \\
= \int \left[ a_{w,23} | b_{6,23}, b_{7,23}, b_{8,23}, H_w, C_w \right] \left[ b_{8,23} | b_{7,23}, B_8, D_8 \right] \left[ b_{7,23} | b_{6,23}, B_7, D_7 \right] \\
\times \left[ b_{6,23} | b_{5,23}, B_6, D_6 \right] \left[ b_{5,23}, \theta, r_w, s_6, s_7, s_8 | Y_{1:T}, X_{1:T} \right] db_{5,23} d\theta \quad (4.40)
\end{align*}
\]

After we have fitted the Bayesian model in Sections 4.5 using a Markov Chain Monte Carlo (MCMC) method, we have samples \( b_{i,23} \) and \( \theta^{(i)} \), \( i = 1, \ldots, M \), from the last distribution in (4.40). Then, by Lemma 4.6.1, we can obtain samples from (4.40) by sampling in the following order:

\[
\begin{align*}
b_{6,23}^{(i)} & \text{ from } N(B_6^{(i)} b_{5,23}^{(i)}, D_6^{(i)}), \\
b_{7,23}^{(i)} & \text{ from } N(B_7^{(i)} b_{6,23}^{(i)}, D_7^{(i)}), \\
b_{8,23}^{(i)} & \text{ from } N(B_8^{(i)} b_{7,23}^{(i)}, D_8^{(i)}) \quad \text{and} \\
a_{w,23}^{(i)} & \text{ from } N(H_6^{(i)} b_{6,23}^{(i)} + H_7^{(i)} b_{7,23}^{(i)} + H_8^{(i)} b_{8,23}^{(i)}, C_w^{(i)}) \quad \text{for } i = 1, \ldots, M. \quad (4.41)
\end{align*}
\]

If we wish to forecast further out in time we can obtain samples from the posterior predictive distribution of the corresponding amplitudes by continuing the sequential sampling scheme above. That is, we continue to sample \( b_{m,t}^{(i)} \) from \( N(B_m^{(i)} b_{m-1,t}^{(i)}, D_m^{(i)}) \) and \( a_{l,t}^{(i)} \) from \( N(H_{m_1}^{(i)} b_{m_1,t}^{(i)} + H_{m_2}^{(i)} b_{m_2,t}^{(i)} + H_{m_3}^{(i)} b_{m_3,t}^{(i)}, C_l^{(i)}) \) for months \( m \), seasons \( l \) and years \( t \) as far as we wish to forecast.

### 4.6.2 Downscaling ERA-40 data

Here we consider the case where ERA-40 data for the three winter months 2001, \( X_{6,23}, X_{7,23} \) and \( X_{8,23} \), are available and we wish to predict Polar MM5 temperatures
for winter 2001, $Y_{w,23}$. This setting is the essence of statistical downscaling, that is to predict the high-resolution Polar MM5 temperatures using coarse-resolution ERA-40 data. Here we use the posterior predictive distribution of $Y_{w,23}$ given all observed data. Using the assumption that given the amplitude vector $a_{w,23}$ and $r_w$, $Y_{w,23}$ is independent of other data we can write this distribution as

$$
[Y_{w,23}|Y_{1:T}, X_{1:T}, X_{6,23}, X_{7,23}, X_{8,23}]
= \int [Y_{w,23}, a_{w,23}, r_w|Y_{1:T}, X_{1:T}, X_{6,23}, X_{7,23}, X_{8,23}] da_{w,23} dr_w.
$$

As before, if we can obtain samples from the second distribution of the integral, $a_{w,23}^{(i)}$ and $r_{w}^{(i)}$, we can obtain samples from the predictive distribution of $Y_{w,23}$ by sampling $N(U_w a_{w,23}^{(i)}, r_w^{(i)} R_w)$. The remarks made after equation (4.38) also apply here so we consider the use of ensembles $\{U_w a_{w,23}^{(i)}\}$ instead.

We obtain samples from the second distribution of the integral in (4.42) by extending the model in Section 4.5 to include the new ERA-40 data. We add the distributions

$$
[X_{m,23}|b_{m,23}, S_m] = N(V_m b_{m,t}, S_m) \text{ for } m = 6, 7, 8,
$$

(4.43)

to the data model, assuming conditional independence as in (4.7). Note that $S_m$ is not a new parameter and that we use the same basis vectors $V_m$. We also add the new amplitude vectors introduced in (4.43) to the process model in Section 4.5, i.e.

$$
[b_{m,23}|b_{m-1,23}, B_m, D_m] = N(B_m b_{m-1,t}, D_m) \text{ for } m = 6, 7, 8.
$$

(4.44)
Furthermore, we include the $a_{m,23}$ amplitude vector in the process model, with

$$[a_{m,23} | b_{6,23}, b_{7,23}, b_{8,23}, H_6, H_7, H_8, C_w]$$

$$= N(H_6 b_{6,23} + H_7 b_{7,23} + H_8 b_{8,23}, C_w). \quad (4.45)$$

We can fit the model in Section 4.5 with the additions in (4.43) - (4.45) using MCMC to obtain samples from the posterior predictive distribution of winter 2001 Polar MM5 amplitude vectors, i.e. $a_{i,w,23}$, $i = 1, \ldots, M$.

For our example, running the MCMC again with the new data is not a difficult task and takes only a few hours. However, we might be in the situation where running the MCMC again is not an option, due to computational constraints, and in that case we could consider the method of sequential importance sampling or particle filters (see for example Robert and Casella, 2004; Doucet et al., 2001; Berliner and Wikle, 2007). In that approach we would use the MCMC samples obtained for the parameters in the model in Section 4.5 and obtain “updated” posterior predictive samples of $a_{w,23}$ that take the new observed data $X_{6,23}$, $X_{7,23}$ and $X_{8,23}$ into account. We will not pursue that approach here.

### 4.7 Results

We obtained samples from the posterior distribution of the model described in Section 4.5 via Gibbs sampling. The full conditional distributions of the Gibbs sampler are listed in Appendix C. There are 1,844 unknown parameters in the model, see Table C.1. We obtained 30,000 MCMC samples and discarded the first 1,000 as burn-in, leaving 29,000 MCMC samples to use for inference. We examined trace-plots of all parameters (not shown here) and they showed no indication of convergence problems. Inferences of unknown parameters are summarized in Figures 4.18 - 4.28. Estimated
marginal posterior densities were obtained from the MCMC samples using (Gaussian) kernel density estimation. Posterior means, standard deviations and posterior credible intervals were estimated by calculating the sample mean, sample standard deviation and sample percentiles of the MCMC samples.

First we examine the marginal posterior distributions of amplitudes, $a_{l,t,k}$ and $b_{m,t,k}$. Figures 4.18 and 4.19 show estimated posterior means and 90% credible intervals of amplitudes. The credible intervals are very narrow and hardly visible in Figure 4.18. While the posterior means of $a_{l,t,k}$ vary from $-171$ to $237.8$, the widths of the 90% credible intervals range from only $1.13$ to $1.65$. The posterior means of $b_{m,t,k}$ vary from $-236.3$ to $176.6$. However, the 90% credible intervals (excluding the first amplitude vector $b_{2,1}$) are wider than for the $a_{l,t,k}$, they range from $1.72$ to $5.71$. This difference in the widths of posterior credible intervals goes together with the differences in size of data vectors. We have larger Polar MM5 data vectors ($N_Y = 14,641$) at each time point than ERA-40 data vectors ($N_X = 2,736$). We also note that the elements of the first ERA-40 amplitude vector, $b_{2,1}$, has the widest posterior credible interval, which is not surprising since there are no data at that time point. In Figures 4.20 and 4.21 we have plotted the time series of the posterior means of all amplitudes. We see that the first amplitudes, i.e. the amplitudes of the first EOF and first OMCP ($a_{l,t,1}$ and $b_{m,t,1}$, respectively) tend to be the most dominant amplitudes; see the black lines in Figures 4.20 and 4.21. These first amplitudes are usually the largest in absolute value. The fourth amplitudes, $a_{l,t,4}$ and $b_{m,t,4}$, on the other hand tend to be the smallest amplitudes; see the blue lines in Figures 4.20 and 4.21.

The parameters $r_l$, $l = s, f, w, p$, and $s_m$, $m = 1, \ldots, 12$, are not easy to interpret. They are the only unknown parts of the covariance matrices, $R_l = r_l \tilde{R}_l$ and $S_m =$
Figure 4.18: Estimated posterior means (‘-‘) and 90% credible intervals (vertical lines) for the Polar MM5 amplitudes, $a_{t,1}$, for all 89 seasons. The top panel shows the first amplitude for each season, the bottom panel shows the fourth amplitude.
Figure 4.19: Estimated posterior means ("-") and 90% credible intervals (vertical lines) for the Polar MM5 amplitudes, $\hat{b}_{m,t}$, for all 268 months. The top panel shows the first amplitude for each season, the bottom panel shows the fourth amplitude.
Figure 4.20: Estimated posterior means of the Polar MM5 amplitudes, $a_{t,t}$, for all 89 seasons.

Figure 4.21: Estimated posterior means of the ERA-40 amplitudes, $b_{m,t}$, for all 268 months.
Posterior densities of $\sqrt{r_l}$ and $\sqrt{s_m}$, of the data model, see equations (4.16) and (4.19). The matrices $\tilde{R}_l$ and $\tilde{S}_m$ have a rich structure and the parameters $r_l$ and $s_m$ can be thought of as bringing the variances on the diagonal of $\tilde{R}_l$ and $\tilde{S}_m$ to the right scale. We can also think of the $\sqrt{r_l}$ and $\sqrt{s_m}$ as multiplicative scalars of the standard deviations. Our prior on these parameters reflected our prior belief that no correction of scale was needed and that $r_l$ and $s_m$ were close to 1 (see Figure 4.17). Figure 4.22 shows estimated posterior densities of $\sqrt{r_l}$ and $\sqrt{s_m}$. We see that the estimated posterior densities of these parameters are concentrated around values much smaller than 1.

Next we consider posterior inference of the transition matrices $H_m$ and $B_m$, $m = 1, \ldots, 12$. In Figure 4.23 we show estimated posterior densities of the diagonal elements of the $H_m$ matrices (the off-diagonal elements were assumed to be zero). We note that the posterior densities are all more concentrated than the prior distribution (dashed lines) which is very flat compared to the posterior densities. This suggests that the posterior distribution is driven more by data than by the prior distribution. Furthermore, note that the elements of the matrix of the last month
within season (third column in Figure 4.23) tend to be positive, with most of the posterior density mass above zero. For the other two months within each season (first and second column in Figure 4.23), many amplitudes tend to have posterior densities concentrated around zero, especially September and October amplitudes. This indicates that given the November amplitudes, \( b_{11,t} \), the September and October amplitudes, \( b_{9,t} \) and \( b_{10,t} \), do not contribute much in the modeling of the spring Polar MM5 amplitudes \( a_{p,t} \). The same can be said about many of the other amplitudes of the first two months within each season. This is, in fact, in concert with our findings in the exploratory analysis in Section 4.4; that is, the models that included only the \( b \) amplitudes of the last month of each season, often resulted in a \( R^2 \) value close to the \( R^2 \) value of models that included all three amplitudes (see Table 4.2).

In Figure 4.24 we show estimated posterior densities of all 16 elements of the \( B_m \) matrices. In general, the lower triangular elements, i.e. the elements below the diagonal of \( B_m \) (blue lines) tend to be concentrated around zero. On the other hand, most of the posterior densities of the diagonal elements (red lines) and the elements of the upper triangular of \( B_m \) are concentrated around non-zero values. This indicates that the \( B_m \) matrices could be taken to be upper triangular matrices. That means that amplitudes from one month before, \( b_{m-1,t,k_1} \), are only helpful in predicting next months amplitudes of higher order, i.e. \( b_{m,t,k_2} \) where \( k_2 > k_1 \). That is, amplitudes for basis vectors of larger scales are not helpful in predicting amplitudes for basis vectors of smaller scales.

Lastly, we consider inference of the covariance matrices of the process model, \( C_l \) for \( l = s, f, w, p \) and \( D_m \) for \( m = 1, \ldots, 12 \). Figures 4.25 and 4.27 show estimated posterior densities of the standard deviations, i.e. the square root of the diagonal
Figure 4.23: Estimated posterior densities of the diagonal elements of $H_m$ for each month.
Figure 4.24: Estimated posterior densities of the 16 elements of the transition matrix $B_m$ for each month.
elements of $C_l$ and $D_m$, respectively. For both $C_l$ and $D_m$, the standard deviations of amplitudes of higher order ($k = 1$ or 2) tend to be larger (red and black lines) than the standard deviations of amplitudes of lower order ($k = 3$ or 4, green and blue lines). In most cases the posterior densities are concentrated around values lower than the prior mean of the standard deviations, 60 for the standard deviations of $C_l$ and 70 for the standard deviations of $D_m$.

Figures 4.25 and 4.27 show estimated posterior means and 90% posterior credible intervals for the correlations in $C_l$ and $D_m$. MCMC samples of these correlations are obtained from the MCMC samples $C_{l,i,j,k}^{(i)}$ and $D_{m,i,j,k}^{(i)}$, $i = 1, \ldots, M$, by calculating

$$
\rho_{l,j,k}^{(i)} = \frac{C_{l,j,k}^{(i)}}{\sqrt{C_{l,j,j}^{(i)} C_{l,k,k}^{(i)}}} \quad \text{and} \quad \rho_{m,j,k}^{(i)} = \frac{D_{m,j,k}^{(i)}}{\sqrt{D_{m,j,j}^{(i)} D_{m,k,k}^{(i)}}},
$$

where $C_{l,j,k}^{(i)}$ and $D_{m,j,k}^{(i)}$ are the $(j,k)^{th}$ elements of $C_{l,i,j,k}^{(i)}$ and $D_{m,i,j,k}^{(i)}$, respectively. Only one credible interval in Figure 4.25 does not cover zero and most posterior means are close to zero, indicating that $C_l$ could be taken as a diagonal matrix. On the other hand, in Figure 4.27 some credible intervals are far from zero, indicating high (positive) correlations between some $b$ amplitudes.

### 4.7.1 Forecasting the next year, without new ERA-40 data

We now consider predicting both the Polar MM5 and ERA-40 data for the following year, winter of 2001 through fall of 2002. The actual Polar MM5 and ERA-40 temperature anomalies, $Y_{l,t}$ and $X_{m,t}$, are shown in Figure 4.29. Here we assume that we have no new data and we proceed with the methods introduced in Section 4.6.1. From the previously described Gibbs sampling, we have 29,000 MCMC posterior samples of

$$
b_{i,23}^{(i)}, D_{m}^{(i)}, B_{m}^{(i)}, H_{m}^{(i)}, C_{l}^{(i)} \quad \text{for } m = 1, \ldots, 12 \text{ and } l = s, f, w, p.
$$
Figure 4.25: Estimated posterior densities of the standard deviations of the $C_l$ matrices, i.e. the square root of the diagonal elements of $C_l$.

Following the sampling scheme of (4.41), we obtain predictive samples $b^{(i)}_{m,t}$ and $a^{(i)}_{l,t}$ for $m$, $l$ and $t$ that correspond to the time period winter of 2001 through fall of 2002.

Figure 4.30 shows the predicted means and prediction intervals of each element of the amplitude vectors, calculated from these samples. The predicted means of the $b_{m,t,k}$ amplitudes are all close to zero, except the first amplitudes (black lines) for the first five months. The predicted means of the $a_{l,t,k}$ amplitudes are also all close to zero, except the first amplitude for the first season. Furthermore, the 90% prediction intervals for both $b_{m,t,k}$ and $a_{l,t,k}$ are very wide, they cover almost the same range as the estimated amplitudes shown in Figures 4.18 and 4.19. Using the data for the last year we can obtain estimated amplitude vectors

$$\hat{a}_{l,t} = U'_l Y_{l,t} \quad \text{and} \quad \hat{b}_{m,t} = V'_m X_{m,t},$$

(4.48)
Figure 4.26: Estimated 90% posterior credible intervals and posterior means (circles) for the correlations in $C_l$. The labels, $j,k$, indicate the $(j,k)^{th}$ element of $C_l$. 
Figure 4.27: Estimated posterior densities of the standard deviations of the $D_m$ matrices, i.e. the square root of the diagonal elements of $D_m$. 
Figure 4.28: Estimated 90% posterior credible intervals and posterior means (circles) for the correlations in $D_m$. The labels, $j.k$, indicate the $(j,k)^{th}$ element of $D_m$. 
for \( l, m \) and \( t \) that correspond to the last year. The estimated amplitudes (x-marks in Figure 4.30) are very different from the predicted amplitudes (circles in Figure 4.30). However, the estimated amplitudes are all within the 90% prediction intervals, indicating that the uncertainty of the predictions is indeed realistic.

Although the predicted amplitudes do not bode well for predicting future data, we proceed to inference of the ensembles \( V_m b_{m,t}^{(i)} \) and \( U_l a_{l,t}^{(i)} \). We note that due to their size, we do not wish to store all \( M = 29,000 \) ensembles of the \( N_X \) and \( N_Y \)-dimensional vectors \( V_m b_{m,t}^{(i)} \) and \( U_l a_{l,t}^{(i)} \). The posterior predicted mean of \( U_l a_{l,t} \) and \( V_m b_{m,t} \) can be estimated as

\[
\hat{E}(U_l a_{l,t} | Y_{1:T}, X_{1:T}) = \frac{1}{M} \sum_{i=1}^{M} U_l a_{l,t}^{(i)} = U_l \hat{\mathbf{a}}_{l,t}
\]

and

\[
\hat{E}(V_m b_{m,t} | Y_{1:T}, X_{1:T}) = V_m \hat{\mathbf{b}}_{m,t}.
\] (4.49)

We also estimate the posterior predictive variance of each element of \( U_l a_{l,t} \) and \( V_m b_{m,t} \) by calculating

\[
\hat{\text{Var}}(U_{l,j} a_{l,t} | Y_{1:T}, X_{1:T}) = \frac{1}{M} \sum_{i=1}^{M} (U_{l,j} a_{l,t}^{(i)} - U_{l,j} \hat{a}_{l,t})^2 \text{ for } j = 1, \ldots, N_Y
\]

and

\[
\hat{\text{Var}}(V_{m,j} b_{m,t} | Y_{1:T}, X_{1:T}) = \frac{1}{M} \sum_{i=1}^{M} (V_{m,j} b_{m,t}^{(i)} - V_{m,j} \hat{b}_{m,t})^2 \text{ for } j = 1, \ldots, N_X
\] (4.50)

where \( U_{l,j} \) and \( V_{m,j} \) denote the \( j^{th} \) rows of \( U_l \) and \( V_m \). Note that the ensembles \( U_l a_{l,t}^{(i)} \) and \( V_m b_{m,t}^{(i)} \) do not need to be stored for the calculations in (4.50). Figures 4.31 and 4.32 show maps of the posterior predicted means and standard deviations of \( U_l a_{l,t} \) and \( V_m b_{m,t} \) for the next year. For comparison, the observed data for the same period are shown in Figure 4.29. Note that the color coding are different for the predicted fields (Figure 4.31) and the data fields (Figure 4.29), the predicted fields have a much shorter range that the data fields. As was to be expected, with predicted mean amplitudes
Figure 4.29: ERA-40 and Polar MM5 temperature anomalies, in degrees Kelvin, for the last year.
Figure 4.30: Forecasted amplitudes, $b_{m,t,k}$ (left) and $a_{l,t,k}$ (right), without new ERA-40 data. The circles show the predicted means, the thick and thin lines show the middle 50% and 90% prediction intervals, respectively. The x-marks show the estimated amplitudes.

Close to zero, the predicted fields (Figure 4.31) do not show much structure and do not look very much like the data from this period (Figure 4.29). The Polar MM5 field for winter essentially shows the first EOF (see Figure 4.11), and indeed the only predicted $a_{l,t,k}$ amplitude that is not close to zero is the first amplitude for winter 2001 (see Figure 4.30). The maps of standard deviations (Figure 4.32) show that the predictive uncertainty tends to be lowest over ocean than over land/permanent ice. However, we note that the prediction standard deviation is very large, up to 10 degrees Kelvin at some locations.

As a simple overall measure of how close the predicted fields are to the observed data we calculate the squared error for each location and average that over all locations:

$$MSE(U\bar{a}_{it}) \equiv \frac{1}{NY}(Y_{it} - U_t\bar{a}_{it})'(Y_{it} - U_t\bar{a}_{it})$$ (4.51)

$$MSE(V\bar{b}_{m,t}) \equiv \frac{1}{NX}(X_{m,t} - V_m\bar{b}_{m,t})'(X_{m,t} - V_m\bar{b}_{m,t})$$ (4.52)
Figure 4.31: Forecast without new ERA-40 data. The maps show estimated posterior predicted mean ERA-40 and Polar MM5 temperature-anomaly fields, $U_l\alpha_{t,t}$ (squares) and $V_m\beta_{m,t}$ (circles), for winter 2001 - fall 2002 in degrees Kelvin.
Figure 4.32: Forecast without new ERA-40 data. The maps show estimated posterior predicted standard deviations for ERA-40 and Polar MM5 temperature fields, in degrees Kelvin.
Figure 4.33: $MSE(U_l\bar{a}_{l,t})$ (upper panel) for Polar MM5 temperatures and $MSE(V_m\bar{b}_{m,t})$ (lower panel) for ERA-40 data. Red points represent MSE for forecasted fields without new ERA-40 data.

Figure (4.33) shows $MSE(U_l\bar{a}_{l,t})$ and $MSE(V_m\bar{b}_{m,t})$ for all time points. The MSE up until fall season 2001 (black points) are calculated using the average of MCMC samples $a_{l,t}^{(i)}$ and $b_{m,t}^{(i)}$, while the MSE for the last year (red points) are calculated using the average of the predicted ensembles. We see that the $MSE(U_l\bar{a}_{l,t})$ for the Polar MM5 field is much higher the last four seasons than any season before. Also, $MSE(V_m\bar{b}_{m,t})$ is higher the last twelve months than most months before.

Another overall measure of how good the predictions are, is to see if the data falls within predictions intervals. Here we will only consider the primary field of interest, the Polar MM5 anomaly temperatures. Ideally we would like to obtain predictions of $Y_{l,t}$. The problem is, as mentioned in Section 4.6.1, that jointly forecast the whole vector $Y_{l,t}$ involves sampling a joint normal distribution with a $14641 \times 14641$
Table 4.9: Percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals, forecast without new ERA-40 data.

<table>
<thead>
<tr>
<th></th>
<th>Winter 2001</th>
<th>Spring 2001</th>
<th>Summer 2002</th>
<th>Fall 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% Prediction Interval</td>
<td>78.16%</td>
<td>51.84%</td>
<td>63.26%</td>
<td>54.05%</td>
</tr>
<tr>
<td>90% Prediction Interval</td>
<td>97.96%</td>
<td>95.05%</td>
<td>96.05%</td>
<td>92.08%</td>
</tr>
<tr>
<td>95% Prediction Interval</td>
<td>98.65%</td>
<td>97.98%</td>
<td>97.90%</td>
<td>95.28%</td>
</tr>
</tbody>
</table>

dimensional covariance matrix $r_i^{(i)} \tilde{R}_l$. However, we can obtain marginal predictions of each element of $Y_{l,t}$. Recalling the definition of $\tilde{R}_l$ in equation (4.16) the $j^{th}$ diagonal element of $r_i^{(i)} \tilde{R}_l$ is $r_i^{(i)}(c_l + \tilde{U}_{l,j} \tilde{D} \tilde{U}'_{l,j})$ where $\tilde{U}_{l,j}$ is the $j^{th}$ row vector of $\tilde{U}_l$. For each location $j$, $j = 1, \ldots, 14641$, we sample

$$y_{l,t,j}^{(i)} \sim N \left( U_{l,j} a_{l,t}^{(i)}, r_i^{(i)}(c_l + \tilde{U}_{l,j} \tilde{D} \tilde{U}'_{l,j}) \right), \quad i = 1, \ldots, M, \quad (4.53)$$

where $U_{l,j}$ is the $j^{th}$ row of $U_l$. For each location we then obtain the 50%, 90% and 95% prediction intervals. By performing the above calculations one location at a time we only work with one $M$ dimensional vector each time, but do not need to store all $M N_Y$-dimensional vectors $y_{l,t}^{(i)}$. We then see if the actual data falls within the prediction intervals for each location. Table 4.9 shows the percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals. The percentage of data that fall within each prediction interval are close to the level of that interval.
4.7.2 Statistical downscaling of ERA-40 data onto the Polar MM5 grid 2001-2002

We now consider predicting the Polar MM5 for the last year, winter of 2001 through fall of 2002 having ERA-40 data available for that period. Using Gibbs sampling, we obtained samples from the posterior distribution of the model described in Section 4.5 with the additions described in Section 4.6.2. The additional full conditional distributions needed for the Gibbs sampler are listed in Appendix C.2. With the additional amplitude vectors, for the time period winter of 2001 through fall of 2002, there are 1,908 unknown parameters in this model. We obtained 30,000 MCMC samples and discarded the first 1,000 as burn-in, leaving 29,000 MCMC samples to use for inference. We examined trace-plots of all parameters (not shown here) and they showed no indication of convergence problems.

Figure 4.34 shows estimated posterior means and credible intervals of each element of the amplitude vectors of the last year. Note that for this time period the $a_{l,t,k}$ amplitudes are “predicted” since the Polar MM5 data for this period are not included in the model. However, the $b_{m,t,k}$ amplitudes are “estimated” since the ERA-40 data for this period are included in the model. The credible intervals for the $b_{m,t,k}$ amplitudes are very narrow (left of Figure 4.34), their widths are similar to the widths of the credible intervals seen in Figure 4.19. Furthermore, the estimated amplitudes ($b_{m,t,k}$, see the x-marks in Figure 4.34) are almost identical to the posterior means. This is not surprising since the ERA-40 data for this period are included in the model. The posterior means of the predicted $a_{l,t,k}$ amplitudes (right of Figure 4.34) now have some structure to them, instead of all being close to zero as in Figure 4.30. The the 90% prediction intervals are still rather wide, especially for the first amplitude of
Figure 4.34: Estimated $b_{m,t,k}$ amplitudes (left) and predicted $a_{l,t,k}$ amplitudes (right) for downscaling 2001-2002 without new ERA-40 data. The circles show the predicted means, the thick and thin lines show the middle 50% and 90% prediction intervals, respectively. The x-marks show the estimated amplitudes.

winter 2001 but they are considerably narrower than those in Figure 4.30. Most of the estimated amplitudes ($a_{l,t,k}$, see the x-marks in Figure 4.34) fall within the 90% prediction intervals.

We estimate the posterior predicted means and standard deviations of $U_l a_{l,t}$ and the posterior mean and standard deviations of ERA-40 temperatures, $V_m b_{m,t}$, using the formulas in (4.49) and (4.50). Figures 4.35 and 4.36 show maps of the estimated means and standard deviations of $U_l a_{l,t}$ and $V_m b_{m,t}$ for winter 2001 - fall 2002. For comparison, the observed data for the same period are shown in Figure 4.29. Note that here the color coding for the predicted fields (Figure 4.35) and the data fields (Figure 4.29) are the same. This time the predicted fields (Figure 4.35) do resemble the data from this period (Figure 4.29).

For an easier visual comparison we show, in Figure 4.37, maps of the Polar MM5 data for winter 2001 - fall 2002 together with the predicted fields from both this and last Section. That is, we show both the predictions where no new ERA-40 data was
Figure 4.35: Downscaling 2001-2002 with new ERA-40 data. The maps show the estimated posterior mean ERA-40 fields and the estimated posterior predicted mean of Polar MM5 temperature-anomaly fields for winter 2001 - fall 2002 in degrees Kelvin.
Figure 4.36: Downscaling 2001-2002 with new ERA-40 data. The maps show the estimated posterior standard deviations of ERA-40 and Polar MM5 temperature-anomaly fields for winter 2001 - fall 2002 in degrees Kelvin.
available (also shown in Figure 4.35) and the downscaled temperatures (also shown in Figure 4.31). Note that in Figure 4.37 the color coding is scaled to the range of temperatures of only that time period. We see that the predicted fields (first column of Figure 4.37) do not resemble the structure seen in the data. The predicted fields for spring 2001 - fall 2002 are mostly zero. The downscaled temperatures (second column of Figure 4.37) show some of the same features as the data. The downscaled field for winter 2001 shows similar warm areas in West Antarctica (over Marie Byrd Land and the Ross Ice Shelf) and in Queen Maud Land. (The locations of these different geographic regions can be found on the map in Figure B.1.) The colder patch over East Antarctica (Queen Mary Land) is also represented in the downscaled field although slightly shifted, but the colder patch over the lower part of the Antarctic Peninsula is not present in the downscaled field. The three colder areas in the spring 2001 data are not present in the downscaled field for that season, but the warmer areas over the Ronne Ice Shelf and the Antarctic Peninsula are. The downscaled fields for summer 2002 does have the colder areas over West Antarctica and Queen Maud Land as the data show but they are not as large. Finally, the downscaled fields for fall 2002 have a very similar pattern of warmer temperatures over West Antarctica but does not quite catch the large area with very warm temperatures over East Antarctica.

For an overall measure of how well the predicted fields match the actual data we again use the averaged squared error, $MSE(U_{t})$ and $MSE(V_{t})$ from equation (4.52), see Figure 4.38). The MSEs for downscaled fields are lower than those in Figure 4.33 (except for spring season).

We also predicted the Polar MM5 temperatures (marginally) at each location as in (4.53) and calculated the percentage of locations where the Polar MM5 temperature
Figure 4.37: Estimated posterior predicted mean Polar MM5 temperature fields using no new ERA-40 data (first column), estimated posterior mean of downscaled Polar MM5 temperature fields using the ERA-40 data from the last year (second column) and the actual Polar MM5 temperature data.
4.7.3 Statistical downscaling of ERA-40 data onto the Polar MM5 grid 2000-2001

In an effort to assess the downscaling performance of our model further, we here consider predicting the Polar MM5 temperatures of the time period winter of 2000 through fall of 2001. Here we obtain the EOF and OMCP basis vectors using data only through fall of 2000 but we include the ERA-40 data from winter of 2000 through fall of 2001 in the model.
<table>
<thead>
<tr>
<th></th>
<th>Winter 2001</th>
<th>Spring 2001</th>
<th>Summer 2002</th>
<th>Fall 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% Prediction Interval</td>
<td>70.72%</td>
<td>38.19%</td>
<td>52.52%</td>
<td>50.67%</td>
</tr>
<tr>
<td>90% Prediction Interval</td>
<td>97.40%</td>
<td>77.58%</td>
<td>89.39%</td>
<td>80.60%</td>
</tr>
<tr>
<td>95% Prediction Interval</td>
<td>98.60%</td>
<td>86.45%</td>
<td>94.58%</td>
<td>85.31%</td>
</tr>
</tbody>
</table>

Table 4.10: Percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals, for downscaling 2001-2002 with new ERA-40 data.

The first five EOFs and OMCPs obtained from data up until fall of 2000 are shown in Figures B.36 - B.39 in Appendix B.4. These patterns are similar as those obtained with data through fall of 2001, shown in Figures 4.9 - 4.10, except that some have opposite sign. Also, the first four patterns explain similar proportions of the total sample variance in each field as before, see Table 4.11 versus Table 4.11.

Using Gibbs sampling, we obtained samples from the posterior distribution of the same model as in Section 4.7.1 except here we include Polar MM5 data through fall of 2000 and ERA-40 data through fall of 2001. We obtained 30,000 MCMC samples and discarded the first 1,000 as burn-in, leaving 29,000 MCMC samples to use for inference. We examined trace-plots of all parameters (not shown here) and they showed no indication of convergence problems.

Figure 4.39 shows estimated posterior means and credible intervals of each element of the amplitude vectors of the last year. Note that, as for Figure 4.34, the $a_{l,t,k}$ amplitudes are “predicted” since the Polar MM5 data for this period are not included in the model. However, the $b_{m,t,k}$ amplitudes are “estimated” since the ERA-40 data for this period are included in the model. Similar to what we saw in Figure 4.34 the credible intervals for the $b_{m,t,k}$ amplitudes are very narrow (left of Figure 4.39) but
EOFs of Polar MM5 data

<table>
<thead>
<tr>
<th>Season</th>
<th>Summer</th>
<th>Fall</th>
<th>Winter</th>
<th>Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. expl.</td>
<td>72.65%</td>
<td>68.55%</td>
<td>81.42%</td>
<td>77.26%</td>
</tr>
</tbody>
</table>

OMCPs of ERA-40 data

<table>
<thead>
<tr>
<th>Month</th>
<th>December</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. expl.</td>
<td>48.7%</td>
<td>73.65%</td>
<td>85.14%</td>
<td>63.12%</td>
<td>54.84%</td>
<td>55.23%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Month</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. expl.</td>
<td>51.46%</td>
<td>37.57%</td>
<td>56.08%</td>
<td>65.02%</td>
<td>56.95%</td>
<td>62.47%</td>
</tr>
</tbody>
</table>

Table 4.11: Proportion of total variance explained by the first four EOFs for each season and the first four OMCPs for each month, using only data through fall 2000.

the 90% prediction intervals for the predicted $a_{l,t,k}$ amplitudes (right of Figure 4.39) are rather wide. Most of the estimated amplitudes fall within the 90% prediction intervals (see the x-marks in Figure 4.39).

We estimate the posterior predicted means and standard deviations of $U_{l}a_{l,t}$ using the formulas in (4.49) and (4.50). Figure 4.40 shows maps of the estimated means and standard deviations of $U_{l}a_{l,t}$ as well as the actual Polar MM5 anomaly data for the last year. The overall agreement between actual data 2000 - 2001 and downscaled temperatures is rather good and similar to what we saw for the 2001 - 2002 period (Figure 4.37). Some patterns in the data are present in the downscaled temperatures but others are not. For example, both the warmer temperatures over West Antarctica in spring 2000 data and the colder temperatures over the same area in the summer 2001 data are also present in the downscaled temperatures for those seasons. However, the downscaled temperatures for fall 2001 do not show any of the rich structure of the data from the same time period. This is related to the fact that the first predicted amplitude for this season (black circle in Figure 4.39) is very different.
Figure 4.39: Estimated $b_{m,t,k}$ amplitudes (left) and predicted $a_{l,t,k}$ amplitudes (right) for downscaling 2000-2001 without new ERA-40 data. The circles show the predicted means, the thick and thick lines show the middle 50% and 90% prediction intervals, respectively. The x-marks show the estimated amplitudes.

to the first amplitude estimated from the data (black x-mark in Figure 4.39) The estimated standard deviations (first column of Figure 4.40) do to some extent reflect the difference in actual and downscaled temperatures, see for example the Ross Ice Shelf in fall 2001 and both Wast and East Antarctica in winter 2000.

We calculated the averaged squared error, $MSE(U_{l,t})$ and $MSE(V_{m,t})$ from equation (4.52), see Figure 4.41. The MSEs here are on the same order as those in Figure 4.38 indicating a similar predictive ability of the model for these time periods. Table 4.12 shows the percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals of marginally predicted Polar MM5 data. As in Table 4.10 the percentage of data that fall within each prediction interval is close to the corresponding levels in most cases.
Figure 4.40: Estimated posterior predicted standard deviations (first column) and means (second column) of Polar MM5 temperature fields from downscaling the 2000-2001 ERA-40 data and the actual Polar MM5 temperature data (third column).
Figure 4.41: $MSE(U_{\bar{a}_{t,t}})$ (upper panel) for Polar MM5 temperatures and $MSE(V_{m_{t},b_{m,t}})$ (lower panel) for ERA-40 data. Red points represent MSE for down-scaled fields with new ERA-40 data 2000-2001.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50% Prediction Interval</td>
<td>56.27%</td>
<td>57.33%</td>
<td>58.36%</td>
<td>53.54%</td>
</tr>
<tr>
<td>90% Prediction Interval</td>
<td>92.59%</td>
<td>92.28%</td>
<td>92.15%</td>
<td>89.97%</td>
</tr>
<tr>
<td>95% Prediction Interval</td>
<td>96.48%</td>
<td>96.54%</td>
<td>95.97%</td>
<td>93.89%</td>
</tr>
</tbody>
</table>

Table 4.12: Percentage of locations where the Polar MM5 temperature data fall within the 50%, 90% and 95% prediction intervals, for downscaling 2000-2001 with new ERA-40 data.
Chapter 5: Discussion

In this thesis we have outlined a general Bayesian hierarchical approach to modeling of two spatio-temporal processes. We discussed several aspects of the general framework in Chapter 2. Furthermore, we have proposed new methods of obtaining data-dependent basis vectors, namely MCPs and OMCPs (Chapter 3). We have successfully applied these methods to statistical downscaling of surface temperatures in the Antarctic (Chapter 4). We close with a brief discussion of a few points.

The focus of this thesis can be viewed as a dimension-reduced regression approach. Our goal has been to predict one spatio-temporal process from another, using a dimension-reduced representation of both processes. The motivation for using the dimension reduction approach is the need to be able to handle large data sets where a full representation of the processes is not feasible due to computational constraints. We recognize, however, that the dimension-reduced approach may lead to sub-optimal predictors. We partially addressed this issue in Section 2.1.3. Assuming a linear model between two spatial processes we investigated the difference in mean square error (MSE) between the optimal predictor and predictors based on dimension reduced approach. We demonstrated that the increase in MSE depends on how well the dimension-reduced model captures the full linear model (see equation 2.29). Furthermore, we showed that the MSE of a dimension-reduced predictor depends on
how good the dimension-reduced representation of the target process is and the MSE of the predictor of the amplitudes (see equation 2.34).

We demonstrated in Chapter 3 that both MCA and CCA can be used as tools in dimension-reduced modeling of two spatio-temporal processes. However, when the goal is to predict one process ($Y$) from the other ($X$) we prefer the new approach of MCPs or OMCPs for the process ($X$) used to predict the other. The main reason is the hierarchical setup of first choosing basis vectors for the target process ($Y$) and then obtaining the vectors that best (linearly) predict these basis vectors. In Chapter 4 we use EOFs because of their property of maximizing the proportion of total variance explained, but any other set of basis vectors could be considered. For example, if the target process is a combination of known physical patterns that we wish to predict, we can do so with MCPs or OMCPs. We prefer OMCPs over MCPs both because of computational convenience and because the MCPs are not guaranteed to be linearly independent.

The statistical downscaling example in Chapter 4 demonstrated the effectiveness of the hierarchical dimension-reduced modeling approach for two spatio-temporal processes. The Polar MM5 and ERA-40 data are both very large and a direct modeling approach would be computationally very difficult. The dimension-reduced hierarchical model produced very satisfactory results when downscaling ERA-40 data onto the Polar MM5 grid, as was summarized in Sections 4.7.2 and 4.7.3. However, the forecasting performance of the model left something to be desired (see Section 4.7.1). In essence, the OMCPs of ERA-40 data were very effective for predicting Polar MM5 data (i.e. downscaling), but they were not effective in capturing the temporal evolution of the ERA-40 data. This is perhaps not surprising since the focus of the
OMCPs is on the spatial variation of the data but not the temporal variation. This
same concern is true for EOFs, see for example Section 7.3.1 in Cressie and Wikle
(2011). But as a tool for predicting one process from another the OMCPs and MCPs
are very promising.
Appendix A: Alternative Proofs and Useful Theorems

A.1 Some useful theorems

In this appendix section we list a few theorems (and some proofs) that are referred to in the text. These are neither new facts nor new proofs but are provided here for completeness.

A.1.1 Normal-Normal updates

In general, even if \(X|Y\) and \(Y\) are both Normal distributions the joint distributions \([X,Y]\) is not necessarily Normal, but for the following linear structure it is.

Theorem A.1.1. **Conditional Normal Structure that leads to Joint Normality**

Let \(X\) and \(Y\) be \(n\) and \(m\) dimensional random vectors. If \([X|Y]\) is \(N_n(AY, \Gamma)\) and \([Y]\) is \(N_m(\mu, \Sigma)\) then

\[
[X, Y] = N_{n+m} \left( \begin{pmatrix} A \mu \\ \mu \end{pmatrix}, \Sigma = \begin{pmatrix} \Gamma + A \Sigma A' & A \Sigma \\ \Sigma A' & \Sigma \end{pmatrix} \right) \quad (A.1)
\]

156
Proof. The joint distribution of $X$ and $Y$ is

$$
[X, Y] = [X|Y][Y]
$$

$$
= \frac{1}{(2\pi)^{m/2}|\Gamma|^{1/2}} \exp \left\{ -\frac{1}{2}(X - AY)'\Gamma^{-1}(X - AY) \right\}
$$

$$
= \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(Y - \mu)'\Sigma^{-1}(Y - \mu) \right\}
$$

$$
= \frac{1}{(2\pi)^{(m+n)/2}|\Gamma|^{1/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} \left[ X\Gamma^{-1}X - 2X\Gamma^{-1}AY + Y'\Gamma^{-1}AY + Y'\Sigma^{-1}Y - 2Y'\Sigma^{-1}(\mu + \mu'\Sigma^{-1}\mu) \right] \right\}
$$

(A.2)

The density function for the distribution in (A.1) is

$$
\frac{1}{(2\pi)^{(m+n)/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} \left[ \left( \begin{array}{c} X \\ Y \end{array} \right) - \left( \begin{array}{c} A\mu \\ \mu \end{array} \right) \right]' \Sigma^{-1} \left( \begin{array}{c} X \\ Y \end{array} \right) - \left( \begin{array}{c} A\mu \\ \mu \end{array} \right) \right] \right\}
$$

(A.3)

Using part (i) of lemma A.1.3 the determinant of the covariance matrix can be written as

$$
|\tilde{\Sigma}| = |\Sigma| |\Gamma + A\Sigma A' - A\Sigma\Sigma^{-1}\Sigma A'| = |\Sigma||\Gamma|
$$

so the ratios in front the exponential terms in (A.2) and (A.3) are equal. Using part (i) of lemma A.1.4 the inverse of $\tilde{\Sigma}$ is

$$
\tilde{\Sigma}^{-1} = \begin{pmatrix} \Gamma^{-1} & -\Gamma^{-1}A\Sigma\Sigma^{-1} \\ -\Sigma^{-1}\Sigma A\Gamma^{-1} & \Sigma^{-1} + \Sigma^{-1}\Sigma A\Gamma^{-1}A\Sigma\Sigma^{-1} \end{pmatrix}
$$

$$
= \begin{pmatrix} \Gamma^{-1} & -\Gamma^{-1}A \\ -A\Gamma^{-1} & \Sigma^{-1} + A\Gamma^{-1}A \end{pmatrix}
$$

Now, the term in the square brackets "\[\]" in (A.3) is

$$
\begin{aligned}
\left[ & \right] = (X - A\mu)'\Gamma^{-1}(X - A\mu) - (X - A\mu)'\Gamma^{-1}A(Y - \mu) \\
& - (Y - \mu)'A\Gamma^{-1}(X - A\mu) + (Y - \mu)'(\Sigma^{-1} + A\Gamma^{-1}A)(Y - \mu) \\
= & X\Gamma^{-1}X - 2X\Gamma^{-1}A\mu + \mu' A\Gamma^{-1}A\mu \\
& - X\Gamma^{-1}AY + X\Gamma^{-1}A\mu + \mu' A\Gamma^{-1}AY - \mu' A\Gamma^{-1}A\mu \\
& - Y'A\Gamma^{-1}X + Y'A\Gamma^{-1}A\mu + \mu' A\Gamma^{-1}X - \mu' A\Gamma^{-1}A\mu \\
& + Y'\Sigma^{-1}Y - 2Y'\Sigma^{-1}(\mu + \mu'\Sigma^{-1}\mu) + Y'A\Gamma^{-1}AY - 2Y'A\Gamma^{-1}A\mu + \mu' A\Gamma^{-1}A\mu
\end{aligned}
$$
Noting that since all terms are scalars we can transpose each of them and get, for example, $Y'\Gamma^{-1}X = X\Gamma^{-1}AY$. So by re-arranging terms we get

$$[\cdot] = X\Gamma^{-1}X - 2X\Gamma^{-1}AY + Y'\Gamma^{-1}AY + Y'\Sigma^{-1}Y - 2Y'\Sigma^{-1}\mu + \mu'\Sigma^{-1}\mu$$

$$- 2X\Gamma^{-1}A\mu + 2X\Gamma^{-1}A\mu + 2\mu'\Gamma^{-1}A\mu - 2\mu'\Gamma^{-1}A\mu$$

$$+ 2Y'\Gamma^{-1}A\mu - 2Y'\Gamma^{-1}A\mu$$

$$= X\Gamma^{-1}X - 2X\Gamma^{-1}AY + Y'\Gamma^{-1}AY + Y'\Sigma^{-1}Y - 2Y'\Sigma^{-1}\mu + \mu'\Sigma^{-1}\mu$$

So what is in the square brackets in (A.3) is the same as what is in the square brackets in (A.2) and hence the density functions in (A.3) and (A.2) are the same.

A useful corollary:

**Corollary A.1.2. The other marginal and the other conditional**

If $[X|Y] = N_n(AY, \Gamma)$ and $[Y] = N_m(\mu, \Sigma)$ then

(i) the other marginal distribution is $[X] = N_n(A\mu, \Gamma + A\Sigma A')$ and

(ii) the other conditional distribution is

$$[Y|X] = N_m\left(\mu + \Sigma A' (\Gamma + A\Sigma A')^{-1} (X - A\mu), \Sigma - \Sigma A' (\Gamma + A\Sigma A')^{-1} A\Sigma\right)$$

$$= N_m\left((\Gamma^{-1} - \Sigma)^{-1} (\Gamma^{-1}X + \Sigma^{-1}\mu), (\Gamma^{-1} - \Sigma)^{-1}\right).$$

*Proof.* (i) and (ii) follow directly as the marginal and conditional distributions of the joint distribution $[X, Y]$ from Theorem A.1.1.

**Lemma A.1.3. Determinant of a partitioned matrix.** Let $A$ be a matrix partitioned as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

158
(i) If $|A_{22}| \neq 0$ then $|A| = |A_{22}| |A_{11} - A_{12}A^{-1}_{22}A_{21}|$

(ii) If $|A_{11}| \neq 0$ then $|A| = |A_{11}| |A_{22} - A_{21}A^{-1}_{11}A_{12}|$

Proof. See page 36 in (Ravishanker and Dey, 2002)

Lemma A.1.4. Inverse of a partitioned matrix. Suppose a non-singular matrix $A$ is partitioned as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

and $B = A^{-1}$ is partitioned the same way.

(i) if $|A_{22}| \neq 0$ then

$$B_{11} = (A_{11} - A_{12}A^{-1}_{22}A_{21})^{-1}$$

$$B_{12} = -B_{11}A_{12}A^{-1}_{22}$$

$$B_{21} = -A^{-1}_{22}A_{21}B_{11}$$

$$B_{22} = A^{-1}_{22} + A^{-1}_{22}A_{21}B_{11}A_{12}A^{-1}_{22}$$

(ii) if $|A_{11}| \neq 0$ then

$$B_{11} = A^{-1}_{11} + A^{-1}_{11}A_{12}B_{22}A_{21}A^{-1}_{11}$$

$$B_{12} = -A^{-1}_{11}A_{12}B_{22}$$

$$B_{21} = -B_{22}A_{21}A^{-1}_{11}$$

$$B_{22} = (A_{22} - A_{21}A^{-1}_{11}A_{12})^{-1}$$

Proof. See page 37 in (Ravishanker and Dey, 2002)
A.1.2 Kalman Filter as Bayesian Filtering

It is a known fact that Kalman Filter gives the mean vectors and covariance matrices in Bayesian Filtering when the distributions are Normal and parameters are known, see for example West and Harrison (1997) and Wikle and Berliner (2007).

Theorem A.1.5. Kalman Filter as Bayesian Filtering

Suppose we have the state space model in (2.102) and (2.103) where the transition matrix $M_t$, the model matrix $\Phi_t$ and covariance matrices $\Sigma_t$ and $\Gamma_t$ are known for all $t$. Let

$$
\mu^f_t = M_t \mu^a_{t-1}, \quad P^f_t = \Gamma_t + M_t P^a_{t-1} M_t',
$$

$$
\mu^a_t = \mu^f_t + K_t \left( (Y_t - H_t \mu^f_t) - H_t \mu^f_t \right) \quad \text{and} \quad P^a_t = (I - K_t \Phi_t) P^f_t.
$$

where $K_t = P^f_t \Phi_t' (\Phi_t P^f_t \Phi_t' + \Sigma_t)^{-1}$. Given the assumptions of conditional independence and Markov structure in (2.59) and (2.62) the following holds for all $t = 1, \ldots, T$:

(i) The forecasting distribution at time $t$ is Normal with mean $\mu^f_t$ and covariance matrix $P^f_t$

(ii) The analysis distribution (updated posterior) at time $t$ is Normal with mean $\mu^a_t$ and covariance matrix $P^a_t$

Proof. I will show this by induction. Also, I will leave out the parameter vector $\theta$ in the notation since it is assumed to be known.

Starting with $t = 1$ observe that $[a_1, b_1]$ is the marginal distribution of

$$
[a_1, b_1, a_0, b_0] = [a_1, b_1 | a_0, b_0] [a_0, b_0]
$$
where \([a_1, b_1|a_0, b_0] = N(M_1 (a_{b_0}^{a_0}), \Gamma_1)\) and we assume that \([a_0, b_0] = N(\mu_0, \Gamma_0). So by Corollary A.1.2 (i) we get that
\[
[a_1, b_1] = N(M_1 \mu_0 = \mu_1^f, \Gamma_1 + M_1 \Gamma_0 M_1^f)
\]
and we let \(P_1^f = \Gamma_1 + M_1 \Gamma_0 M_1^f\). Also, \([a_1, b_1|Y_1, X_1] = \text{the "other conditional" distribution of } [a_1, b_1, Y_1, X_1] = [Y_1, X_1]a_1, b_1 \text{ where } [Y_1, X_1]a_1, b_1 \text{ is Normal with mean and covariance matrix}
\]
\[
\mu_1^f + P_1^f \Phi_1'(\Sigma_1 + \Phi_1 P_1^f \Phi_1')^{-1}(Z_1 - \Phi_1 \mu_1^f) = \mu_1^f + K_1((\dot{X}_1) - \Phi_1 \mu_1^f) = \mu_1^o
\]
\[
P_1^f + P_1^f \Phi_1'(\Sigma_1 + \Phi_1 P_1^f \Phi_1')^{-1} \Phi_1 P_1^f = P_1^f + K_1 \Phi_1 P_1^f = (I + K_1 \Phi_1) P_1^f = P_1^o
\]
Now I will show for \(t \geq 2\) that if \([a_t-1, b_t-1|Y_{1:(t-1)}, X_{1:(t-1)}] = N(\mu_t-1^o, P_t-1^o)\) then \([a_t, b_t|Y_{1:(t-1)}, X_{1:(t-1)}] = N(\mu_t^f, P_t^f)\) and \([a_t, b_t|Y_{1:t}, X_{1:t}] = N(\mu_t, P_t^o)\). First note that \([a_t, b_t|Y_{1:(t-1)}, X_{1:(t-1)}]\) is the marginal of
\[
[a_t, b_t, a_{t-1}, b_{t-1}|Y_{1:(t-1)}, X_{1:(t-1)}] = [a_t, b_t|a_{t-1}, b_{t-1}, Y_{1:(t-1)}, X_{1:(t-1)}] [a_{t-1}, b_{t-1}|Y_{1:(t-1)}, X_{1:(t-1)}]
\]
where the second step follows from lemma A.1.6. Since
\[
[a_t, b_t|a_{t-1}, b_{t-1}] = N(M_t (a_{b_{t-1}}^{a_{t-1}}), \Gamma_t)
\]
we get from Corollary A.1.2 (i) that
\[
[a_t, b_t|Y_{1:(t-1)}, X_{1:(t-1)}] = N(M_t \mu_t-1^o, \Gamma_t + M_t P_t-1^o M_t^f) = N(\mu_t^f, P_t^f)
\]
Also, note that \([a_t, b_t|Y_{1:t}, X_{1:t}]\) is the “other” conditional distribution of

\[
[a_t, b_t, Y_t, X_t|Y_{1:(t-1)}, X_{1:(t-1)}] = \\
[Y_t, X_t|a_t, b_t, Y_{1:(t-1)}, X_{1:(t-1)}][a_t, b_t|Y_{1:(t-1)}, X_{1:(t-1)}] \\
= [Y_t, X_t|a_t, b_t][a_t, b_t|Y_{1:(t-1)}, X_{1:(t-1)}]
\]

where the second step follows from lemma A.1.7. Since

\[
[Y_t, X_t|a_t, b_t] = N(\Phi_t(a, b), \Sigma_t)
\]

we get from Corollary A.1.2 (ii) that \([a_t, b_t|Y_{1:t}, X_{1:t}]\) is normal with mean vector

\[
\mu^f_t + P^f_t\Phi'(\Sigma_t + \Phi_tP^f_t\Phi'_t)^{-1}\left(\begin{bmatrix} Y_t \\ X_t \end{bmatrix} - \Phi_t\mu^f_t\right) = \mu^a_t \\
K_t\left(\begin{bmatrix} Y_t \\ X_t \end{bmatrix} - \Phi_t\mu^f_t\right) = \mu^a_t
\]

and covariance matrix

\[
P^f_t + P^f_t\Phi'(\Sigma_t + \Phi_tP^f_t\Phi'_t)^{-1}\Phi_tP^f_t = P^f_t + K_t\Phi_tP^f_t = (I + K_t\Phi_t)P^f_t = P^a_t
\]

\[\square\]

**Lemma A.1.6.** Given the conditional independence assumptions in (2.59) and the Markov structure in (2.62) we get

\[
[a_t, b_t|a_{t-1}, b_{t-1}, Y_{1:(t-1)}, X_{1:(t-1)}] = [a_t, b_t|a_{t-1}, b_{t-1}]
\]
Proof.

\[
\begin{aligned}
[a_t, b_t | a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)] &= \frac{[a_t, b_t, a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)]}{[a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)]} \\
&= \int \frac{[a_{1:t}, b_{1:t}, Y_1:(t-1), X_1:(t-1)]da_{1:t-1}db_{1:t-1}}{[a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)]} \\
&= \int \frac{[Y_1:(t-1), X_1:(t-1) | a_{1:t}, b_{1:t}] [a_t, b_t | a_{t-1}, b_{t-1}] [a_{1:t-1}, b_{1:t-1}] da_{1:t-2}db_{1:t-2}}{[a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)]} \\
&= \frac{[a_t, b_t | a_{t-1}, b_{t-1}]}{[a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)]} \int \frac{[Y_1:(t-1), X_1:(t-1) | a_{1:t-1}, b_{1:t-1}, a_{1:t-1}, b_{1:t-1}] da_{1:t-2}db_{1:t-2}}{[a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)]} \\
&= \frac{[a_t, b_t | a_{t-1}, b_{t-1}]}{[a_{t-1}, b_{t-1}, Y_1:(t-1), X_1:(t-1)]} [Y_1:(t-1), X_1:(t-1) | a_{t-1}, b_{t-1}] \\
&= [a_t, b_t | a_{t-1}, b_{t-1}]
\end{aligned}
\]

\[\square\]

Lemma A.1.7. Given the conditional independence assumptions in (2.59) we get

\[
[Y_t, X_t | a_t, b_t, Y_1:(t-1), X_1:(t-1)] = [Y_t, X_t | a_t, b_t]
\]

Proof.

\[
\begin{aligned}
[Y_t, X_t | a_t, b_t, Y_1:(t-1), X_1:(t-1)] &= \frac{[Y_1:t, X_1:t | a_t, b_t]}{[a_t, b_t, Y_1:(t-1), X_1:(t-1)]} \\
&= \int \frac{[Y_1:t, X_1:t | a_{1:t}, b_{1:t}] da_{1:t-1}db_{1:t-1}}{[a_t, b_t, Y_1:(t-1), X_1:(t-1)]} \\
&= \int \frac{[Y_t, X_t | a_{1:t}, b_{1:t}] [Y_1:(t-1), X_1:(t-1) | a_{1:t}, b_{1:t}] [a_{1:t}, b_{1:t}] da_{1:t-1}db_{1:t-1}}{[a_t, b_t, Y_1:(t-1), X_1:(t-1)]} \\
&= \frac{[Y_t, X_t | a_t, b_t]}{[a_t, b_t, Y_1:(t-1), X_1:(t-1)]} [Y_1:(t-1), X_1:(t-1) | a_t, b_t] \\
&= [Y_t, X_t | a_t, b_t]
\end{aligned}
\]

\[\square\]
A.1.3 Singular Value Decomposition

A mathematical technique heavily relied upon in this thesis is singular value decomposition (SVD). See for example (Johnson and Wickern, 2002, p. 101)

Theorem A.1.8. Singular Value Decomposition.

If $X$ is an $N \times T$ dimensional matrix of real numbers there exist orthonormal matrices $U_{N \times N}$ and $V_{T \times T}$ such that

$$X_{N \times T} = U_{N \times N} \Lambda_{N \times T} V_{T \times T}'$$ (A.4)

where $\Lambda$ is a non-square diagonal matrix. That is, for $i = 1, \ldots, d = \min(N, T)$ the $(i, i)$ entries of $\Lambda$ are $\lambda \geq 0$ and other entries are zero. The diagonal entries in $\Lambda$ are called singular values and the columns of $U$ and $V$ are called left and right singular vectors.

Properties of SVD:

(i) Since $U$ and $V$ are orthonormal we have $U'U = UU' = I$ and $V'V = VV' = I$

(ii) SVD has a direct connection to the spectral decomposition of $XX'$ and $X'X$:

- The matrix $XX'$ has eigenvalues $\lambda_i^2$ and it’s eigenvectors are the column vectors of $U$ $(u_1, \ldots, u_N)$ since $XX' = U \Lambda V' V \Lambda U' = U \Lambda^2 U'$

- The matrix $X'X$ has eigenvalues $\lambda_i^2$ and it’s eigenvectors are the column vectors of $V$ $(v_1, \ldots, v_T)$ since $X'X = V \Lambda U' U \Lambda V' = V \Lambda^2 V'$

164
A.1.4 Density of the square-root of an IG random variable

A convenient and conjugate prior for a variance, say $\sigma^2$ is the Inverse gamma distribution. The density function for $IG(a,b)$ is

$$f_{\sigma^2}(\sigma^2) = \frac{b^a}{\Gamma(a)} \frac{1}{(\sigma^2)^{a+1}} \exp\left\{ -\frac{b}{\sigma^2} \right\}$$

(A.5)

When displaying posterior inference about variances it is usually more natural to make inference about the standard deviation which has a more straightforward interpretation. And often we would like to compare estimated posterior densities of $\sigma$ to the prior on $\sigma$ and so we need the density function for $\sigma$.

In general if $X \sim f_X(x)$ and $Y \sim g(X)$ where $g$ is a monotone function then

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{dg^{-1}(y)}{dy} \right|$$

(A.6)

In our case we have $\sigma = g(\sigma^2) = \sqrt{\sigma^2}$ and $g^{-1}(\sigma) = \sigma^2$ and then

$$\frac{dg^{-1}(\sigma)}{d\sigma} = 2\sigma .$$

Therefore the density function for $\sigma$ is

$$f_{\sigma}(\sigma) = \frac{b^a}{\Gamma(a)} \frac{2\sigma}{(\sigma^2)^{a+1}} \exp\left\{ -\frac{b}{\sigma^2} \right\} = \frac{2b^a}{\Gamma(a)} \frac{1}{\sigma^{2a+1}} \exp\left\{ -\frac{b}{\sigma^2} \right\}$$

(A.7)

A.2 Alternative proofs for MCPs and OMCPs

Here I offer alternative proofs of Theorem 3.4.1 and Theorem 3.5.1 based on Lagrange multipliers.

A.2.1 Alternative proof of Theorem 3.4.1

Proof. We define the Lagrange function

$$\Lambda(\mathbf{v}, \lambda) = c'\mathbf{v} + \lambda(\mathbf{v}'\mathbf{v} - 1) = \sum_{i=1}^{n} c_i v_i + \lambda \left( \sum_{i=1}^{n} v_i^2 - 1 \right)$$

(A.8)
Setting the gradient vector equal to zero yields

\[
\frac{\partial \Lambda}{\partial v_i} = c_i + 2\lambda v_i = 0 \quad \Rightarrow \quad v_i = -\frac{c_i}{2\lambda} \quad \text{for } i = 1, \ldots, n \quad (A.9)
\]

\[
\frac{\partial \Lambda}{\partial \lambda} = v'v - 1 = 0 \quad \Rightarrow \quad v'v = 1 \quad (A.10)
\]

Putting the \( n \) expressions in (A.9) into (A.10) gives

\[
1 = v'v = \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} \left( -\frac{c_i}{2\lambda} \right)^2 = \frac{1}{4\lambda^2} \sum_{i=1}^{n} c_i^2
\]

\[
\Rightarrow \lambda^2 = \frac{1}{4} \sum_{i=1}^{n} c_i^2 = \frac{1}{4} c'c
\]

\[
\Rightarrow \lambda = \pm \frac{1}{2} \sqrt{c'c} = \pm \frac{1}{2} ||c|| \quad (A.11)
\]

Inserting (A.11) into (A.9) gives

\[
v_i = \mp \frac{c_i}{||c||} \quad \text{for } i = 1, \ldots, n \quad (A.12)
\]

or in other words, \( v = \mp \frac{c}{||c||} \). Now note that if \( v = \frac{c}{||c||} \) the value of the objective function is positive: \( f(v) = \frac{c'c}{||c||} > 0 \). If \( v = -\frac{c}{||c||} \) the value of the objective function is negative: \( f(v) = -\frac{c'c}{||c||} < 0 \). Therefore the maximum is obtained for \( v = \frac{c}{||c||} \) and the minimum is obtained for \( v = -\frac{c}{||c||} \).

\[\square\]

**A.2.2 Alternative proof of Theorem 3.5.1**

**Proof.** The Lagrange function is

\[
\Lambda(v_{k+1},\lambda,\delta) = c'v_{k+1} + \lambda(v'_{k+1}v_{k+1} - 1) + \sum_{j=1}^{k} \delta_j v'_{k+1}v_i \quad (A.13)
\]

where \( \delta = (\delta_1, \ldots, \delta_k) \). The derivatives are set equal to zero:

\[
\frac{\partial \Lambda}{\partial v_{k+1}} = c + \lambda 2v_{k+1} + \sum_{j=1}^{k} \delta_j v_j = 0 \quad \Rightarrow \quad v_{k+1} = -\frac{1}{2\lambda} \left( c + \sum_{j=1}^{k} \delta_j v_j \right) \quad (A.14)
\]

\[
\frac{\partial \Lambda}{\partial \lambda} = v'_{k+1}v_{k+1} - 1 = 0 \quad (A.15)
\]

\[
\frac{\partial \Lambda}{\partial \delta_j} = v'_{k+1}v_j = 0, \quad j = 1, \ldots, k \quad (A.16)
\]
Note that $\frac{\partial \Lambda}{\partial v_{k+1}}$ stands for $(\frac{\partial \Lambda}{\partial v_{k+1,1}}, \ldots, \frac{\partial \Lambda}{\partial v_{k+1,N}})'$. Insert (A.14) into (A.16) and we get for $j = 1, \ldots, k$

$$0 = v'_{k+1} v_j = -\frac{1}{2\lambda} \left(c'v_j + \left(\sum_{i=1}^{k} \delta_i v_i\right)'v_j\right) = -\frac{1}{2\lambda} (c'v_j + \delta_j)$$

The last step follows from the fact that $v_1, \ldots, v_k$ are orthonormal vectors, so $v_j'v_i$ is 1 if $i = j$ but 0 otherwise. Now, assuming that $\lambda \neq 0$ we get

$$\delta_j = -c'v_j \quad (A.17)$$

From (A.14) we see that for $v_{k+1}$ to satisfy $v'_{k+1} v_{k+1} = 1$ we have

$$\lambda = \pm \frac{1}{2} \left\| c + \sum_{j=1}^{k} \delta_j v_j \right\| = \pm \frac{1}{2} \left\| c - \sum_{j=1}^{k} (c'v_j)v_j \right\| \quad (A.18)$$

Note that both the minus and plus sign give $v'_{k+1} v_{k+1} = 1$.

Now, inserting the expressions for $\lambda$ (A.18) and $\delta_j$ (A.17) into (A.14) we get

$$v = \mp \frac{1}{\left\| c - \sum_{j=1}^{k} (c'v_j)v_j \right\|} \left(c - \sum_{j=1}^{k} (c'v_j)v_j\right) \equiv \mp v^* \quad (A.19)$$

The only thing left is to determine is whether the optimal solution is the positive ($v^*$) or negative part ($-v^*$) of (A.19). Note that $v_1, \ldots, v_k$ is an orthonormal basis for $\mathbb{R}^k$. Let $w_{k+1}, \ldots, w_N$ be orthonormal vectors that together with $v_1, \ldots, v_k$ form an orthonormal basis for $\mathbb{R}^N$. Then we can write

$$c = \sum_{j=1}^{k} (c'v_j)v_j + \sum_{j=k+1}^{N} (c'w_j)w_j \quad (A.20)$$
and then

$$f(v^*) = c' \left( c - \sum_{j=1}^{k} (c'v_j)v_j \right) = c'c - c' \sum_{j=1}^{k} (c'v_j)v_j$$

$$\geq c'c - c' \sum_{j=1}^{k} (c'v_j)v_j - \sum_{j=k+1}^{N} (c'w_j)^2$$

$$= c'c - c' \sum_{j=1}^{k} (c'v_j)v_j - c' \sum_{j=k+1}^{N} (c'w_j)w_j$$

$$= c'c - c' \left( \sum_{j=1}^{k} (c'v_j)v_j + \sum_{j=k+1}^{N} (c'w_j)w_j \right)$$

$$= c'c - c'c = 0 . \quad (A.21)$$

So \( f(v^*) \geq 0 \) and also \( f(-v^*) \leq 0 \) so we have that the optimal vector is the positive part of (A.19). \( \square \)
Appendix B: Exploratory Data Analysis

This part of the Appendix shows extra Figures related to the exploratory analysis in Chapter 4. These include maps of the Polar MM5 and ERA-40 temperature fields (see Section B.1, Figures B.2 - B.7), maps of centered temperatures (section B.2, Figures B.8 - B.13) and various exploratory analysis of the OMCP and EOF amplitudes (Section B.3).

The first Figure, however, shows an overview map of Antarctica that shows the major geographical features on the Antarctic continent, see Figure B.1.

B.1 Polar MM5 and ERA-40 2-meter temperature fields

Figures B.2 - B.7 show the ERA-40 and Polar MM5 2 meter temperature fields for all the 24 years from 1979 to 2002. Temperatures are in Kelvin \( (273.15 \text{K} = 0^\circ \text{C}) \) and the color pallet is the same across all maps so that the same colors on different maps correspond to the same temperature range.
Figure B.1: Antarctica overview map, obtained from Landsat Image Mosaic of Antarctica (LIMA), http://lima.usgs.gov.
Figure B.2: Temperature fields, ERA-40 and Polar MM5, in Kelvin, years 1979-1982.
Figure B.3: Temperature fields, ERA-40 and Polar MM5, in Kelvin, years 1983-1986.
Figure B.4: Temperature fields, ERA-40 and Polar MM5, in Kelvin, years 1987-1990.
Figure B.5: Temperature fields, ERA-40 and Polar MM5, in Kelvin, years 1991-1994.
Figure B.6: Temperature fields, ERA-40 and Polar MM5, in Kelvin, years 1995-1998.
Figure B.7: Temperature fields, ERA-40 and Polar MM5, in Kelvin, years 1999-2002.
B.2 Centered Polar MM5 and ERA-40 2-meter temperature fields

Figures B.8 - B.13 show the centered ERA-40 and Polar MM5 temperature fields. As before, temperatures are in Kelvin. The color palette here is separate for the ERA-40 and Polar MM5 fields but it is the same across all years so that the same colors on different ERA-40 maps correspond to the same temperature range and the same goes for Polar MM5.

Note that the average that is used to center the data does not include the summer months of 1979 (ERA-40 data) nor the data after fall season of 2001. Temperatures of winter 2001 through fall of 2002, should therefore be thought of as anomalies rather than centered data.
Figure B.8: Centered temperature fields ERA-40 (circles) and Polar MM5 (squares) in degrees Kelvin, year 1979-1982.
Figure B.9: Centered temperature fields ERA-40 (circles) and Polar MM5 (squares) in degrees Kelvin, year 1983-1986.
Figure B.10: Centered temperature fields ERA-40 (circles) and Polar MM5 (squares) in degrees Kelvin, year 1987-1990.
Figure B.11: Centered temperature fields ERA-40 (circles) and Polar MM5 (squares) in degrees Kelvin, year 1991-1994.
Figure B.12: Centered temperature fields ERA-40 (circles) and Polar MM5 (squares) in degrees Kelvin, year 1995-1998.
Figure B.13: Centered temperature fields ERA-40 (circles) and Polar MM5 (squares) in degrees Kelvin, year 1999-2002.
B.3 Exploratory analysis of estimated EOF and OMCP amplitudes

This Appendix section contains plots used for exploratory analysis of amplitudes (see equation (4.3)) Section B.3.1, contains scatterplots of estimated EOF amplitudes \( \hat{a}_{l,t,k_X} \) for \( k_X = 1, \ldots, 5 \), versus the first five estimated OMCP amplitudes, \( \hat{b}_{m,t,k_X} \) for \( k_X = 1, \ldots, 5 \), where \( m \) is a month in season \( l \), see figures B.14 - B.24. Note that plots for \( \hat{a}_{l,t,k_Y} \) (summer) versus \( \hat{b}_{12,t,k_X} \) (December) are shown in figure 4.15 in section 4.4.

Section B.3.2 contains plots of the first five estimated OMCP amplitudes, \( \hat{b}_{m,t,k_1} \) for \( k_1 = 1, \ldots, 5 \), versus the first five estimated OMCP amplitudes from the month before, \( \hat{b}_{m-1,t,k_2} \) for \( k_2 = 1, \ldots, 5 \), see figures B.25 - B.35. Plots for \( \hat{b}_{12,t,k_1} \) (December) versus \( \hat{b}_{11,t,k_2} \) (November) are shown in figure 4.16 in section 4.4.

In all the plots the simple regression line is added, in red color if the amplitudes are from the same group (\( k_X = k_Y \) or \( k_1 = k_2 \)) but otherwise in blue.
B.3.1 Plots of $a_{t,t}$ vs. $b_{m,t}$

Figure B.14: Estimated EOF amplitudes for summer plotted against estimated OMCP amplitudes for January for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 

185
Figure B.15: Estimated EOF amplitudes for summer plotted against estimated OMCP amplitudes for February for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
Figure B.16: Estimated EOF amplitudes for fall plotted against estimated OMCP amplitudes for March for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
Figure B.17: Estimated EOF amplitudes for fall plotted against estimated OMCP amplitudes for April for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 

188
Figure B.18: Estimated EOF amplitudes for fall plotted against estimated OMCP amplitudes for May for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
Figure B.19: Estimated EOF amplitudes for winter plotted against estimated OMCP amplitudes for June for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
Figure B.20: Estimated EOF amplitudes for winter plotted against estimated OMCP amplitudes for July for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
Figure B.21: Estimated EOF amplitudes for winter plotted against estimated OMCP amplitudes for August for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 

192
Figure B.22: Estimated EOF amplitudes for spring plotted against estimated OMCP amplitudes for September for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
Figure B.23: Estimated EOF amplitudes for spring plotted against estimated OMCP amplitudes for October for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
Figure B.24: Estimated EOF amplitudes for spring plotted against estimated OMCP amplitudes for November for every combination of $k_Y = 1, \ldots, 5$ and $k_X = 1, \ldots, 5$. 
B.3.2 Plots of $b_{m,t}$ vs. $b_{m-1,t}$

Figure B.25: Estimated January OMCP amplitudes plotted against estimated December OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.26: Estimated February OMCP amplitudes plotted against estimated January OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 

197
Figure B.27: Estimated March OMCP amplitudes plotted against estimated February OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.28: Estimated April OMCP amplitudes plotted against estimated March OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.29: Estimated May OMCP amplitudes plotted against estimated April OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.30: Estimated June OMCP amplitudes plotted against estimated May OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.31: Estimated July OMCP amplitudes plotted against estimated June OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.32: Estimated August OMCP amplitudes plotted against estimated July OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.33: Estimated September OMCP amplitudes plotted against estimated August OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.34: Estimated October OMCP amplitudes plotted against estimated September OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
Figure B.35: Estimated November OMCP amplitudes plotted against estimated October OMCP amplitudes, for every combination of $k_1 = 1, \ldots, 5$ and $k_2 = 1, \ldots, 5$. 
B.4 EOFs and OMCPs for data through fall of 2001

Figure B.36: First five groups of OMCPs and EOF patterns for summer season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001.
Figure B.37: First five groups of OMCPs and EOF patterns for fall season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001.
Figure B.38: First five groups of OMCPs and EOF patterns for winter season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001.
Figure B.39: First five groups of OMCPs and EOF patterns for spring season ($v_{12,k}$, $v_{1,k}$, $v_{2,k}$ and $u_{s,k}$ for $k = 1, \ldots, 5$), for data only through fall of 2001.
Appendix C: Gibbs Sampler

In this Appendix Chapter we provide full conditional distributions for Gibbs samplers used to obtain MCMC samples for inference for the models in Section 4.5.

Table C.1 gives an overview of the unknown parameters in the model in Section 4.5. Note that if we include amplitudes for the last year, as we do to obtain predictions in Section 4.7.1, there will be more four more $a_{l,t}$ amplitude vectors and twelve more $b_{m,t}$ amplitudes vectors, that is 64 more parameters than are listed in Table C.1.

Given the assumptions stated in Sec. 4.5 the joint posterior can be written as:

\[
[a_{1:T}, b_{1:T}, r_{s,f,w,p}, s_{1:12}, h_{1:12}, B_{1:12}, C_{s,f,w,p}, D_{1:12} | Y_{1:T}, X_{1:T}] \\
\propto \prod_{l=s,f,w,p}^{T_l} \prod_{t=\tau_l}^{T_l} [Y_{l,t} | a_{l,t}, R_l] \prod_{m=1}^{T_m} \prod_{t=\tau_m}^{T_m} [X_{m,t} | b_{m,t}, S_m] \prod_{l=s,f,w,p} \prod_{t=\tau_l}^{T_l} [r_l] [C_l] \\
\times \prod_{t=\tau_s}^{T_s} \prod_{t=\tau_s}^{T_s} [a_{s,t} | b_{12,t-1}, b_{1,t}, b_{2,t}, H_s, C_s] \prod_{t=\tau_f}^{T_f} \prod_{t=\tau_f}^{T_f} [a_{f,t} | b_{3,t}, b_{4,t}, b_{5,t}, H_f, C_f] \\
\times \prod_{t=\tau_w}^{T_w} [a_{w,t} | b_{6,t}, b_{7,t}, b_{8,t}, H_w, C_w] \prod_{t=\tau_p}^{T_p} [a_{p,t} | b_{9,t}, b_{10,t}, b_{11,t}, H_p, C_p] \\
\times \prod_{m=1}^{12} \prod_{t=\tau_m}^{T_m} [b_{m,t} | b_{m-1,t}, B_m, D_m] \prod_{m=1}^{12} [h_m] [vec(B_m)] [s_m] [D_m]. \quad (C.1)
\]

Note that $H_m = diag(h_m)$, $R_l = r_l \tilde{R}_l$ and $S_m = s_m \tilde{S}_m$ where $\tilde{R}_l$ and $\tilde{S}_m$ are known matrices.
Table C.1: List of unknown parameters in the model. Note that there are 89 season of data included in the model. The last column shows the number of parameters for the chosen number of basis vectors, $K_Y = 4$ and $K_X = 4$. Note that, in comparison, the total number of observations is $14641 \times 89 + 2736 \times 89 \times 3 = 2,033,561$.

C.1 Full conditional distributions

In the following we list the full conditional distributions of all parameters in Table C.1. We use the notation $\left[X|\text{rest}\right]$ to denote the conditional distribution of a random variable $X$ given both the data and all other unknown parameters in the model.

**Polar MM5 amplitudes $a_{l,t}$**

Let $b_{l,t}$ denote the three ERA amplitude vectors that correspond to season $l$. The full conditional distribution of $a_{l,t}$ for each year and season is

$$
\left[a_{l,t}|\text{rest}\right] \propto \left[Y_{l,t}|a_{l,t}, R_l\right] \left[a_{l,t}|b_{l,t}, H_l, C_l\right]
$$

where $\left[Y_{l,t}|a_{l,t}, R_l\right] = N(U_l a_{l,t}, R_l)$ and $\left[a_{l,t}|b_{l,t}, H_l, C_l\right] = N(H_l b_{l,t}, C_l)$. Using corollary A.1.2 we get that $\left[a_{l,t}|\text{rest}\right]$ is $N(A^{-1}d, A^{-1})$ with

$$
A = U_l' R_l^{-1} U_l + C_l^{-1} \quad \text{and} \quad d = U_l' R_l^{-1} Y_{l,t} + C_l^{-1} H_l b_{l,t}
$$
Note that from (4.26) we have \( U_l' R_l^{-1} U_l = \frac{1}{r_l c_l} I_{K_Y} \). Note also that from (4.25) we have

\[
U_l' R_l^{-1} Y_{l,t} = \frac{1}{r_l c_l} U_l' Y_{l,t}
\]

where \( U_l, c_l \) and \( Y_{l,t} \) are constants and \( \frac{1}{r_l c_l} U_l' Y_{l,t} \) can be calculated (and stored) outside of the MCMC iterations.

**ERA amplitudes \( b_{m,t} \)**

The full conditional distribution of \( b_{m,t} \) for each year and month, except the first and last month, is

\[
[b_{m,t}|\text{rest}] \propto \begin{bmatrix} X_{m,t} | b_{m,t}, S_m \end{bmatrix} \begin{bmatrix} b_{m,t} | b_{m-1,t}, B_m, D_m \end{bmatrix} \begin{bmatrix} b_{m+1,t} | b_{m,t}, B_{m+1}, D_{m+1} \end{bmatrix} \begin{bmatrix} a_{l,t} | b_{l,t}, H_l, C_l \end{bmatrix}
\]

The three distributions with \( b_{m+1,t} \) on the right side of “|” are the following:

\[
[X_{m,t} | b_{m,t}, S_m] = N(V_m b_{m,t}, S_m)
\]

\[
[b_{m+1,t} | b_{m,t}, B_{m+1}, D_{m+1}] = N(B_{m+1} b_{m,t}, D_{m+1}) \quad \text{and}
\]

\[
[a_{l,t} | b_{l,t}, H_l, C_l] = N(H_l b_{l,t}, C_l)
\]

Let \( m_1 \) and \( m_2 \) denote the two other months in season \( l \) and let \( b = H_{m_1} b_{m_1,t} + H_{m_2} b_{m_2,t} \). Then (C.8) can written as

\[
[a_{l,t} | b_{l,t}, H_l, C_l] = N(H_m b_{m,t} + b, C_l)
\]

Using the assumption of conditional independence we get

\[
\begin{pmatrix}
X_{m,t} \\
B_{m+1,t} \\
H_m
\end{pmatrix}
\begin{pmatrix}
b_{m,t} \\
a_{l,t}
\end{pmatrix}
\sim N
\begin{pmatrix}
V_m \\
B_{m+1} \\
H_m
\end{pmatrix}
\begin{pmatrix}
b_{m,t} \\
a_{l,t}
\end{pmatrix}
+
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix}
+ \begin{pmatrix}
S_m & 0 & 0 \\
0 & D_{m+1} & 0 \\
0 & 0 & C_l
\end{pmatrix}
\]

213
We also have $[b_{m,t}|b_{m-1,t}, B_m, D_m] = N(B_m b_{m-1,t}, D_m)$ so using corollary A.1.2 we get that $[b_{m,t}|\text{rest}]$ is $N(A^{-1}d, A^{-1})$ with

$$A = V'_m S_m^{-1} V_m + B'_{m+1} D_{m+1}^{-1} B_{m+1} + H' C_l^{-1} H_m + D_m^{-1}$$

$$d = V'_m S_m^{-1} X_{m,t} + B'_{m+1} D_{m+1}^{-1} b_{m+1,t} + H' C_l^{-1} (a_{l,t} - b) + D_m^{-1} B_m b_{m-1,t}.$$  \hfill (C.11)

Some of the calculations can be simplified by noting that from (4.27) we have

$$V'_m S_m^{-1} V_m = \frac{1}{s_m c_m} I_K$$  \hfill and  \hfill $$V'_m S_m^{-1} X_{m,t} = \frac{1}{s_m c_m} V'_m X_{l,t}.$$  \hfill (C.12)

$V_m$, $c_m$ and $X_{m,t}$ are constants and $\frac{1}{c_m} V'_m X_{m,t}$ can be calculated (and stored) outside of the MCMC iterations.

**The last ERA amplitude $b_{m,t}$, with data**

The full conditional distribution of the last $b_{m,t}$ amplitude vector is different from others because there is no $b_{m+1,t}$ amplitude after it. Hence

$$[b_{m,t}|\text{rest}] \propto [X_{m,t}|b_{m,t}, S_m][b_{m,t}|b_{m-1,t}, B_m, D_m][a_{l,t}|b_{l,t}, H_l, C_l]$$  \hfill (C.13)

Using similar steps as before equation (C.11) we get that $[b_{m,t}|\text{rest}]$ is $N(A^{-1}d, A^{-1})$ with

$$A = V'_m S_m^{-1} V_m + H' C_l^{-1} H_m + D_m^{-1}$$

$$d = V'_m S_m^{-1} X_{m,t} + H' C_l^{-1} (a_{l,t} - b) + D_m^{-1} B_m b_{m-1,t}.$$  \hfill (C.14)

**The first ERA amplitude $b_{2,1}$**

The full conditional distribution of the first ERA amplitude is different from the others because there is neither data, nor a $a_{l,t}$ amplitude at that timepoint and no $b_{m-1,1}$ amplitude before it. Hence

$$[b_{2,1}|\text{rest}] \propto [b_{3,1}|b_{2,1}, B_3, D_3][b_{2,1}]$$  \hfill (C.15)
where \([b_{3,1}|b_{2,1}, B_3, D_3] = N(B_3 b_{2,1}, D_3)\) and \([b_{2,1}] = N(\mu_b, \Sigma_b)\). Using corollary A.1.2 we get that \([b_{2,1}|\text{rest}] = N(A^{-1}d, A^{-1})\) with

\[
A = B_3' D_3^{-1} B_3 + \Sigma_b^{-1} \quad \text{and} \quad d = B_3' D_3^{-1} b_{3,1} + \Sigma_b^{-1} \mu_b \tag{C.16}
\]

**Transition matrices** \(H_m\)

Recall that the transition matrices \(H_m\) are diagonal matrices and that \(h_m = (H_{m,1}, \ldots, H_{m,K_Y})'\) is a vector of the diagonal elements of \(H_m\). Also recall the \(K_Y \times 3K_X\) dimensional matrices \(H_l\) from (4.11). Let \(h_l\) be the \(3K_Y\) dimensional vector of non-zero elements of \(H_l\):

\[
h_l = (h'_{m_{l1}}, h'_{m_{l2}}, h'_{m_{l3}})' \tag{C.17}
\]

where \(m_{li}, i = 1, 2, 3\) indicate the three months of season \(l\). The full conditional distribution of \(h_l\) is

\[
[h_l|\text{rest}] \propto [h_l] \prod_{t=\tau_l}^{T_l} [a_{t,t}|b_{t,t}, H_l, C_l] \tag{C.18}
\]

We have \([a_{t,t}|b_{t,t}, H_l, C_l] = N(H_l b_{t,t}, C_l)\). We can write \(H_l b_{t,t}\) as a linear function of \(h_l\) by noting that

\[
H_l b_{l,t} = (H_{m_{l1}} H_{m_{l2}} H_{m_{l3}}) \begin{pmatrix} b_{m_{l1},t} \\ b_{m_{l2},t} \\ b_{m_{l3},t} \end{pmatrix} = \begin{pmatrix} H_{m_{l1},t} b_{m_{l1},t,1} + H_{m_{l2},t} b_{m_{l2},t,1} + H_{m_{l3},t} b_{m_{l3},t,1} \\ \\ \\ \\ \\ H_{m_{l1},K_X} b_{m_{l1},t,K_X} + H_{m_{l2},K_X} b_{m_{l2},t,K_X} + H_{m_{l3},K_X} b_{m_{l3},t,K_X} \end{pmatrix} = \begin{pmatrix} \text{diag}(b_{m_{l1},t}) \\ \text{diag}(b_{m_{l2},t}) \\ \text{diag}(b_{m_{l3},t}) \end{pmatrix} h_l \equiv J_{l,t} h_l \tag{C.19}
\]

where \(H_{m,k}\) is the \(k\)th diagonal element of \(H_m\) and \(J_{l,t}\) is the \((K_Y \times 3K_Y)\)-dimensional matrix defined above. We also have

\[
[a_{t,\tau_l:T_l}|b_{t,\tau_l:T_l}, H_l, C_l] = \prod_{t=\tau_l}^{T_l} [a_{t,t}|b_{t,t}, H_l, C_l] = N (J_l h_l, I_{T_l} \otimes C_l) . \tag{C.20}
\]
where $J_l$ is the $T_l K_Y \times 3K_Y$ matrix

$$J_l = \begin{pmatrix} J_{l,\tau_l} & \\
J_{l,\tau_l+1} & \\
\vdots & \\
J_{l,T_l} & \\ 
\end{pmatrix}. \tag{C.21}$$

The prior distribution for $h_l$ is

$$[h_l] = N(\mu_{1,l}, \Sigma_{1,l}) = N\left(\begin{pmatrix} \mu_{1,m_{l1}} \\
\mu_{1,m_{l2}} \\
\mu_{1,m_{l3}} \end{pmatrix}, \begin{pmatrix} \Sigma_{1,m_{l1}} & 0 & 0 \\
0 & \Sigma_{1,m_{l2}} & 0 \\
0 & 0 & \Sigma_{1,m_{l3}} \end{pmatrix}\right). \tag{C.22}$$

so using corollary A.1.2 we get that $[h_l|\text{rest}]$ is $N(A^{-1}d, A^{-1})$ with

$$A = \Sigma_{l,1}^{-1} + J'_l(I_{T_l} \otimes C_l^{-1})J_l \quad \text{and} \quad d = J'_l(I_{T_l} \otimes C_l^{-1})a_{l,T_l} + \Sigma_{l,1}^{-1} \mu_{l,1} \tag{C.23}$$

For ease of calculations note that

$$J'_l(I_{T_l} \otimes C_l^{-1})J_l = \sum_{t=\tau_l}^{T_l} J'_{l,t}C_l^{-1}J_{l,t} \quad \text{and} \quad J'_l(I_{T_l} \otimes C_l^{-1})a_{l,T_l} = \sum_{t=\tau_l}^{T_l} J'_{l,t}C_l^{-1}a_{l,t}. \tag{C.24}$$

**Transition matrices $B_m$**

The full conditional distribution of the $K_X^2$ dimensional vector $\text{vec}(B_m)$ is:

$$[\text{vec}(B_m)|\text{rest}] \propto [\text{vec}(B_m)] \prod_{t=\tau_m}^{T_m} [b_{m,t}|b_{m-1,t}, B_m, D_m]. \tag{C.25}$$

We have $[\text{vec}(B_m)] = N(\mu_{2,m}, \Sigma_{2,m})$ and $[b_{m,t}|b_{m-1,t}, B_m, D_m] = N(B_m b_{m-1,t}, D_m)$ and we note that

$$B_m b_{m-1,t} = (b'_{m-1,t} \otimes I_{K_X})\text{vec}(B_m) \equiv J_{m,t-1} \text{vec}(B_m) \tag{C.26}$$
where $J_{m,t-1}$ is the $(K_X \times K^2_X)$-dimensional matrix defined above.

\[
[\text{vec}(B_m)_{\text{rest}}] \propto [\text{vec}(B_m)] \prod_{t=\tau_m}^{T_m} [b_{m,t}|b_{m-1,t}, \text{vec}(B_m), D_m]
\]

\[
\propto \exp \left\{ -\frac{1}{2} \sum_{t=\tau_m}^{T_m} (b_{m,t} - J_{m,t-1} \text{vec}(B_m))' D^{-1}_m (b_{m,t} - J_{m,t-1} \text{vec}(B_m)) \\
+ (\text{vec}(B_m) - \mu_{2,m})' \Sigma_{2,m}^{-1} (\text{vec}(B_m) - \mu_{2,m}) \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left[ \text{vec}(B_m)' \left( \sum_{t=\tau_m}^{T_m} J'_{m,t-1} C^{-1}_m J_{m,t-1} + \Sigma_{m,2}^{-1} \right) \text{vec}(B_m) \\
- 2 \text{vec}(B_m)' \left( \sum_{t=\tau_m}^{T_m} J'_{m,t-1} D^{-1}_m b_{m,t} + \Sigma_{m,2}^{-1} \mu_{2,m} \right) \right] \right\} 
\]

(C.27)

Therefore, $[\text{vec}(B_m)_{\text{rest}}]$ is $N(A^{-1}d, A^{-1})$ with

\[
A = \sum_{t=\tau_m}^{T_m} J'_{m,t-1} D^{-1}_m J_{m,t-1} + \Sigma_{m,2}^{-1}
\]

\[
d = \sum_{t=\tau_m}^{T_m} J'_{m,t-1} D^{-1}_m b_{m,t} + \Sigma_{m,2}^{-1} \mu_{2,m}
\]

(C.28)

**Covariance Matrices of the data model ($r_l$ and $s_m$)**

Recall that $R_l^{-1} = \frac{1}{r_l} E_l$ and therefore

\[
[r_l|\text{rest}] \propto \prod_{t=\tau_l}^{T_l} [Y_{l,t}|a_{l,t}, r_l][r_l]
\]

\[
\propto \frac{1}{r_l^{T_l N_Y/2}} \exp \left\{ -\frac{1}{2 r_l} \sum_{t=\tau_l}^{T_l} (Y_{l,t} - U_l a_{l,t})' E_l (Y_{l,t} - U_l a_{l,t}) \right\} \frac{1}{r_l^{\alpha_1+1}} \exp \left\{ -\frac{\beta_1}{r_l} \right\}
\]

\[
\propto IG \left( T_l N_Y/2 + \alpha_1, \beta_1 + \frac{1}{2} \sum_{t=\tau_l}^{T_l} (Y_{l,t} - U_l a_{l,t})' E_l (Y_{l,t} - U_l a_{l,t}) \right). 
\]

(C.29)
Recall also that \( S_m^{-1} = \frac{1}{s_m} E_m \) and therefore

\[
[s_m | \text{rest}] \propto \prod_{t = \tau_m}^{T_m} [X_{m,t} | b_{m,t}, s_m] [s_m]
\]

\[
\propto \frac{1}{s_m^{T_m N X / 2}} \exp \left\{ -\frac{1}{2 s_m} \sum_{t = \tau_m}^{T_m} (X_{m,t} - V_m b_{m,t})' E_m (Y_{m,t} - V_m b_{m,t}) \right\} \times \frac{\beta_2}{s_m^{\alpha_2 + 1}} \exp \left\{ -\frac{\beta_2}{s_m} \right\}
\]

\[
\propto IG (T_m N X / 2 + \alpha_2, \beta_2 + \frac{1}{2} \sum_{t = \tau_m}^{T_m} (X_{m,t} - V_m b_{m,t})' E_m (X_{m,t} - V_m b_{m,t})) . \quad (C.30)
\]

Note that \( E_l \equiv \left( \frac{1}{c_l} I_{N_Y} - \tilde{U}_l \Lambda_l \tilde{U}_l' \right) \) and \( E_m \equiv \left( \frac{1}{c_m} I_{N_X} - \tilde{V}_m \Lambda_m \tilde{V}_m' \right) \) are constants.

Note that since \( U_l' \tilde{U}_l = 0, V_m' \tilde{V}_m = 0, U_l' U_l = I \) and \( V_m' V_m = I \) we get for each \( t \):

\[
(Y_{l,t} - U_l a_{l,t})' E_l (Y_{l,t} - U_l a_{l,t}) = (Y_{l,t} - U_l a_{l,t})' \left( \frac{1}{c_l} I_{N_Y} - \tilde{U}_l \Lambda_l \tilde{U}_l' \right) (Y_{l,t} - U_l a_{l,t})
\]

\[
= \frac{1}{c_l} Y_{l,t}' Y_{l,t} - \frac{2}{c_l} Y_{l,t}' U_l a_{l,t} - Y_{l,t}' \tilde{U}_l \Lambda_l \tilde{U}_l' Y_{l,t} + \frac{1}{c_l} a_{l,t}' a_{l,t} \quad (C.31)
\]

and

\[
(X_{m,t} - V_m b_{m,t})' E_m (X_{m,t} - V_m b_{m,t})
\]

\[
= \frac{1}{c_m} X_{m,t}' X_{m,t} - \frac{2}{c_m} X_{m,t}' V_m b_{m,t} - X_{m,t}' \tilde{V}_m \Lambda_m \tilde{V}_m' X_{m,t} + \frac{1}{c_m} b_{m,t}' b_{m,t} \quad (C.32)
\]

Note that terms that do not involve \( a_{l,t} \) or \( b_{m,t} \) are constants and can be stored before running the MCMC.

**Covariance Matrices of the process model** \((C_l \text{ and } D_m)\)

The prior distribution for both \( C_l \) and \( D_m \) is an inverse-Wishart distribution. The prior density function, for \( C_l \), is proportional to

\[
|C_l|^{-(\nu_Y + K_Y + 1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(W_l C_l^{-1}) \right\} \quad (C.33)
\]
We also note that for any sequence of \(K_Y\) dimensional vectors \(x_1, \ldots, x_T\)
\[
\sum_{t=1}^{T} x_t' C^{-1} x_t = \sum_{t=1}^{T} \text{tr}(x_t' C^{-1} x_t) = \sum_{t=1}^{T} \text{tr}(x_t x_t' C^{-1})
\]
\[
= \text{tr} \left( \left( \sum_{t=1}^{T} x_t x_t' \right) C^{-1} \right). \tag{C.34}
\]

The full conditional distribution of \(C_l\) is therefore
\[
[C_l | \text{rest} ] \propto \prod_{t=\tau_l}^{T_l} [a_{l,t} | H_l, b_{l,t}, C_l] [C_l]
\]
\[
\propto |C_l|^{(T_l-\tau_l+1)/2} \exp \left\{ -\frac{1}{2} \text{tr} \left( \left( \sum_{t=\tau_l}^{T_l} (a_{l,t} - H_l b_{l,t})(a_{l,t} - H_l b_{l,t})' \right) C_l^{-1} \right) \right\}
\]
\[
|C_l|^{-(\nu_Y + K_Y + 1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(W_l C_l^{-1}) \right\}
\]
\[
\propto IW \left( \nu_Y + T_l - \tau_l + 1, W_l + \sum_{t=\tau_l}^{T_l} (a_{l,t} - H_l b_{l,t})(a_{l,t} - H_l b_{l,t})' \right) \tag{C.35}
\]

Similarly we find that the full conditional distribution of \(D_m\) is
\[
[D_m | \text{rest} ] \propto \prod_{t=\tau_m}^{T_m} [b_{m,t} | B_m, b_{m-1,t}, D_m] [D_m]
\]
\[
\propto \exp \left\{ -\frac{1}{2} \text{tr} \left( \left( \sum_{t=\tau_m}^{T_m} (b_{m,t} - B_m b_{m-1,t})(b_{m,t} - B_m b_{m-1,t})' \right) D_m^{-1} \right) \right\}
\]
\[
|D_m|^{(T_m-\tau_m+1)/2} |D_m|^{-(\nu_X + K_X + 1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(W_m D_m^{-1}) \right\}
\]
\[
\propto IW \left( \nu_X + T_m - \tau_m + 1, W_m + \sum_{t=\tau_m}^{T_m} (b_{m,t} - B_m b_{m-1,t})(b_{m,t} - B_m b_{m-1,t})' \right) \tag{C.36}
\]

**C.2 Full conditional distributions for predictive model**

The following full conditional distributions are need for the model that includes the last year of ERA-40 data (\(X\)) but not the Polar MM5 data (\(Y\)). Besides sampling the additional full conditional distributions listed below, we note that the new \(a_{l,t}\)
and $b_{m,t}$ need to be accounted for in other distributions. Specifically, the summations in (C.28) and (C.28) need to incorporate the added time points. Also, the sum in (C.30) needs to account for the added ERA-40 data. However, since we do not include any new Polar MM5 data we don not need any changes to the distribution in (C.29). Finally, we need to update the $T_m$ constants and the summations in (C.35) and (C.36) to include the new set of amplitudes.

**Polar MM5 amplitudes $a_{l,t}$, without data**

The full conditional distribution of $a_{l,t}$ for timepoints where we do not have Polar MM5 data is simply the process model for $a_{l,t}$:

$$[a_{l,t}|\text{rest}] = [a_{l,t}|b_{l,t}, H_l, C_l] = N(H_l b_{l,t}, C_l) .$$  \hspace{1cm} (C.37)

**The last ERA amplitude $b_{m,t}$, without data**

The full conditional distribution of the last $b_{m,t}$ when there is no ERA-40 data at the last time point is

$$[b_{m,t}|\text{rest}] \propto [b_{m,t}|b_{m-1,t}, B_m, D_m] [a_{l,t}|b_{l,t}, H_l, C_l] \hspace{1cm} (C.38)$$

Using similar steps as before equation (C.11) we get that $[b_{m,t}|\text{rest}]$ is $N(A^{-1}d, A^{-1})$ with

\begin{align*}
A &= H'_m C_l^{-1} H_m + D_m^{-1} \\
\mathbf{d} &= H'_m C_l^{-1} (a_{l,t} - b) + D_m^{-1} B_m b_{m-1,t} \hspace{1cm} (C.39)
\end{align*}
Bibliography


ECMWF (2002), *The ERA-40 Archive*, European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom, on-line version available from the ECMWF ERA-40 Project Plan http://www.ecmwf.int/research/era/Project/Plan/.


