On Statistical Approaches to Climate Change Analysis

by

Terry Chun Kit Lee

B.Sc., Simon Fraser University, 2001
M.Sc., University of Victoria, 2003

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Mathematics and Statistics

On Statistical Approaches to Climate Change Analysis

by

Terry Chun Kit Lee

B.Sc., Simon Fraser University, 2001
M.Sc., University of Victoria, 2003

Supervisory Committee

Dr. M. Tsao, Supervisor
(Department of Mathematics and Statistics)

Dr. F.W. Zwiers, Supervisor
(Climate Research Division, Environment Canada/Adjunct, Department of Mathematics and Statistics)

Dr. W.J. Reed, Departmental Member
(Department of Mathematics and Statistics)

Dr. A.J. Weaver, Outside Member
(Department of Earth and Ocean Sciences)

Dr. D.W. Nychka, External Examiner
(Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research)

Abstract

Supervisory Committee

Dr. M. Tsao, Supervisor
(Department of Mathematics and Statistics)

Dr. F.W. Zwiers, Supervisor
(Climate Research Division, Environment Canada/Adjunct, Department of Mathematics and Statistics)

Dr. W.J. Reed, Departmental Member
(Department of Mathematics and Statistics)

Dr. A.J. Weaver, Outside Member
(Department of Earth and Ocean Sciences)

Dr. D.W. Nychka, External Examiner
(Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research)

Evidence for a human contribution to climatic changes during the past century is accumulating rapidly. Given the strength of the evidence, it seems natural to ask whether forcing projections can be used to forecast climate change. A Bayesian method for post-processing forced climate model simulations that produces probabilistic hindcasts of inter-decadal temperature changes on large spatial scales is proposed. Hindcasts produced for the last two decades of the 20^{th} century are shown to be skillful. The suggestion that skillful decadal forecasts can be produced on large regional scales by exploiting the response to anthropogenic forcing provides additional evidence that anthropogenic change in the composition of the atmosphere has influenced our climate. In the absence of large negative volcanic forcing on the climate system (which cannot presently be forecast), the global mean temperature for the decade 2000-2009 is predicted to lie above the 1970-1999 normal with probability 0.94. The global mean temperature anomaly for this decade relative to 1970-1999 is predicted to be 0.35°C (5-95% confidence range: 0.21°C−0.48°C).

Reconstruction of temperature variability of the past centuries using climate proxy data can also provide important information on the role of anthropogenic forcing in the observed 20^{th} century warming. A state-space model approach that allows incorporation of additional

non-temperature information, such as the estimated response to external forcing, to reconstruct historical temperature is proposed. An advantage of this approach is that it permits simultaneous reconstruction and detection analysis as well as future projection. A difficulty in using this approach is that estimation of several unknown state-space model parameters is required. To take advantage of the data structure in the reconstruction problem, the existing parameter estimation approach is modified, resulting in two new estimation approaches. The competing estimation approaches are compared based on theoretical grounds and through simulation studies. The two new estimation approaches generally perform better than the existing approach.

A number of studies have attempted to reconstruct hemispheric mean temperature for the past millennium from proxy climate indicators. Different statistical methods are used in these studies and it therefore seems natural to ask which method is more reliable. An empirical comparison between the different reconstruction methods is considered using both climate model data and real-world paleoclimate proxy data. The proposed state-space model approach and the RegEM method generally perform better than their competitors when reconstructing interannual variations in Northern Hemispheric mean surface air temperature. On the other hand, a variety of methods are seen to perform well when reconstructing decadal temperature variability. The similarity in performance provides evidence that the difference between many real-world reconstructions is more likely to be due to the choice of the proxy series, or the use of difference target seasons or latitudes, than to the choice of statistical method.

Contents

Supervisory Committee ii

Abstract iii

Contents v

List of Tables vii

List of Figures viii

Acknowlegements xi

1 Introduction 1

2 Decadal climate prediction skill from climate models 9

2.1 Bayesian forecasting model and forecast updating procedure 10

2.2 Climate change hindcast and skill evaluation 15

2.3 Additional hindcasts 16

2.4 Application 18

2.4.1 Covariance matrix estimation and dimension reduction 20

2.4.2 Determining the significance of the BSS 22

2.4.3 Hindcast results 23

2.5 Concluding remarks 31

2.6 Appendix: Derivation of posterior distribution 33

3 State-space model for historical temperature reconstruction 34

3.1 State-space model and the Kalman filter 36

3.1.1 Derivation of the Kalman filter and smoother 44

3.2 Maximum likelihood estimation for a state-space model with an unknown state process 50

3.2.1 The $\mathbb{PXY}$ approach 50

3.2.2 Numerical estimation procedure 53
3.3 Maximum likelihood estimation for a state-space model with a partially known state process 57
3.3.1 The $\mathbb{ALL}$ approach 57
3.3.2 The $\mathbb{CAL}$ approach 62
3.3.3 Proof of asymptotic distribution 66
3.4 Comparing the different estimators 92
3.4.1 Performance of estimator under the assumed state-space model 93
3.4.2 Robustness of estimators with respect to observational noise structure . 105
3.4.3 Robustness of estimators with respect to state equation specification . . 112
3.5 Concluding remarks 115
3.6 Appendix 117
3.6.1 Derivation of the EM estimation procedure with an unknown state process 117
3.6.2 Derivation of the EM estimation procedure with a partially known state process 119
3.6.3 Miscellaneous results 121

4 Comparing historical temperature reconstruction methods 126
4.1 A survey of existing reconstruction methods 126
4.1.1 Composite plus scale (CPS) approaches 127
4.1.2 Climate field reconstruction (CFR) approaches 129
4.2 Applying the state-space model for reconstruction 133
4.3 Comparison using climate model simulations 135
4.3.1 Analysis with red pseudo-proxy noise 150
4.4 Comparison using real-world paleoclimate proxy data 152
4.5 Concluding remarks 155

5 References 157

List of Tables

2.1 Summary of simulations used in the analysis of decadal predictability. 19

3.1 Standard error of the estimators obtained from simulations with 1000 runs with sample size of 200 or 998. 98

3.2 Comparison between the sample standard error from simulations with $N = 998$ and the theoretical asymptotic standard error for the three estimation approaches. 99

3.3 Percentage change in estimates sample variance in the red noise simulation relative to the white noise simulation with $N = 998$. 110

4.1 Estimated parameter values of the state-space model obtained with 100 pseudo-proxy series using two analysis periods. Boldfaced values are significant. 148

4.2 Estimated parameter values of the state-space model obtained with paleoclimate proxy data for three reconstruction periods. The 95% confidence interval is listed in brackets. Boldfaced values are significant. 154

List of Figures

1.1 Annual global mean surface temperature anomalies relative to the period 1961 to 1990 of the HadCRUT3v data set. 2

1.2 Reconstructions of northern hemispheric mean temperature of the last millennium using climate proxies. 8

2.1 Brier skill scores for the above normal event with 15 EOFs retained. 24

2.2 Global mean hindcast probabilities of above normal decadal mean temperatures in $30^o \times 40^o$ latitude-longitude regions and the observed proportion of regions with above normal temperatures together with its estimated uncertainty. 26

2.3 Hindcasts of global decadal mean surface temperature anomalies and their 5-95% confidence bounds 28

2.4 Mean and the 5-95 percentile of the posterior distribution of the scaling factor $\boldsymbol{\beta}_t$ from the full Bayesian hindcast with 15 EOFs retained. 30

3.1 Structure of the reconstruction problem. 35

3.2 Conditional independence structure of a state-space model. 36

3.3 Autocorrelation and partial autocorrelation function of a 1000 years northern hemispheric mean temperature series obtained from a control run of CCCma CGCM2 climate model. 39

3.4 Absolute bias of estimators obtained using the three different likelihood functions for three sample sizes. 96

3.5 Estimated efficiency of estimators obtained using the three different likelihood functions for three sample sizes. 97

3.6 Mean square error between the simulated state process and the Kalman filter estimated state process over 1000 simulation runs for the different estimation approaches 99

3.7 Normal probability plots for estimates obtained from the $\mathbb{PXY}$ approach with varying sample sizes. The p-values from the Shapiro-Wilk test are shown in the top left corner of each plot. Small p-value indicates evidence against the null hypothesis of normality. 102

3.8 Normal probability plots for estimates obtained from the $\mathbb{ALL}$ approach with varying sample sizes. The p-values from the Shapiro-Wilk test are shown in the top left corner of each plot. Small p-value indicates evidence against the null hypothesis of normality. 103

3.9 Normal probability plots for estimates obtained from the $\mathbb{CAL}$ approach with varying sample sizes. The p-values from the Shapiro-Wilk test are shown in the top left corner of each plot. Small p-value indicates evidence against the null hypothesis of normality. 104

3.10 Absolute bias of the estimators obtained using the three different likelihood functions for three sample sizes. The simulated data is generated with red observational noise. 106

3.11 Mean square error between the simulated state process and the Kalman filter estimated state process over 1000 simulation runs with red observational noise. 111

3.12 Absolute bias of the estimators obtained using the three different likelihood functions for three sample sizes. The simulated data is generated under an AR(2) model. 113

3.13 Mean square error between the simulated state process and the Kalman filter estimated state process over 1000 simulation runs with an AR(2) model. 114

4.1 Visualization of the testing procedures proposed by von Storch et al. (2004). . 136

4.2 Examples of reconstructed CSM northern hemisphere mean temperature series. 141

4.3 Relative root mean squared error (RRMSE) of the reconstruction error, expressed relative to the variability of the simulated hemispheric temperature. . . 144

4.4 Example of the difference in temperature anomalies between the ECHO-G simulation and the reconstructed ECHO-G series. 145

4.5 95% confidence bounds for the coefficients used to scale the EBM simulated responses to external forcing in the state-space model when the pseudo-proxy SNR is 0.5 and using the 1860-1970 calibration period. 146
4.6 Hindcasts of annual mean NH temperature based on the estimated state equation from 100 proxy series for two analysis periods. 148
4.7 Comparison of the reconstructed hemispheric mean temperature series obtained using the MBH method with non-detrended or detrended data at two signal-to-noise ratios. 149
4.8 Examples of reconstructed CSM northern hemisphere mean temperature series. Experiments were run using SNR=0.5 and the 1860-1970 calibration period with a) 15 and b) 100 red pseudo-proxy series. 151
4.9 Reconstructed series for the 30N-90N mean using the variance matching method and state-space model approach for real-world paleoclimate proxy data. 153
4.10 State equation hindcast of decadally smoothed 30N-90N mean temperature with real-world paleoclimate proxy data. 154

Acknowledgments

I would first like to acknowledge and thank my supervisors, Dr. Min Tsao and Dr. Francis Zwiers for their support, guidance and encouragement during my Master and Ph.D studies. Thank you Dr. Min Tsao, for your invaluable advice and support on both academic and personal matters. Thank you Dr. Francis Zwiers, for introducing me to the world of climate change science and for generously sharing time from your busy schedule to provide advice and help on my research.

Secondly, I would like to thank the members of the supervisory committee for their insightful comments and suggestions.

Finally, I would like to gratefully acknowledge the funding I have received from the Canadian Foundation for Climate and Atmospheric Science through the Canadian CLIVAR Research Network as well as that received from the Department of Mathematics and Statistics.

1 Introduction

Evidence for an human contribution to climatic changes during the past century is accumulating rapidly. The evolution of the climate system is influenced by its own internal variability and by external factors (called ***forcings***). The increase in concentration of anthropogenic greenhouse gases, such as CO_2 and CFCs, has been one of the main hypothesized external anthropogenic forcings that influence the evolution of the climate system. Indeed, measurement of atmospheric CO_2 concentration indicates that CO_2 concentration has been increased exponentially to 367 ppm in 1999 and to 379 ppm in 2005 (Sornerville et al. 2007), from about 280 ppm in the pre-industrial era (AD 1000-1750). Increase in greenhouse gases concentration will reduce Earth's efficiency to radiate heat back to the space, thereby intensifying the Earth's greenhouse effect. Not only that, many greenhouse gases tend to be long-lived and have a long-term effect on the climate system. There are also other anthropogenic forcings that influence the evolution of the climate system, such as aerosols. Aerosols are small airborne particles that result mainly from fossil fuel and biomass burning. They affect the climate by changing the energy balance of the Earth's atmosphere through the reflection of incoming solar radiation. Natural external forcings are also thought to play a role in influencing the climate system, such as those that arise from solar changes and explosive volcanic eruptions. In addition to the cyclic 11-year solar cycle, the sun's energy output has increased gradually in the industrial era (Sornerville et al. 2007). Solar energy directly heats the climate system and thus variation in solar output will thereby change the climate system. Explosive volcanic activity can inject large amounts of short-lived (2-3 years) aerosols into the stratosphere. These aerosols have been shown to have a cooling effect on the climate system. Natural forcing on its own probably would have cooled the climate system during the latter half of the 20^{th} century (IPCC 2007).

Figure 1.1 shows the observed annual global mean surface temperature anomalies relative to the period 1961 to 1990 of the HadCRUT3v data set (Brohan et al. 2006). In practice, temperatures are generally expressed as departure from some baseline value, such as the average of the whole or part of the data. The baseline value is termed the ***base climatology***, and

anomalies refers to values obtained by subtracting the base climatology from the observed values. For this data set, warming since 1979 has been estimated to be 0.163 ± 0.046°C per decade for the globe, but 0.234 ± 0.070°C and 0.092 ± 0.038°C for northern hemisphere and southern hemisphere respectively (Brohan et al. 2006). The question as to how much of these observed changes are due to anthropogenic causes has been a topic of increasing interest. A large body of research suggests that the evolution of the climate system since the industrial revolution is unlikely to be the result of just natural internal variation of the climate system, but rather that a combination of natural and anthropogenic forcings have had a considerable influence (see, e.g., the reviews of Hegerl et al., 2007b and IDAG 2005).

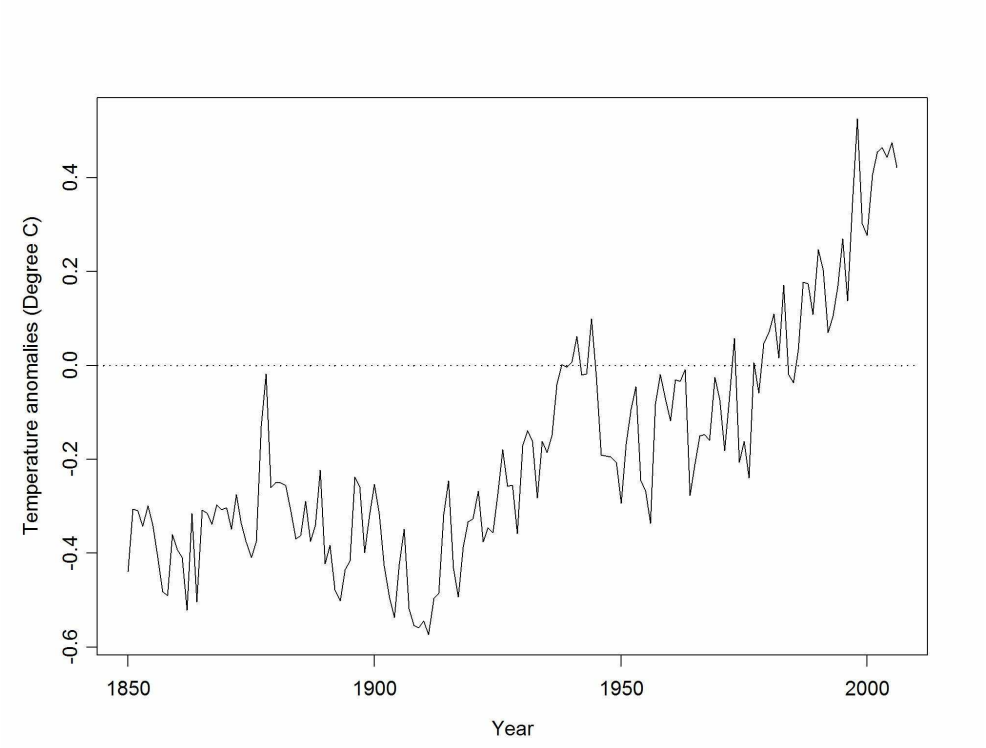

Figure 1.1: Annual global mean surface temperature anomalies relative to the period 1961 to 1990 of the HadCRUT3v data set.

Detection and attribution are two key concepts in the analysis of climate change. ***Detection*** is the process of demonstrating that an observed change is significantly different from what one can expect with internal climate variability alone. However, the detection of a change

in the climate does not necessarily imply that its causes are well understood. ***Attribution*** refers to the process of establishing cause and effect relationships between the observed change and forcings. This involves statistical analysis and the assessment of multiple lines of evidence to show that the observed changes are consistent with the anticipated responses to a combination of external forcings, and at the same time, inconsistent with alternative, physically plausible explanations.

Ideally, detection and attribution studies would require controlled experimentation with the Earth's climate system. However, with no spare Earth to carry out such experiments, the use of climate models is required. A ***climate model*** is a numerical representation of the Earth's climate system. Based on the physical, chemical and biological properties of the Earth's components, climate models have been developed to predict a number of environmental factors, for example, the temperature at the Earth's surface, circulation of ocean and wind currents and the development of cloud cover. Different models are developed to represent the climate system with varying complexity. Two types of models at opposite ends of the complexity scale are utilized in this thesis, namely the coupled atmosphere-ocean general circulation models (CGCM) and the simple energy balanced models (EBM). CGCMs are the most complex climate model currently available. In general, this type of model consists of two three-dimensional sub-models coupled together, with one modeling the atmosphere and land and the other one modeling the dynamics of the ocean and sea ice. Some CGCMs also include representation of important biogeochemical cycles, such as the global carbon cycle. A CGCM is computationally expensive to run due to its complexity. In contrast, energy balanced models are much simpler models which simulate only surface temperature by modeling the incoming and outgoing radiation budget of the Earth. Unlike CGCMs, an EBM produces estimates of surface temperature change that are free of internal climate variability. EBMs are computationally inexpensive to run but may omit some of the important feedback processes that exist in more complex models. Thus, most detection and attribution studies utilize data obtained from CGCM simulations.

Two types of simulations from the climate models are of interest in climate change analysis. Controlled simulations, also called ***control runs***, are obtained by fixing the external forcing

factors at some specific level, usually the pre-industrial level. This type of simulation therefore simulates only internal variability of the climate system in the case of CGCM. In contrast, forced simulations, also called ***forced runs***, are obtained by changing the level of the forcing factors over time. This type of simulation allows one to investigate the climate system's response to changes in forcing. Since runs from CGCMs also simulate natural internal climate variability, the average of a group of parallel simulations (called ***ensemble averages***) is usually used in detection and attribution analysis to reduce the impact of such variability.

A widely used statistical tool for detection and attribution is optimal fingerprinting (see, for example, Hasselmann 1979, 1997; Allen and Tett 1999). ***Optimal fingerprinting***, in principle, is just linear least square regression, in which one would estimate the scaling factors that are needed to best fit the climate model output to the observed data. The problem can be cast with the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is a vector containing the observed data in space and time, and $\mathbf{X} = (\mathbf{x}_1|\cdots|\mathbf{x}_m)$ is a matrix composed of m model simulated response (signal) patterns to external forcings that are under investigation. Each $\mathbf{x}_i$ is organized in the same manner in space and time as the observed data. The vector $\boldsymbol{\varepsilon}$ is a random vector that represents natural internal climate variability which is assumed to be normally distributed with covariance matrix $\mathbf{C}$. Vector $\boldsymbol{\beta}$ consists of m scaling factors that adjust the amplitudes of the signal patterns and is estimated by least square regression where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}$. The matrix $\mathbf{C}$ is unknown and is usually estimated from a long control run of a climate model. Once the vector $\boldsymbol{\beta}$ is estimated, a residual consistency test is conducted to detect model inadequacy that may arise from a dimension reduction step that is required to overcome the non-invertibility of the estimate of $\mathbf{C}$. Such non-invertibility arise because the number of independent vectors that can be extracted from a control run to estimate $\mathbf{C}$ is less than the length of $\mathbf{y}$ (see Allen and Scott (2003) for further detail).

In practice, the fields in $\mathbf{y}$ usually contain observed surface temperature anomalies at different latitude and longitude grids of the globe over a 30-100 year time scale, arrayed in

space and time. If the grid size is 5°(latitude)×5°(latitude), there will be $36 \times 72 = 2592$ spatial grid points for the entire globe at a each time point. However, many grid values are missing in the early years due to the fact that no measurements of temperature were recorded for that grid box at those times. The fields in $\mathbf{y}$ are therefore truncated so that they represent only those grid boxes with adequate data (this process is referred to as ***masking***). Depending on the researchers and the type of studies, the data might be aggregated into different grid box sizes (e.g., $30° \times 40°$) and decadal or annual values might be used. Other than surface air temperature, other variables, such as precipitation and mean sea level pressure, have also been used in detection and attribution studies (see, for example, the review of Hegerl et al. 2007b and references therein; see also Zhang et al. 2007).

Detection and attribution questions are assessed by testing specific hypotheses on $\boldsymbol{\beta}$. The detection of a postulated climate change signal occurs when its amplitude in observations is shown to be significantly different from zero, thus, implying the patterns in the observation cannot be explained by natural internal climate variability alone. This is handled by testing the null hypothesis

$$H_D : \ \boldsymbol{\beta} = \vec{0}$$

where $\vec{0}$ is a vector of zeros. Rejection of H_D leads to detection at a specified level of significance.

The requirements for making an attribution claim are detection, elimination of other plausible causes, and evidence that the observed change is consistent with the climate model estimated response to external forcing, i.e., $\boldsymbol{\beta} = \vec{1}$ where $\vec{1}$ is a vector of units. Evidence in support of attribution is thus obtained by the ***attribution consistency test***, which test the null hypothesis

$$H_A : \ \boldsymbol{\beta} = \vec{1}.$$

Formally, consistency between the observed and climate model simulated response to forcing can be claimed when H_A cannot be rejected. However, a failure to reject indicates only a lack of evidence against H_A. It does not constitute evidence for the hypothesis, which is what is really needed to support an attribution assessment. Hence, a more complete attribution

assessment would require the same test on the scaling factors $\boldsymbol{\beta}$ that are obtained using other physically plausible combinations of simulated signal patterns.

Bayesian approaches to address the questions of detection and attribution have also been consider in the literature, for example, Berliner et al. (2000), Lee (2003), Min et al. (2004), Lee et al. (2005) and Schnur and Hasselmann (2005). The approach of Berliner et al. (2000), Lee (2003) and Lee et al. (2005) is similar to the optimal fingerprinting approach. Instead of the classical hypothesis testing procedure, inferences on the scaling factor $\boldsymbol{\beta}$ is approached through the posterior distribution of $\boldsymbol{\beta}$. The posterior distribution is obtained based on the observed data, climate model simulated signal patterns and one's prior knowledge on $\boldsymbol{\beta}$. Detection and attribution assessment are made by calculating the posterior probabilities that $\boldsymbol{\beta}$ lies in a predefined detection ($\mathfrak{D}$) and attribution ($\mathfrak{A}$) regions, respectively. In Lee (2003) and Lee et al. (2005), the detection region $\mathfrak{D}$ is defined as $[0.1, \infty)$ and detection is claimed when the posterior probability that $\boldsymbol{\beta} \in \mathfrak{D}$ is large. The attribution region $\mathfrak{A}$ is defined as (0.8,1.2) and a large posterior probability that $\boldsymbol{\beta} \in \mathfrak{A}$ provides evidence in support of attribution. Other methods have also been used in detection and attribution analysis, such as centered and uncentered pattern correlation statistics (e.g, Santer et al. 1995). Readers are referred to the authoritative review of Mitchell et al. (2001) and Hegerl et al. (2007b) for results and further discussion on detection and attribution analysis. See also IDAG (2005).

The climate change detection and attribution problem is investigated from a different angle in Chapter 2. Simulations of the 20^{th} century from climate models can in fact be interpreted as predictions of the past climate (called ***hindcasts***), provided that these simulations are driven with realistic estimate of historical changes in external radiative forcing. A simple Bayesian method is proposed for post-processing such simulations to produce probabilistic hindcasts of inter-decadal temperature changes on large spatial scales. The skill of these probabilistic hindcasts are subsequently assessed to provide answers to the detection problem, where skill is defined as a statistical evaluation of the accuracy of hindcast. Such skill, if present, would provide evidence that changes in the composition of the atmosphere from external forcings have influenced the climate. The work presented in Chapter 2 is from Lee et al. (2006).

Another topic that is considered in this thesis relates to the study of paleoclimate. ***Paleoclimate*** is the climate during periods prior to the development of measuring instruments. The late 20^{th} century warming trend that is exhibited in Figure 1.1 is clearly unusual over the past 150 years. However, there still remains the question as to how significant is this warming trend compared to climate variability prior to the industrial era. Gaining knowledge about the behaviour of the climate system on multi-centennial or longer timescales provides an answer to this question. Even though the thermometer was invented in the early 1600s, systematic records of temperature were not widely kept until the latter half of the 19^{th} century. To better understand the climate system prior to the period when systematic instrumental records are available, one therefore has to rely on climate proxy data. A ***climate proxy*** is a local record that is able to capture climate variability of the past and is interpreted as a climate variable, such as, temperature or precipitation, using a transfer function and recently observed relations between the proxy and the climate variable under investigation. Examples of climate proxies are thickness of tree rings, pollen of different species and ice cores. Some currently available proxy data goes back more than a thousand years, but with decreasing availability for earlier periods.

Recently, there has been considerable interest in using networks of climate proxies to reconstruct the northern hemisphere (NH) mean temperature of the last millennium. A wide variety of techniques have been used and Figure 1.2 displays reconstructed NH mean temperature from some recent studies. These reconstructions suggest that the past 50 years are indeed the warmest in at least the last 1000 years. However, there are considerable differences between these reconstructed series. Such discrepancy can be due to the use of different statistical methods, the choice of proxies, the quality of the proxy records, the target season or latitude band or other reasons. Different statistical methods are employed in these reconstructions and it therefore seems natural to ask how much of this discrepancy is caused by the variations in methods. The reliability of some of the reconstruction methods has been looked at in different studies (e.g., Zortia et al. 2003; von Storch et al. 2004; Bürger and Cubasch 2005; Esper et al. 2005; Mann et al. 2005; Zorita and von Storch 2005; Bürger et al. 2006; Juckes et al. 2006; Mann et al., 2007). However, a comprehensive comparison between a range of existing

methods has not yet been attempted. In Chapter 4, an empirical comparison between different reconstruction methods is provided. Analyses are carried out using both climate model data and real-world paleoclimate proxy data. A new method for reconstruction that is based on a state-space time series model and Kalman filter algorithm will also be proposed. This approach allows one to simultaneously reconstruct the unknown temperature and conduct a detection assessment of the importance of the response to forcings on historical temperature. To better suit the temperature reconstruction problem, the existing statistical methods used in state-space modelling are modified. Details of the modifications are given in Chapter 3. The new approach will also be compared to the existing methods in Chapter 4. The work presented in Chapter 4 is from Lee et al. (2008).

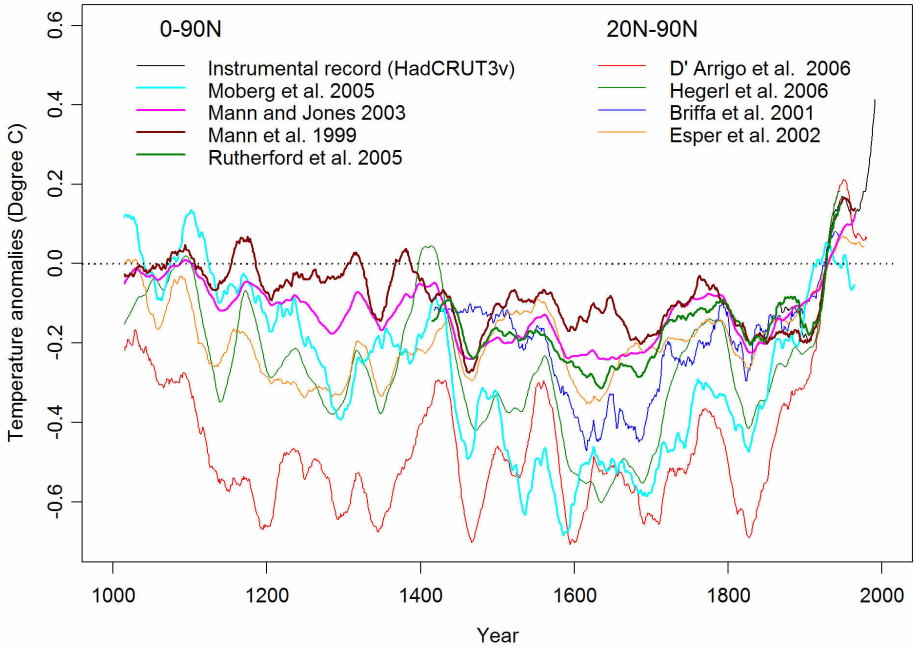

Figure 1.2: Reconstructions of northern hemispheric mean temperature of the last millennium using climate proxies. 31-yr moving averages are shown. The reconstructions represent either the mean of the northern hemisphere (0-90N) or the region 20N-90N as indicated in the legend. The instrumental NH mean from HadCRUT3v is shown in black. Reconstructions are shown as coloured line. All series are expressed as anomalies relative to the 1900-1960 period. Units (°C). Details of each series can be found in the corresponding reference and in the review by Jansen et al. (2007).

2 Decadal climate prediction skill from climate models

There is a large body of evidence that changes in the external radiative forcing of the climate system have had a substantial impact on its evolution since the industrial revolution. These forcing changes have been caused by the changing composition of the atmosphere, mainly as a result of anthropogenic emissions of greenhouse gases and aerosol precursors from fossil fuel burning, and secondarily by natural decadal scale variations in volcanic and solar forcing. Evidence of the effects of these forcing changes on the climate system has been detected in surface air temperature at global scales (Mitchell et al., 2001; IDAG, 2005; Hegerl et al. 2007b) and recently also at continental and subcontinental scales (Zwiers and Zhang 2003; Stott 2003; Barganza et al. 2004; Gillett et al. 2004a; Zhang et al. 2005; Karoly and Wu 2005; Hegerl et al. 2007b). A consistent picture of change that is related to the change in external forcing is also emerging in several other aspects of the climate system, such as ocean heat content, snow and sea-ice cover extent, growing season length, tropopause height, precipitation and mean sea level pressure (see, for example, IDAG 2005, Hegerl et al. 2007b, and references therein; see also Zhang et al. 2007).

Given the strength of the evidence, it seems natural to ask whether forcing projections can be used to forecast large scale climate change on decadal time scales. Given their potential applications, many of which would involve some type of hedging, it is desirable that any decadal scale forecast should be probabilistic rather than deterministic. One approach for obtaining such forecasts would be to produce large ensembles of forced climate simulations which would then be interpreted in a probabilistic manner, either directly or after some type of post-processing to adjust for model biases. Unfortunately, such an approach would be expensive to implement given the need for large ensembles and the complexity of climate system models that have been used to study the evolution of the climate of the 20th century. However, recent methodological developments, notably the application of Bayesian techniques to climate change detection (Berliner et al. 2000; Min et al. 2004; Schnur and Hasselmann 2005; Lee et al. 2005), provide means by which this possibility can be evaluated using small

ensembles of simulations.

Forecasts of decadal climate change have several potential sources of skill. These include the initial ocean state (e.g., Boer 2004), possibly the initial land surface state, the response to changes in external forcing conditions during the forecast period, and the continued adjustment of the climate during the forecast period towards a new equilibrium consistent with previous changes in forcing that continue to persist, such as those that result from previous changes in atmospheric composition. Climate simulations of the 20th century with specified historical changes in radiative forcing can be thought of as climate hindcasts that attempt to exploit the latter two sources of skill. The term hindcast refers to a prediction statement about the past that is made using only information, such as initial conditions, that was available prior to the prediction period, while forecast corresponds to a statement about the future. The extension of such simulations into the 21st century using scenarios of future emission change can further be considered as forecasts of future climate change, at least for relatively short 1-2 decade periods into the future (e.g., Zwiers, 2002) because at these forecast leads, the climate response appears not to be very sensitive to the details of the future forcing change scenario that must be specified.

The remainder of this chapter explains the technique and describes results obtained when an ensemble of simulations of the 20th century using only the history of anthropogenic forcing change are evaluated as climate hindcasts on decadal time scales. A Bayesian method is proposed for making decadal hindcasts of temperature change. The hindcasts are then verify with the observed values to assess the skill of the hindcast. Description of the forecasting model and the techniques used will be given in section 2.1 to 2.3. Data analysis results will be presented in section 2.4. Concluding remarks are given in section 2.5. Statistical derivations are given in section 2.6.

2.1 Bayesian forecasting model and forecast updating procedure

In order to describe the statistical forecasting model, some notation and assumptions are first introduced. Let $\mathbf{O}_t$ be a $n \times 1$ observed decadal mean temperature vector containing n spatial grid points for a decade t, such as that of the 1970's. One can then think of the departure of

$\mathbf{O}_t$ from some base climatology $\overline{\mathbf{O}}_t$ as being the sum of the response to external forcing during decade t relative to the base period, plus the effects of internal natural variability. With these assumptions, a reasonable statistical model for $\mathbf{Y}_t = \mathbf{O}_t - \overline{\mathbf{O}}_t$ is,

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \tag{2.1.1}$$

where vector $\mathbf{X}_t = \mathbf{S}_t - \overline{\mathbf{S}}_t$ contains the model simulated response in decade t to past external forcing change ($\mathbf{S}_t$) relative to the climatological base period ($\overline{\mathbf{S}}_t$), $\boldsymbol{\varepsilon}_t$ is random noise results from internal variability, and $\boldsymbol{\beta}_t$ is a scaling factor that accounts for errors in the magnitude of the simulated response to the specified external forcing changes. Such a model can be used to make a decadal climate forecast by making a suitable choice of a climatological base period. Following the standard World Meteorological Organization convention for defining the current mean climate, the base period for decade t was chosen to be the previous three decades $t-10$, $t-20$ and $t-30$. This choice is made because it is an approach often used in seasonal forecasting (see, for example, Derome et al. 2001).

Decades are indicated by their first year. Thus for the 1970's, the observed anomaly field is computed as

$$\mathbf{Y}_{1970} = \mathbf{O}_{1970} - (\mathbf{O}_{1960} + \mathbf{O}_{1950} + \mathbf{O}_{1940})/3$$

and the corresponding model simulated field of response anomalies is given by

$$\mathbf{X}_{1970} = \mathbf{S}_{1970} - (\mathbf{S}_{1960} + \mathbf{S}_{1950} + \mathbf{S}_{1940})/3$$

where, in order to reduce the effects of internal variability on the simulated response pattern, $\mathbf{S}_t$ is in fact the mean of an ensemble of simulations of the 20th century, each using the same forcing prescription. With such a convention and given an appropriate scaling factor estimate $\hat{\beta}_t$, a point-value hindcast (or forecast depending upon the choice of t) can be made for decade t by calculating

$$\hat{\mathbf{O}}_t = \overline{\mathbf{O}}_t + \hat{\mathbf{X}}_t \hat{\beta}_t + \hat{\varepsilon}_t \tag{2.1.2}$$

where $\hat{\mathbf{X}}_t = \mathbf{X}_t$ is the model forecast anomaly response and $\hat{\varepsilon}_t = \underline{0}$ is the forecast of the internal

variability in decade t. The third term, $\hat{\boldsymbol{\varepsilon}}_t$, is zero because it is assumed that internal variability is random and unpredictable on decadal time scales. Even though internal variability might itself be predictable on decadal time scales as an initial value problem, with skill obtaining from ocean and perhaps land surface initial conditions (e.g., Grotzner et al. 1999; Collins and Allen 2002; Pohlmann et al. 2004), but for the purposes of this study, it is assumed that there is no climate predictability from internal sources on decadal time scales. This is likely to be a suboptimal assumption (e.g., Boer 2004; Pohlmann et al. 2004).

Equation (2.1.2) constitutes a valid hindcast in the sense that it depends only on the evolution of the forcing specified to the climate model prior to and during the forecast decade t. This evolution is known from historical observations, albeit with uncertainty (Ramaswamy et al. 2001) to the end of the 20th century, and subsequently can be specified from a forcing scenario, such as one of the scenarios in the IPCC Special Report on Emissions Scenarios (Nakicenovic et al. 2000). The point-value forecasts obtained in this way can be extended to probabilistic forecasts by noting (i) that the simple forecasting model described above is, in fact, the same statistical model that is used in optimal fingerprinting; and (ii) that this model can be given a Bayesian interpretation if one considers the parameter $\boldsymbol{\beta}_t$ to be random variable that has its own statistical distributions (e.g., Lee et al. 2005).

Following Berliner et al. (2000) and Lee et al. (2005), $\boldsymbol{\beta}_t$ will assume to have a normal probability density function $\phi(m_t, c_t)$ with mean m_t and variance c_t. The value of m_t and c_t will be chosen according to one's subjective knowledge on $\boldsymbol{\beta}_t$ and information from past observations. Further, $\boldsymbol{\varepsilon}_t$ is assumed to be independent of $\boldsymbol{\beta}_t$ and that $\boldsymbol{\varepsilon}_t$ has a multivariate Gaussian probability density function $\phi_n(\mathbf{0}, \boldsymbol{\Sigma})$ of dimension n. With these assumptions, it follows from the forecast model (2.1.1) that the n dimensional hindcast distribution of $\mathbf{Y}_t$, conditional on the model simulated response anomaly $\mathbf{X}_t$, is given by

$$f(\mathbf{Y}_t \mid \mathbf{X}_t) = \phi_n(m_t \mathbf{X}_t, c_t \mathbf{X}_t \mathbf{X}_t^{\mathrm{T}} + \boldsymbol{\Sigma}). \tag{2.1.3}$$

As discussed in Lee et al. (2005), the distribution on $\boldsymbol{\beta}_t$ can be chosen to reflect prior knowledge about the presence of the simulated response anomaly in the observations, the uncertainty of

the response and other sources of uncertainty. For instance, m_t can be chosen to be zero if the simulated response anomaly is believed to be absent in the observations. The width of the distribution of $\boldsymbol{\beta}_t$ might also be used to reflect the uncertainty in the pattern of response, in the sense that a biased response pattern may require a $\boldsymbol{\beta}_t$ value different from unity to make the best fit with the observations.

Having constructed this multivariate distribution, the hindcast distribution for grid i is easily obtained by noting that the marginal distribution of a multivariate normal distribution is also normal. Hence the distribution for $\mathbf{y}_{ti}$, for $i = 1,\ 2, \cdots,\ n$ is given by

$$f(\mathbf{y}_{ti} \mid \mathbf{X}_t) = \phi(m_t \mathbf{x}_{ti}, c_t \mathbf{x}_{ti}^2 + \boldsymbol{\Sigma}_{ii}). \tag{2.1.4}$$

Point and interval forecasts for point i (i.e., confidence intervals for the forecast temperature change in decade t relative to the current 3-decade climatology) can then be defined as the mean and desired percentiles of this distribution. One can also obtain a probability hindcast for a particular event $\mathbf{E}$ for grid i by integrating $f(\mathbf{y}_{ti} \mid \mathbf{X}_t)$ over the hindcast event of interest. That is, the hindcast probability of event $\mathbf{E}$ at grid i can be obtained by computing,

$$\int_{\mathbf{E}} f(\mathbf{y}_{ti} \mid \mathbf{X}_t) d\mathbf{y}_{ti} \tag{2.1.5}$$

Typically, one would take $\mathbf{E}$ to be an event such as the occurrence of an above "normal" decadal mean temperature (i.e., $\mathbf{y}_{ti} > 0$) where "normal" is defined as the operational 3-decade climatological mean that is current for the decade for which the hindcast is issued.

Hindcast distributions for other quantities such as the global mean $\mathbf{y}_{t.} = w_t^{\mathrm{T}} \mathbf{Y}_t$ are easily obtained as

$$f(\mathbf{y}_{t.} \mid \mathbf{X}_t) = \phi(m_t w_t^{\mathrm{T}} \mathbf{X}_t, c_t w_t^{\mathrm{T}} \mathbf{X}_t \mathbf{X}_t^{\mathrm{T}} w_t + w_t^{\mathrm{T}} \boldsymbol{\Sigma} w_t) \tag{2.1.6}$$

where, in the case of the global mean, w_t is a vector of weights proportional to area. In general, the weights are the cosine of the central latitude of the grid points in $\mathbf{Y}_t$. Note that the dot in place of a subscript indicates that weighted averaging has been performed over that subscript.

Once a hindcast has been produced for an initial decade t and verified against observations

for that period, hindcasts (or forecasts as the case may be) can be produced for subsequent decades by means of a simple Bayesian updating procedure. In particular, the hindcast for decade $t+10$ is obtained by deriving a probability distribution for $\boldsymbol{\beta}_{t+10}$ that is based on the outcome from decade t. The analysis that produced the forecast for decade t was based on observations prior to decade t, a simulation of the response to external forcing on the climate system through to the end of decade t, and a prior distribution $\phi(m_t, c_t)$ on the scaling factor $\boldsymbol{\beta}_t$ that reflects knowledge about the true value of $\boldsymbol{\beta}_t$ before observing $\mathbf{Y}_t$. Once the observations for the initial decade t become available, knowledge regarding the scaling factor can be updated by calculating the posterior distribution on $\boldsymbol{\beta}_t$. According to the Bayes' theorem, this distribution is given by

$$f(\boldsymbol{\beta}_t \mid \mathbf{X}_t, \mathbf{Y}_t) = \phi(m_{t+10}, c_{t+10}) \tag{2.1.7}$$

where $c_{t+10} = (1/c_t + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_t)^{-1}$ and $m_{t+10} = c_{t+10}(m_t/c_t + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}_t)$. Derivation of the posterior distribution is given in section 2.6. This updated version of the distribution on $\boldsymbol{\beta}_t$, which represents a combination of the prior information that was available regarding $\boldsymbol{\beta}_t$ and information that was subsequently extracted from the observations for decade t, can now be used as the prior distribution for $\boldsymbol{\beta}_{t+10}$ to generate the hindcast for decade $t+10$. In another words, the hindcast for decade $t+10$ is derived by using the posterior distribution on $\boldsymbol{\beta}_t$ as the prior distribution for $\boldsymbol{\beta}_{t+10}$ to obtain the hindcast distribution $f(\mathbf{Y}_{t+10} \mid \mathbf{X}_{t+10})$ for the climate state in decade $t+10$. Once observations become available for decade $t+10$, a further updated posterior distribution $f(\boldsymbol{\beta}_{t+10} \mid \mathbf{X}_{t+10}, \mathbf{Y}_{t+10})$ can then be calculated for making the hindcast in decade $t+20$, etc. Thus such a posterior-prior updating process allows us to improve our knowledge, over time, regarding the scaling $\boldsymbol{\beta}$ that is required to best match the model simulated decadal temperature increments with observed temperature increments. For convenience, the posterior-prior updating process is summarized on the next page.

The Bayesian forecasting process

Assume the distribution function of $\boldsymbol{\beta}_t$ is $f(\boldsymbol{\beta}_t) = \phi(m_t, c_t)$, the forecast distribution of $\mathbf{Y}_s$ and the posterior distribution of $\boldsymbol{\beta}_s$, for $s = t,\ t+10,\ t+20,\ \ldots$, is given recursively by,

$$f(\mathbf{Y}_s \mid \mathbf{X}_s) = \phi_n(m_s\mathbf{X}_s, c_s\mathbf{X}_s\mathbf{X}_s^{\mathrm{T}} + \boldsymbol{\Sigma})$$
$$f(\boldsymbol{\beta}_s \mid \mathbf{X}_s, \mathbf{Y}_s) = \phi(m_{s+10}, c_{s+10}) = f(\boldsymbol{\beta}_{s+10})$$

where

$$c_{s+10} = (1/c_s + \mathbf{X}_s^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_s)^{-1}$$
$$m_{s+10} = c_{s+10}(m_s/c_s + \mathbf{X}_s^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}_s).$$

□

2.2 Climate change hindcast and skill evaluation

By defining the forecasting problem in terms of anomalies from a moving climatological base period, one can easily define hindcast events that are related to climate change. Here, only a two category forecast system is considered, that is, either an increase in temperature in decade t relative to the base period (above "normal") or conversely, a decrease (below "normal"). Thus the hindcast probabilities at point i is calculated by defining the event $\mathbf{E}$ in (2.1.5) to be either [0,∞) for above normal, or ($-\infty$,0) for below normal. If one predicts that there will be no climate change (relative to the base period), the probabilities for both events should be equal to 1/2. Otherwise, the probabilities would differ from 1/2, depending on the strength and sign of the effect of the forcing. In contrast, seasonal forecasting systems (e.g., O'Lenic, 1994; Mason et al., 1999; Derome et al., 2001) typically provide three category forecasts of the likelihood of above, near and below normal. Extending the present two category system to three categories would not be difficult, but would increase uncertainty somewhat because the event boundaries, as well as the forecasts themselves, would then become dependent upon our estimates of the internal climate variability. Using a two category system avoids this source of uncertainty by allowing one to define events relative only to the current operational base climatology.

Once the probability hindcasts are generated, one can evaluate the skill of these hindcast

for each decade t over the n spatial grid points contained in the observation vector $\mathbf{Y}_t$. Such an evaluation allows us to assess whether knowledge of forcing change during the decades leading up to decade t can be translated into usable forecast skill on decadal timescales. Such skill, if present, would also provide additional, and very practical, supporting evidence for the attribution of observed 20th century climate change to external forcing change.

A widely used verification statistic for probability forecasts is the Brier score (Brier 1950; see also Wilks 1995). The Brier score, evaluated over n forecasts, is given by

$$B = \frac{1}{n}\sum_{i=1}^{n}(p_i - q_i)^2$$

where p_i is the forecast probability of an event $\mathbf{E}$ at grid i and q_i is an indicator variable that is set to 1 or 0 depending upon whether or not the event occurred in the observation. A forecasting system cannot be considered to be useful if the same score value can be obtained by means of a forecast that is easier to obtain than the forecast under consideration. In the case of Brier Score, to assess the skill of a forecast, it is conventional to look at the Brier skill score (BSS), which is defined as

$$BSS = 1 - B/B_{cli}$$

where B_{cli} is the climatologically expected Brier score of the event, that is, the Brier score for a forecast which always predicts the event $\mathbf{E}$ to occur with its true occurrence probability P_E. Under the assumption of no climate change and with event $\mathbf{E}$ defined as $[0,\infty)$ or $(-\infty,0)$, $P_E = 0.5$ and thus $B_{cli} = P_E(1 - P_E) = 0.25$. The Brier skill score equals to 1 for a perfect probability forecast, 0 for a forecast that performs the same as the climatological forecast and negative for a forecast that performs worse than the climatological forecast.

2.3 Additional hindcasts

In order to investigate whether the Bayesian procedure described above improves or diminishes forecast skill from the models, or whether indeed, the climate models contribute skill beyond a simple straw man approach to forecasting, three additional hindcast variants are considered.

The first, called the raw model hindcast, is produced by not updating the mean of the

prior distribution in the posterior-prior updating process. That is, $m_t = 1$ at each time point t so that the mean of the hindcast distribution is the ensemble mean. However, the width of the prior is still allowed to vary between time periods according to (2.1.7). The effect is that the variance of the the forecast distribution (2.1.3) is inflated by the factor $c_t \mathbf{X}_t \mathbf{X}_t^{\mathrm{T}}$. This is roughly equivalent to the usual practice of adding a factor $1/m$ to the forecast variance, where m is the ensemble size, to account for sampling variability in the ensemble mean (Allen and Tett 1999). Note the details of the treatment of the prior variance have almost no influence on the results.

The second, called the blended hindcast, uses a prior distribution where the mean $\lambda \times m_t + (1-\lambda) \times 1$ is a blend of the posterior mean obtained from the updating process at time $t-10$, and the mean $m_t = 1$ that would be appropriate if the model always responded correctly to external forcing. The variance of the prior distribution continues to be updated as in (2.1.7). The weighting factor λ can be varied between 1 and 0, with $\lambda = 1$ producing the Bayesian hindcast described previously, and $\lambda = 0$ producing the raw model hindcast described above. A λ value of 0.5 is used, thereby allowing the update process to learn partially from previous success by the model in reproducing observed large scale climate variations, but also allowing for the possibility that previous performance may have a detrimental effect on future forecast skill and lead to underestimation of the scaling factor $\boldsymbol{\beta}_t$. Underestimation of $\boldsymbol{\beta}_t$ in a given decade might occur for a variety of reasons. These would include (i) small ensemble sizes, which would lead to contamination of the model ensemble response $\mathbf{X}_t$ by sampling errors and thus underestimation of $\boldsymbol{\beta}_t$ (see, for example, Allen and Tett (1999); see also topics in measurement error model, e.g., Fuller (1987, 2-5)); (ii) poor response to a short time scale forcing such as a volcanic event; or (iii) the occurrence of unusual natural internal variability, such as a strong El-Nino event, during a given decade that one would not expect a model to reproduce.

A third, called the persistence hindcast, is produced by using $\mathbf{Y}_{t-1}$ as $\mathbf{X}_t$ and then generating the hindcast using the Bayesian mechanism describe previously. It is anticipated that the persistence hindcast will be difficult to beat. While it does not benefit from a sophisticated formulation of the anticipated response to external forcing, it does implicitly benefit

from knowledge of the state of the climate system at the start of the forecast period, including the true response to external forcing up to that point. In addition, the Bayesian hindcasting process should be able to learn from aspects of internal climate variability that persist from one decade to the next.

2.4 Application

The observational data set used in this study is the same as that used in Lee et al. (2005), namely, the HadCRUTv dataset (Jones et al. 2001; data available at www.cru.uea.ac.uk). This is a combined dataset of monthly surface air and sea temperature anomalies relative to 1961-90 for the period 1870 to 1999 and is presented on a $5^o \times 5^o$ latitude-longitude grid. Various versions of this dataset have been used extensively in previous climate change detection and attribution studies.

The climate simulations of the 20th century used in this study are from the Canadian Centre for Climate Modelling and Analysis second generation coupled model CGCM2 (Flato and Boer 2001), the Hadley Center second and third generation CGCMs (HadCM2 and HadCM3) and simulations from six models in the IPCC 4th Assessment Report (IPCC AR4) model archives that are driven with estimates of historical forcings for the 20th century (CCSM3.0, GFDL2.0, GFDL2.1, MIROC3.2, MRI and PCM). A summary of the simulations used in this study is displayed in Table 2.1. Ensemble sizes for individual models range from 3 to 8 simulations of the 20th century. Earlier ensembles available for this study include only anthropogenic forcing, with sulphate aerosol forcing limited to the direct effect in some instances, while the IPCC AR4 simulations all include anthropogenic and natural external forcing, and generally include indirect aerosol effects. Three long control simulations from CGCM2, HadCM2 and HadCM3 are also used in this study. All simulations, which are available in a variety of grid sizes (Table 2.1), were interpolated onto the $5^o \times 5^o$ grid of the observations and subsequently averaged into regional decadal means (details to be described below). An analysis is conducted for each individual model and for the ensemble mean of the simulations from the six IPCC AR4 models.

Analysis of decadal predictability are carried out on regional decadal means calculated over

$30^o \times 40^o$ latitude-longitude grid boxes as in Lee et al. (2005). Monthly means in $30^o \times 40^o$ regions were calculated by averaging all available observed, or simulated, $5^o \times 5^o$ monthly means in the region. Observed annual means are treated as missing if even one month within the year is missing. Decadal means are treated as missing if fewer than 6 of the 10 years are present. The base period temperature is treated as missing only if all three decadal means are missing. To avoid systematic bias, missing data are not filled in. Instead, model output is flagged as missing whenever the corresponding observations are missing. In the absence of any missing data, the observational vector $\mathbf{Y}_t$ in a given decade would have length $n = 6 \times 9 = 54$, where 54 is the number of $30^o \times 40^o$ grid boxes that cover the globe. Missing data reduces the dimension length to a number ranging from 44 for the decade of the 1930's to 51 for the decade of the 1990's.

Model	Atmospheric Resolution	Runs	Forcings	Control length (yr)
CGCM2	T32 × L10	3	Anthro	1000
HadCM2	$2.5^o \times 3.75^o \times$ L19	4	Anthro	1019
HadCM3	$2.5^o \times 3.75^o \times$ L19	4	Anthro	1640
CCSM3.0	T85 × L26	8	Anthro+Nat	
GFDL2.0	$2.0^o \times 2.5^o \times$ L24	3	Anthro+Nat	
GFDL2.1	$2.0^o \times 2.5^o \times$ L24	3	Anthro+Nat	
MIROC3.2	T42 × L20	3	Anthro+Nat	
MRI	T42 × L30	5	Anthro+Nat	
PCM	T42 × L26	4	Anthro+Nat	

Table 2.1: Summary of simulations used in this study. 'Anthro' indicates that the simulation is driven by anthropogenic forcing consisting at least of greenhouse gas and direct sulphate aerosol forcing in the case of CGCM2 and HadCM2, but also including forcing from other sources, such as the indirect effects of sulphate aerosols, other non-sulphate aerosols, ozone, and land use in some of the more complete IPCC AR4 models. 'Anthro+Nat' indicates that the simulation is also driven by reconstructions of historical natural forcings such as solar and volcanic. The Control length column is filled in only when control simulation is available from that model. Horizontal atmospheric resolution is indicated either by the model's spectral resolution, or by grid box size expressed in degrees of longitude by degrees of latitude. Further details of the models are available on websites (CGCM2: `http://www.cccma.ec.gc.ca`; HadCM2 and HadCM3: `http://www.metoffice.com/research/hadleycentre/models/modeltypes.html`; IPCC AR4 models: `http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php`).

2.4.1 Covariance matrix estimation and dimension reduction

To generate the probability hindcasts, it is necessary to have an estimate of the natural internal variability of the surface temperature on the decadal and regional scales that are retained in the observation vector $\mathbf{Y}_t$. That is, an estimate of the variance-covariance matrix $\mathbf{\Sigma}$ of the term $\boldsymbol{\varepsilon}_t$ that appears in (2.1.1) is required. Instrumental observations during the past 150 years cannot provide a reliable estimate of $\mathbf{\Sigma}$ because natural climate noise is confounded with the effect of external forcings during the instrumental period. Also, the length of the observed record is not long enough to provide a reliable estimate of decadal scale variability. Hence, as in many climate change studies (e.g., Lee et al. 2005), this matrix is estimated from long control simulations in which only internal variability of the climate system is being simulated. Even though a growing number of long control runs are becoming available, there remains limitations on the spatial scale (number of grid boxes) in which analysis can be conducted. In this analysis, the available control runs are limited to about 1000 years in length. This translates to about 100 independent samples, each sample being a $n \times 1$ vector, for estimating $\mathbf{\Sigma}$ on decadal scale. In general, one can estimate $\mathbf{\Sigma}$ using the sample covariance matrix from the sample of 100. The dimension of the matrix $\mathbf{\Sigma}$ ranges from 44×44 to 51×51, depending on the number of missing observations at the particular analysis period. Thus, a sample size of 100 might not be suffice to provide an accurate estimate of $\mathbf{\Sigma}$ in this case.

A dimension reduction technique that has been widely used in climate change detection analysis is employed here to provide an potentially better estimate for $\mathbf{\Sigma}$. Denote the sample covariance matrix obtained from a control simulation as $\hat{\mathbf{\Sigma}}$. If one assume $\hat{\mathbf{\Sigma}}$ can provide a reliable estimate of the variability in the subspace spanned by the k gravest Empirical Orthogonal Functions (EOFs; or equivalently, the eigenvectors), then one can conduct analysis in the k-dimensional space that retains only the large spatial scales of variation in decadal anomalies. This is done by projecting all the observed and simulated decadal anomalies onto the k gravest EOFs of $\hat{\mathbf{\Sigma}}$. The choice of k is determined by two criteria. First, the retained scales should well represent the model simulated response anomaly $\mathbf{X}_t$. More importantly, the model simulated internal variability should be consistent with observed variability (Allen and Tett, 1999) on these scales. Previous detection work at global and regional scales suggests

that moderate values of k can be used. As in Zwiers and Zhang (2003) and Lee et al. (2005), the results presented here are found to be insensitive to the choice of k for $5 \leq k \leq 25$.

The data in the reduced space is $\mathbf{P}^{(k)}\mathbf{Y}_t$ and $\mathbf{P}^{(k)}\mathbf{X}_t$, where the rows of $\mathbf{P}^{(k)}$ are the first k EOFs of $\hat{\mathbf{\Sigma}}$. Given such a dimension reduction, a natural estimate of the variance structure of the internal variability $\boldsymbol{\varepsilon}_t$ in the reduced space is $\mathbf{P}^{(k)}\tilde{\mathbf{\Sigma}}\mathbf{P}^{(k)\mathrm{T}}$, where $\tilde{\mathbf{\Sigma}}$ is a sample covariance matrix obtained from control run data that has not been used to obtain $\hat{\mathbf{\Sigma}}$. The matrix $\tilde{\mathbf{\Sigma}}$ is used rather than $\hat{\mathbf{\Sigma}}$ to estimate the variance structure of internal variability in the reduced space because biases would creep into the analysis from using only one control run sample to estimate both the EOFs and the variability in the reduced space. Such biases arise because the EOF basis vectors inevitably "adapt" to the specific variations present in the part of the control run from which they are estimated (Allen and Tett 1999; Allen and Stott 2003).

The variance and mean of the posterior distribution of the Bayesian analysis in the reduced space is therefore defined as

$$\begin{aligned} c_{t+10} &= \left[1/c_t + \mathbf{X}_t^{\mathrm{T}}\mathbf{P}^{(k)\mathrm{T}}(\mathbf{P}^{(k)}\tilde{\mathbf{\Sigma}}\mathbf{P}^{(k)\mathrm{T}})^{-1}\mathbf{P}^{(k)}\mathbf{X}_t\right]^{-1} \\ m_{t+10} &= c_{t+10}\left[\{\lambda m_t + (1-\lambda)\}/c_t + \mathbf{X}_t^{\mathrm{T}}\mathbf{P}^{(k)\mathrm{T}}(\mathbf{P}^{(k)}\tilde{\mathbf{\Sigma}}\mathbf{P}^{(k)\mathrm{T}})^{-1}\mathbf{P}^{(k)}\mathbf{Y}_t\right]. \end{aligned} \tag{2.4.1}$$

By projecting $\mathbf{P}^{(k)}\mathbf{X}_t$ back to the original space, the hindcasting distribution (cf. (2.1.3)) now becomes,

$$f(\mathbf{Y}_t \mid \mathbf{X}_t) = \phi_n\left(\{\lambda m_t + (1-\lambda)\}\mathbf{P}^{(k)\mathrm{T}}\mathbf{P}^{(k)}\mathbf{X}_t,\ c_t\mathbf{P}^{(k)\mathrm{T}}\mathbf{P}^{(k)}\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}\mathbf{P}^{(k)\mathrm{T}}\mathbf{P}^{(k)} + \tilde{\mathbf{\Sigma}}\right)$$

and for other quantities such as global mean (cf. (2.1.7)), the distribution is

$$\begin{aligned} &f(\mathbf{y}_{t\cdot} \mid \mathbf{X}_t) \\ &= \phi_n\left(\{\lambda m_t + (1-\lambda)\}w_t^{\mathrm{T}}\mathbf{P}^{(k)\mathrm{T}}\mathbf{P}^{(k)}\mathbf{X}_t,\ c_t w_t^{\mathrm{T}}\mathbf{P}^{(k)\mathrm{T}}\mathbf{P}^{(k)}\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}\mathbf{P}^{(k)\mathrm{T}}\mathbf{P}^{(k)}w_t + w_t^{\mathrm{T}}\tilde{\mathbf{\Sigma}}w_t\right). \end{aligned} \tag{2.4.2}$$

Recall that $\lambda = 1$ for the full Bayesian hindcast, $\lambda = 0.5$ for the blended hindcast, and $\lambda = 0$ for the raw hindcast. Also, the matrices $\tilde{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Sigma}}$ vary slightly from one hindcast period to the next because the masking of the observations varies. Thus the individual EOFs that are

used for the dimension reduction also vary somewhat from one hindcast period to the next.

The details of estimating the covariance matrix $\hat{\mathbf{\Sigma}}$ and $\tilde{\mathbf{\Sigma}}$ are as follows. Three control simulations are available in this study. To avoid bias when estimating the covariance matrix, only the last 1000 years of HadCM2 and HadCM3 control simulations are used so that their length matches with that of the CGCM2 control simulation. Each control simulation is divided into two 500-yr subsets and each 500-yr control run subset is formed into 99 10-yr chunks, each overlapped by 5 years. Decadal means are computed from these chunks and the prior 3-decade mean is subtracted from each chunk. This results in 93 decadal anomalies from their respective prior 3-decade climatologies within each 500-yr subset of each control run. This approach of calculating overlapping decadal anomalies provides somewhat more information for calculating the covariance matrix than would be available from only the 47 non-overlapping decadal anomalies that can be computed from years 31-40, 41-50, ..., 491-500 of the control run. By overlapping decades, an additional 46 decadal anomalies are obtained from years 36-45, 46-55, ..., 486-495. A sample covariance matrix is calculated for each control run from the first collection of 93 anomalies using the standard formula. The average of the three resulting covariance matrices is then used as an estimate of $\hat{\mathbf{\Sigma}}$. A second estimate $\tilde{\mathbf{\Sigma}}$ is similarly calculated from the collections of 93 anomalies obtained from the second 500-yr segment of each of the three control runs. Note that these calculations are repeated for each hindcast because the masking that reflects whether observations are missing varies from one hindcast period to the next.

2.4.2 Determining the significance of the BSS

The BSS that is use to evaluate the forecasts is affected by the specific realization $\boldsymbol{\varepsilon}_t$ of the climate's internal variability that is present during the hindcast period. For example, under the assumption of no climate change, if there are more positive elements in $\boldsymbol{\varepsilon}_t$ than negative by random chances and the hindcast predicts the above normal event to be more likely to happen, then one can obtain a BSS that is greater than zero. However, such skill arise from the hindcast is merely a result of the sampling variability in $\boldsymbol{\varepsilon}_t$. Thus one would expect the BSS of an unskilled hindcast or forecast to vary about zero as a result of sampling variability.

It is therefore necessary to construct an upper critical bound for the BSS in order to identify a skill threshold above which one can reject the null hypothesis that the forecast is not skillful. Such a critical level can be estimated by verifying the hindcast against decadal anomalies calculated from the three available 1000-yr control simulations. A total of $97 \times 3 = 291$ such anomalies can be obtained by using non-overlapping 10-yr chunks. The resulting sample of BSS's reflects the range of skill scores one would obtained if the verification data set consists of only internal climate variability. The upper 5% critical level for the BSS is then easily estimated by calculating the 95^{th} percentile of the sample of 291 Brier Skill Scores. This critical value varies slightly from one decade to the next because the observational masking changes in time.

2.4.3 Hindcast results

Temperature anomaly hindcasts for each model and the AR4 ensemble were produced using the methods described above for seven decades (1930-39, 1940-49, ..., 1990-99). In addition, a forecast for the decade 2000-2009 using the CGCM2 model are also produced.

To produce the first hindcast for the decade 1930-39, it was necessary to specify a prior distribution on the scaling factor $\boldsymbol{\beta}_{1930}$ that starts the Bayesian forecasting process. That is, one need to choose values for the parameters m_{1930} and c_{1930}. Prior distributions for the subsequent hindcasts were then obtained by using the posterior-prior updating process described in previous section. The prior distribution on $\boldsymbol{\beta}_{1930}$ is subjectively chosen to be $\phi(1, 0.25)$, thereby producing an initial 4 standard deviation uncertainty on the scaling factor that ranges from 0 to +2. The robustness of the results to the initial choice of prior is discussed in the latter part of this section.

Figure 2.1 displays the BSS for the above-normal event as a function of decade with 15 EOFs retained for the full Bayesian hindcast (with $\lambda = 1$), together with corresponding 5% critical value for rejecting the null hypothesis of an unskillful hindcast for the CGCM2 hindcast (thick horizontal line segments). The critical values for other models are very similar and thus not shown. The analysis is repeated by retaining 5, 10, 20 and 25 EOFs (not shown) and find that the BSS is not very sensitive to the number of EOFs retained. With 15 EOFs

retained, the BSS for CGCM2 lies above the critical value for the decades 1930-39, 1940-49, 1980-89 and 1990-99, suggesting that the temperature anomaly hindcasts for those decades may have significant skill. Specifically, the BSS's for these periods are 0.523, 0.373, 0.439 and 0.684 respectively. Similar results are obtained for HadCM2 and HadCM3, for the hindcasts produced from the AR4 model simulations and for their ensemble mean, where the BSS is above the critical value for the early and late decades of the 20th century. This suggests that the inclusion of natural forcing in the simulations is not the solution to the apparent lack of skill in 1950's to 1970's.

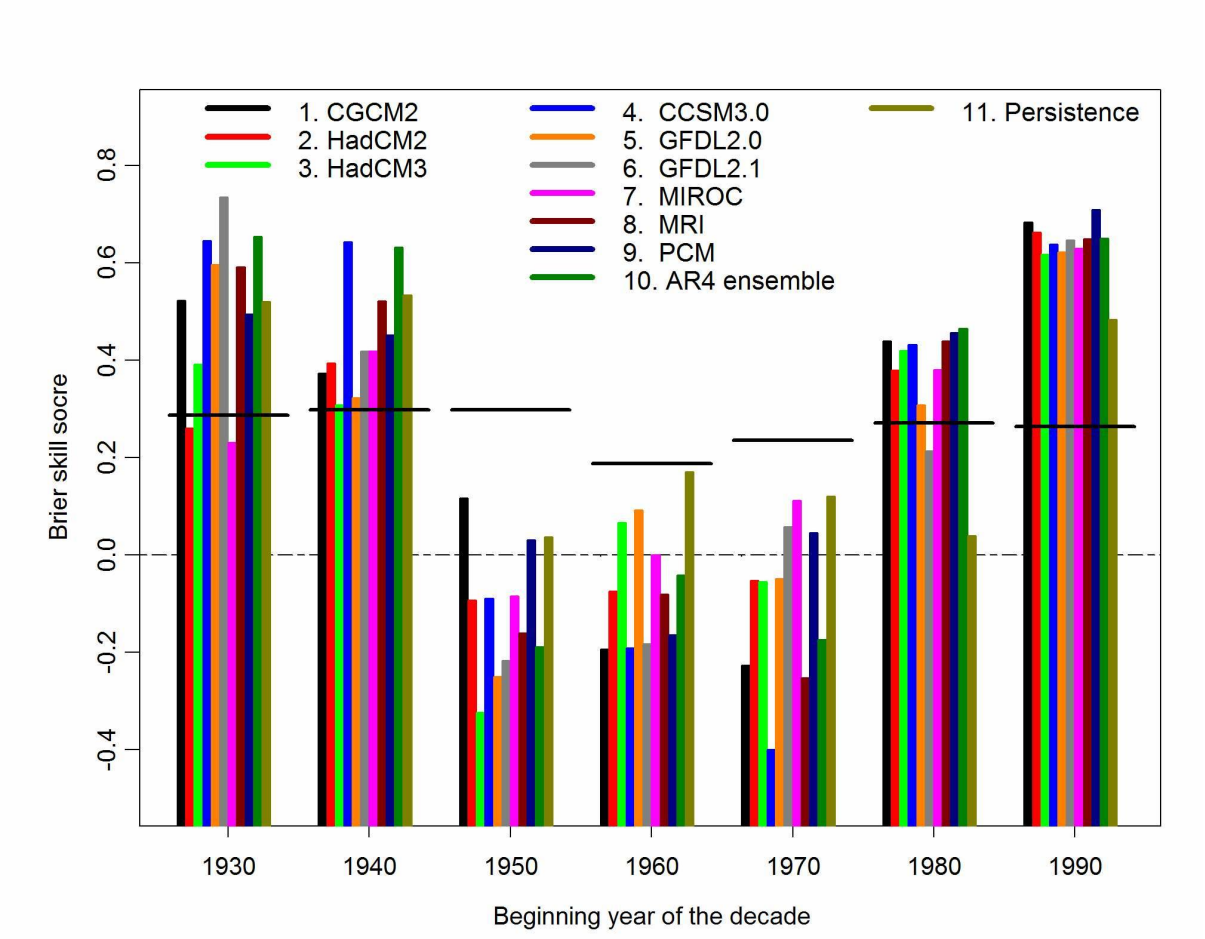

Figure 2.1: Brier skill scores for the above normal event with 15 EOFs retained. The thick horizontal line segments indicate the critical skill threshold for rejecting the no-skill null hypothesis. See text for details. Skill scores are shown only for the full Bayesian analysis with $\lambda = 1$ (see text). Results for the blended and raw hindcasts are very similar.

The skill obtained for the blended and raw hindcasts (not shown) is very similar to that obtained with the full Bayesian hindcast, with minor variations in skill (both increases and decreases) depending upon the model that is used. This indicates that variation in the details of the prior updating process does not have a large influence on forecast skill, at least as measured by the BSS. This is a reasonable result given that the BSS measures the agreement between the spatial pattern of the hindcasts of above normal probability, and the verifying pattern of above normal events. The prior updating process would have some influence on the amplitude of the pattern of hindcast probabilities, but it does not affect its shape, and thus has little influence on the Brier skill scores.

A concern with respect to the skill scores is that they may be sensitive to the choice of prior distribution on the 1930's scaling factor $\boldsymbol{\beta}_{1930}$. Thus, the full Bayesian analysis is repeated using two other classes of priors to evaluate that possibility. The first type of prior is identical to that used above except that the initial mean (m_{1930}) was varied between -1 and 3. The second type of prior has $m_{1930} = 1$ but uses variances c_{1930} that range from 0.1 to 1.1. The BSS for CGCM2 obtained by using these priors ranges from -1.09 to 0.59 for the decade 1930-39 and from 0.184 to 0.460 for the decade 1940-49. However, BSS's in subsequent decades, after the Bayesian updating process commences, are very similar to those shown in Figure 2.1. The hindcast probability, hindcast anomaly and posterior distribution were found to be insensitive to the initial choice of prior after the first three decades, with only a minor impact on the third decade. Similar behavior was also exhibited by the other models that have considered here. The robustness of the results after the initial two to three decades is mainly due to the calibration of the distribution of $\boldsymbol{\beta}_t$ from the posterior-prior updating process. While results for all decades will continue to show in the remainder of this chapter, the sensitivity to the choice of prior during the first two hindcast decades indicates that results for those two decades should be down weighted.

Figure 2.2 displays the global mean hindcast probability for the above-normal event as a function of decade with 15 EOFs retained together with the observed proportion of such events and its confidence bound. An approximate 95% confidence bound for the observed proportion $\hat{p}^o$ can be defined as $\hat{p}^o \pm 2\sqrt{\hat{p}^o(1-\hat{p}^o)/n}$ where n is the number of spatial points

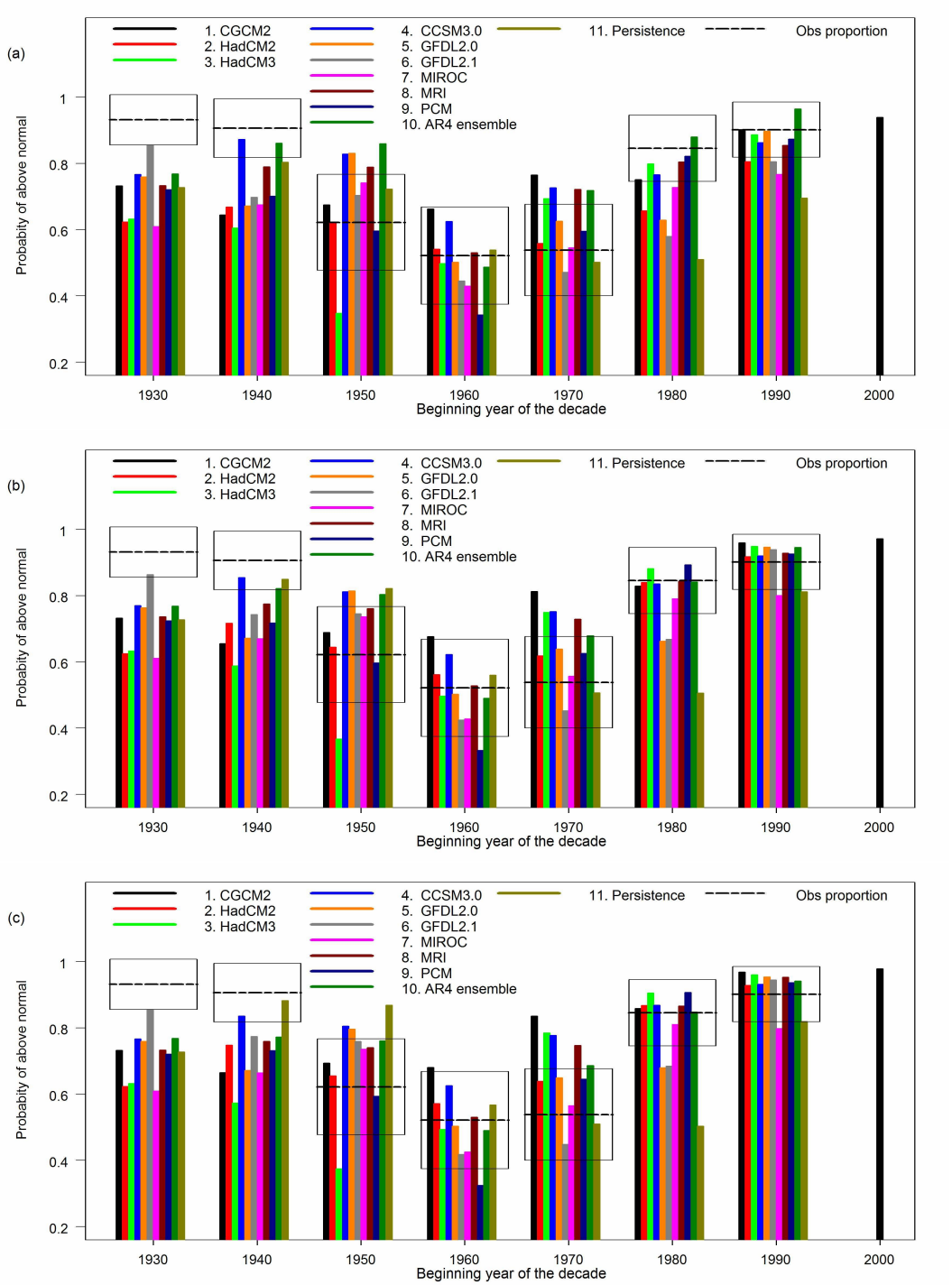

Figure 2.2: Global mean hindcast probabilities of above normal decadal mean temperatures in $30^o \times 40^o$ latitude-longitude regions (bars) and the observed proportion of regions with above normal temperatures (thick horizontal dashed line segments) together with its estimated uncertainty (box). The probability scale is indicated on the left. A regional decadal mean temperature is considered to be above normal in a given decade when it is greater than the corresponding climatological mean temperature in the region for the preceding three decades. A forecast for the 2000-2009 decade using the CGCM2 model is also displayed.(a) Full Bayesian hindcast, (b) blended hindcast, and (c) raw hindcast.

in the analysis. This bound accounts for sampling variations in the observed proportion of above normal events that would be expected under similar conditions, and under the assumption of spatial independence. The actual confidence bound is likely to be wider because of dependence between regions. Figure 2.2a shows that the CGCM2 hindcast probabilities significantly under-estimate the proportion of observed above-normal events for the decades 1930-39 and 1940-49, but are within the uncertainty range for 1950's, 1960's, 1980's and 1990's. Considering all ten model hindcasts together, the hindcast performance is generally poor in the 1930's and 1940's, but starting from the 1950's, a majority of the models "correctly" hindcast the observed proportion in each decade. As noted previously, hindcasts of the first two decades are sensitive to the choice of initial prior, and thus these results should be discounted. Also, as with the results for the BSS, the details of the prior updating process have a relatively small effect, although there is some evidence (compare Figures 2.2b,c with Figure 2.2a) that either the blended or raw hindcast performs slightly better than the full Bayesian hindcast, perhaps because the latter allows the scaling factor $\boldsymbol{\beta}_t$ to be too heavily influenced by past forecast errors.

Figure 2.3 shows the corresponding hindcasts as derived from (2.4.2) for the actual global mean decadal temperature anomalies together with their corresponding 5-95% hindcast confidence intervals and the observed anomalies. This graph also provides a forecast for the 2000-2009 decade using the CGCM2 model. The information provided is essentially the same as that provided in Figure 2.2. In particular, there is good agreement between the observed and hindcast anomalies during the last four decades of the 20th century. Specifically, for the full Bayesian procedure, the CGCM2, HadCM3, CCSM3.0, MRI and PCM hindcasts were able to capture the observed anomalies throughout 1960's to 1990's. In addition, the AR4 ensemble hindcast was able to capture six out of seven observed anomalies. In contrast, the hindcast anomalies of the earlier skillful hindcasts, for the decade of the 1930's and 1940's, were less promising for all the models, except for CCSM3.0 and AR4 ensemble. Based on the anticipated response to anthropogenic forcing and assuming that there will not be a substantial change in natural external forcing, the global mean temperature anomaly for the current decade (2000-2009) is predicted to be 0.35°C with a 5-95% confidence range 0.21°C to 0.48°C

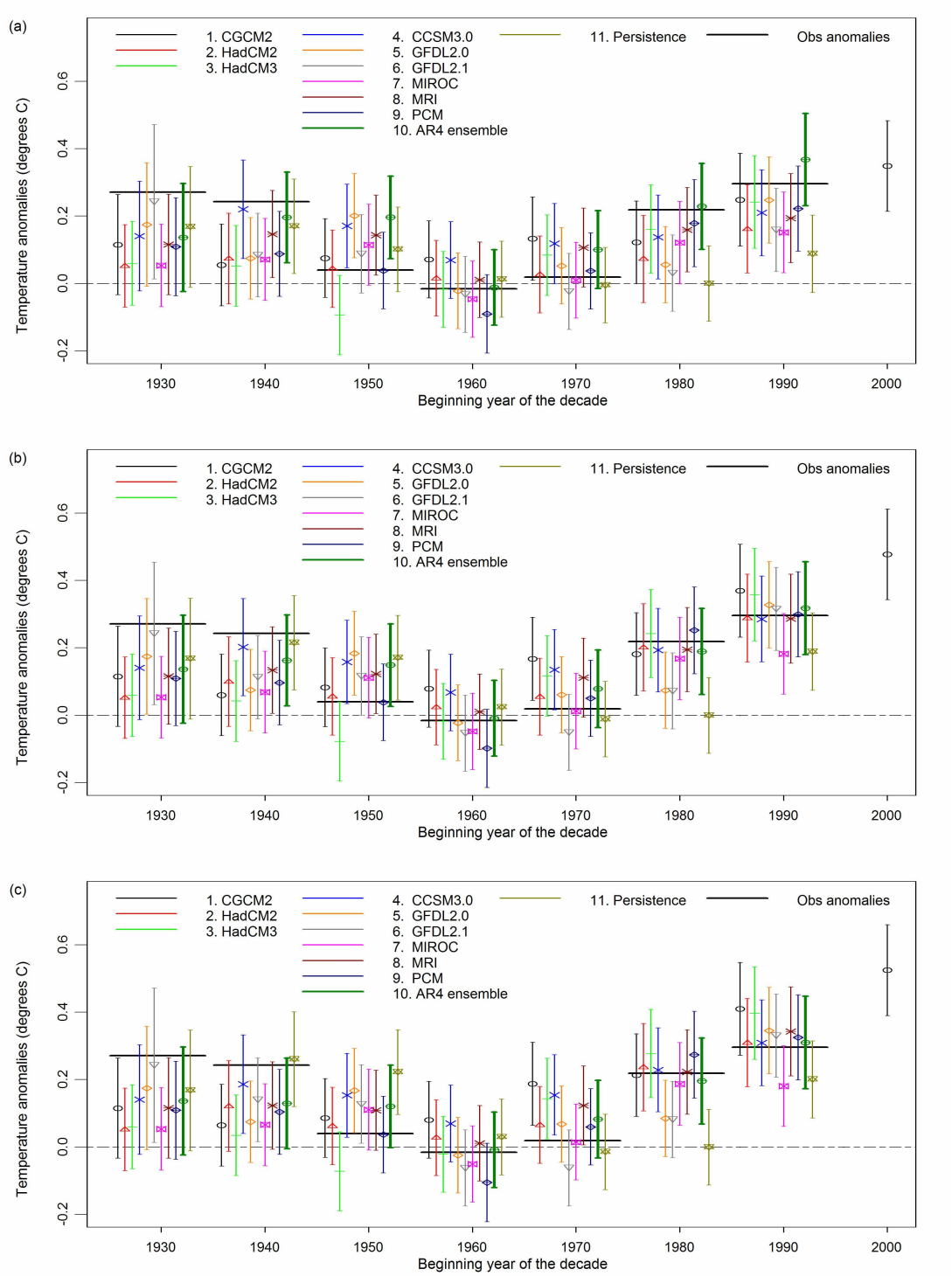

Figure 2.3: Hindcasts of global decadal mean surface temperature anomalies and their 5-95% confidence bounds based on (2.4.2), together with the observed global mean anomalies (thick horizontal line segments), with 15 EOFs retained. A forecast for the decadal global mean anomaly for the decade 2000-2009, relative to the 1970-1999 climatology, is also displayed. Units are ^{o}C. (a) Full Bayesian hindcast, (b) blended hindcast, and (c) raw hindcast.

using the CGCM2 model.

Again, as above, there is some evidence that the blended or raw hindcast (Figures 2.3b and 2.3c) has slightly better performance than the full Bayesian hindcast (Figure 2.3a). The AR4 ensemble, for example, is able to capture all of the observed anomalies in the former two cases. There is some evidence that the full Bayesian hindcast overly constrains the model forecasts (compare, for example, the 1990s in Figure 2.3a with the same period in Figures 2.3b,c), suggesting that the scaling factor $\boldsymbol{\beta}_t$ has been underestimated. However, the blended and raw hindcasts also impose weaker constraints on models that warm quickly, such as CGCM2, with the result that greater warming is forecast for the first decade of the 3$^{\text{rd}}$ millennium. In particular, the global mean temperature anomaly for 2000-2009 is predicted to be 0.48^{o}C (5-95% confidence range: 0.34°C to 0.61°C) for blended hindcast and 0.52°C (5-95% confidence range: 0.39°C to 0.66°C) for raw hindcast using the CGCM2 model.

Figure 2.4 displays the 5-95 percentile for the posterior distribution of the scaling factor $\boldsymbol{\beta}_t$ for the full Bayesian hindcast as in (2.4.1) with 15 EOFs retained. The mean of the posterior distribution has a general downward trend for all the models. For example, for the CGCM2 model, the 5-95 percentile of the posterior confidence interval lies below one for the last five decades. This may be partly due to negative bias in the estimate of $\boldsymbol{\beta}_t$ due to weak decadal scale signals (see Allen and Stott 2003, for a discussion). However, it may also reflect an over simulated response to 20th century forcing, perhaps because of missing forcings in the case of this model (Lee et al. 2005). In contrast, the 5-95 percentile ranges for the AR4 ensemble include the possibility that $\boldsymbol{\beta}_t = 1$ for the 1960's, 1970's, 1980's and 1990's, providing evidence towards attribution of natural and anthropogenic influence on climate for these four decades in the context of optimal fingerprinting. Note that the $\boldsymbol{\beta}_t$ for the AR4 ensemble is bigger than those obtained using the individual AR4 models. This is possibly due to the reduction of sampling uncertainty in the simulated response pattern for the AR4 ensemble hindcast as more ensemble members is used. As would be expected, the scaling factors obtained for the other two hindcasts (not shown) are much more tightly constrained to $\boldsymbol{\beta}_t = 1$.

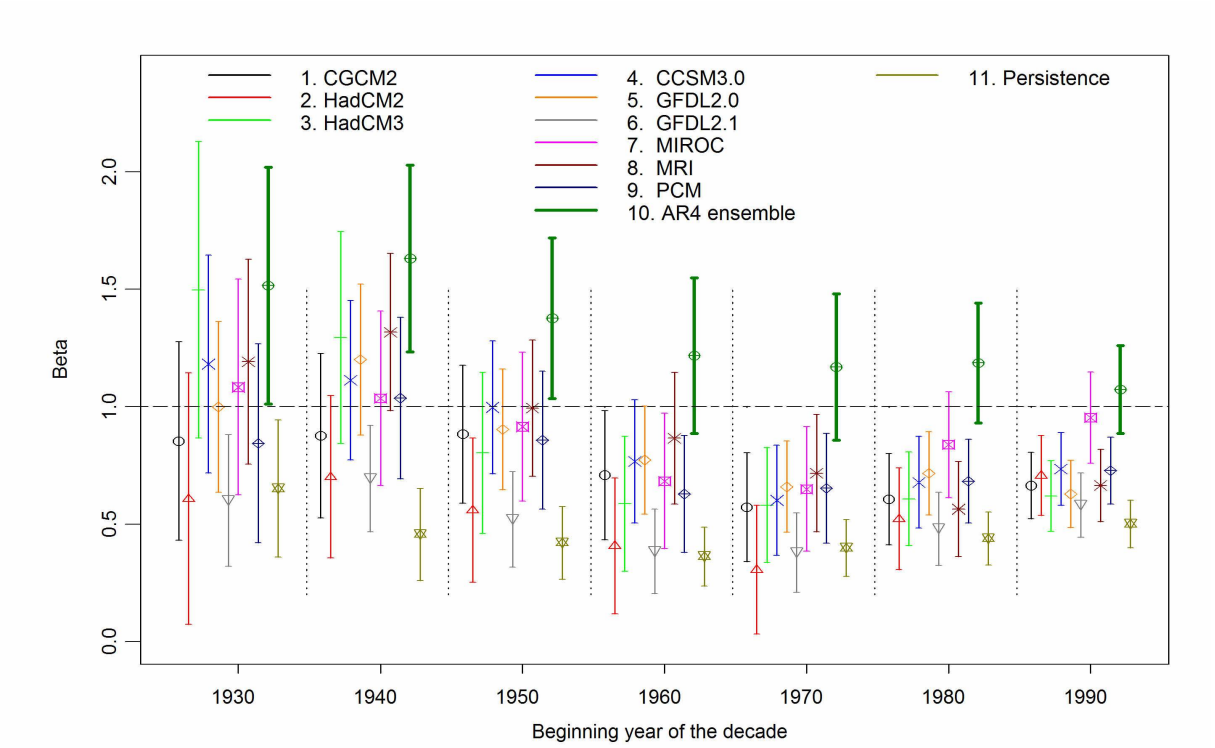

Figure 2.4: Mean and the 5-95 percentile of the posterior distribution of the scaling factor $\boldsymbol{\beta}_t$ from the full Bayesian hindcast with 15 EOFs retained.

Even though the hindcasts are skillful for part of the 20th century, there remains the question of whether this reflects the capability of the climate models to predict the decadal response to external forcing, or whether the skill is really just an artifact arising from natural persistence. Such persistence might arise from two sources - either low frequency natural internal variability, or a persistent forcing disequilibrium resulting in a continuing response of the climate system to that disequilibrium.

The former possibility is, in fact, taken into account in the estimation of the 5% critical value for the BSS (see Section 2.3). BSS's for the decades of the 1980's and 1990's (Figure 2.1), which are not affected by the choice of initial prior, are significantly greater than the estimated critical value, suggesting an external, rather than internal source of skill. This assessment is, of course, subject to the caveat that the control runs used to estimate these critical values correctly simulate natural internal low frequency variability of the climate system on the space and time scales retained in the hindcasting analysis.

The latter possibility is considered by evaluating the performance of the persistence forecast, the results of which are indicated by the olive-green bars in Figures 2.1-2.4. As discussed above, persistence is a very tough forecast to beat at decadal leads given the apparent linearity of the response to forcing (e.g., Gillett et al. 2004b) and the fact that response during any given decade is more reflective of a continuing response to historical forcing, than to forcing change during the hindcast decade. The BSS's for the persistence hindcast are similar to those obtained using climate model hindcasts, except in the last two decades, where the BSS is significantly lower for the persistence hindcast (Figure 2.1). This holds regardless of whether one uses the full Bayesian hindcast procedure (Figure 2.1) or the blended or raw hindcast procedures (not shown). Similarly, the persistence hindcast under-hindcasts the global mean hindcast probability of above normal during the last two decades, regardless of the hindcast procedure used (Figure 2.2), and it under-predicts the global mean temperature anomaly (Figure 2.3). The latter problem is particularly evident in the full Bayesian hindcast (Figure 2.3a), perhaps because the "signal" used in persistence hindcast is heavily contaminated by noise from internal variability (which leads to negative bias in the estimate of $\boldsymbol{\beta}_t$). Thus despite the expectation that the persistence hindcasts would be difficult to beat, it would appear that the model based hindcasts outperform the persistence hindcast during the last two decades when anthropogenic forcing is largest.

2.5 Concluding remarks

In this chapter, another approach has been put forward to climate change detection analysis that is based on the skill of probabilistic decadal hindcasts that are produced from simulations of the climate of the 20th century with a Bayesian technique. Specifically, hindcasts of decadal temperature anomalies on large spatial scales relative to the 3-decade operational climatologies that are current at the time of the hindcast are considered. Consistent with other detection studies, the Bayesian analysis indicates that the effect of greenhouse gas and sulfate aerosols is detectable in the latter part of the 20^{th} century. Statistical characteristics of the hindcasts, such as the global mean hindcast of the probability of above normal and the hindcast of the global mean temperature, are consistent with the characteristics of the verifying observations

from the 1950's onwards. The BSS's indicate that the model based hindcasts demonstrate significant skill during the last two decades of the 20th century. Comparison between the model based hindcast and a persistence hindcast suggest that the models add value during this period relative to simple persistence.On the other hand, the results for 1930's and 1940's should be down weighted because they are sensitive to the initial choice of prior distribution. As in other studies (Derome et al.2001; Kharin and Zwiers 2002; Gillett et al. 2002; Zhang et al. 2005), there is also some evidence that the ensemble model mean approach performs more consistently than do individual models. The inclusion of natural external forcing does not appear to significantly improve short term (i.e., decadal) hindcast skill, perhaps because the response associated with natural forcing is small relative to that associated with anthropogenic forcing.

Further work will be required to more clearly identify the factors that contribute skill on the decadal time scale, to make more sophisticated use of multi-model ensembles as in seasonal forecasting (e.g., Derome et al. 2001; Kharin and Zwiers 2002), and by assimilating observed ocean, and perhaps land surface, state information into the model. A multi-signal analysis that attempts to tease out the effect on skill of the different external forcing factors may be feasible, but requires the specification of a multivariate prior distribution. The corresponding hindcast distribution and posterior distribution can then be derived from (2.1.1) by using the Bayes' theorem based on the prior distribution.

Forecasts for events in the future can be generated using the same methodology. However, one cannot carry out the posterior-prior updating process since observations are not available. Also, simulations of the 20th century must then be extended into the future using a scenario of future emissions. While the simulated response 1-2 decades into the future is not likely to be sensitive to scenario details (e.g., Zwiers 2002), there are forcing uncertainties, such as the possibility of unforeseen volcanic activity, that must be taken into account. Using the CGCM2 simulated anthropogenic signal as $\mathbf{X}_t$ and using the posterior distribution from 1990-99 as the prior distribution, it is predicted that in the absence of large negative volcanic forcing on the climate system (which can not presently be forecasted) the global mean temperature anomaly for the decade 2000-2009 will be above the 1970-1999 normal with probability 0.94. The global

mean temperature increment for this decade is correspondingly predicted to be 0.35°C with a 5-95% confidence range of 0.21°C to 0.48°C (Figure 2.3a). The suggestion that such decadal forecasts are now apparently skillful on large regional scales based only on anthropogenic forcing, and that they can be regularly updated and verified, provides additional evidence for the influence that anthropogenic forcing is having on the climate.

2.6 Appendix: Derivation of posterior distribution

The derivation for the posterior distribution used to produce hindcast will be given below. Assume the prior distribution of $\boldsymbol{\beta}_t$ is

$$f(\boldsymbol{\beta}_t) = \phi(m_t, c_t).$$

Conditional on $\boldsymbol{\beta}_t$, the distribution of $\mathbf{Y}_t$ is,

$$f(\mathbf{Y}_t|\boldsymbol{\beta}_t) = \phi_n(\boldsymbol{\beta}_t\mathbf{X}_t, \Sigma).$$

Hence, by Bayes' Theorem, the posterior distribution of $\boldsymbol{\beta}_t$ is

$$\begin{aligned}
f(\boldsymbol{\beta}_t|\mathbf{Y}_t) &\propto f(\boldsymbol{\beta}_t)f(\mathbf{Y}_t|\boldsymbol{\beta}_t) \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{(\boldsymbol{\beta}_t - m_t)^2}{c_t} + (\mathbf{Y}_t - \boldsymbol{\beta}_t\mathbf{X}_t)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_t - \boldsymbol{\beta}_t\mathbf{X}_t)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[c_t^{-1}\boldsymbol{\beta}_t^2 - 2c_t^{-1}\boldsymbol{\beta}_t m_t + \boldsymbol{\beta}_t^2\mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_t - 2\boldsymbol{\beta}_t\mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}_t\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_t^2(c_t^{-1} + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_t) - 2\boldsymbol{\beta}_t(c_t^{-1}m_t + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}_t\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}(c_t^{-1} + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_t)\left(\boldsymbol{\beta}_t - \frac{c_t^{-1}m_t + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}_t}{c_t^{-1} + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_t}\right)^2\right\}.
\end{aligned}$$

Thus the posterior distribution of $\boldsymbol{\beta}_t$ is $f(\boldsymbol{\beta}_{t+10}) = f(\boldsymbol{\beta}_t|\mathbf{Y}_t) = \phi(c_{t+10}, m_{t+10})$ where $c_{t+10} = (1/c_t + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}_t)^{-1}$ and $m_{t+10} = c_{t+10}(m_t/c_t + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}_t)$.

3 State-space model for historical temperature reconstruction

Knowledge of how climate has changed in the past centuries can provide important information on the role of anthropogenic forcing in the observed 20^{th} century warming. The lack of systematic records of temperature prior to the latter half of the 19^{th} century places emphasis on the need to reconstruct historical temperature from climate proxy data. Local temperature reconstructed from a single climate proxy series is usually of little use as it does not provide an adequate picture on large-scale changes. A more meaningful approach would involve reconstructing large-scale (global or hemispheric) mean temperature from a network of proxy data, where each proxy series comes from a different location on the globe. In most studies, the reconstruction target is the northern hemispheric mean temperature because proxy data, derived mainly from trees, are more widely available from the northern hemisphere land masses than elsewhere. The size of the proxy network used in recent reconstructions varies between ten to over a hundred locations.

Reconstruction of the hemispheric mean temperature requires the use of statistical method to map the proxy data onto the reconstructed temperature. Such mapping involves first finding the relationship between the proxy data and the instrumental temperature record during a ***calibration period***, the period when both records are available. This relationship is then applied to the proxy data that are available for the period prior to the calibration period to obtain a reconstruction of the historical temperature record. In most reconstructions, a composite of the proxy data, instead of an individual proxy, is used in the mapping process. The ***composite*** series is typically a weighted average of all the available proxy series (see Chapter 4). The use of such composite series can reduce the effect of noise in the individual proxy series on the reconstructed temperature.

Figure 3.1 provides a visualization of the structure of the reconstruction problem. In general, one has data available from the proxy composite ($\mathbf{P}_t$) over the pre-calibration ($t = 1, 2, \cdots, n$) and calibration periods ($t = n+1, n+2, \cdots, N = n+m$) and from the hemispheric mean temperature record ($\mathbf{T}_t$) during the calibration period. The vector of exogenous variables

($\mathbf{F}_t$) contains series that are thought to be correlated with the hemispheric mean temperature, for example, the estimated responses to external forcings. Existing reconstruction methods use only information from the first two sources. A new reconstruction method that uses knowledge from all three sources will be consider in this and the next chapter. This new method is based on a state-space time series model and the Kalman filter and smoother algorithm.

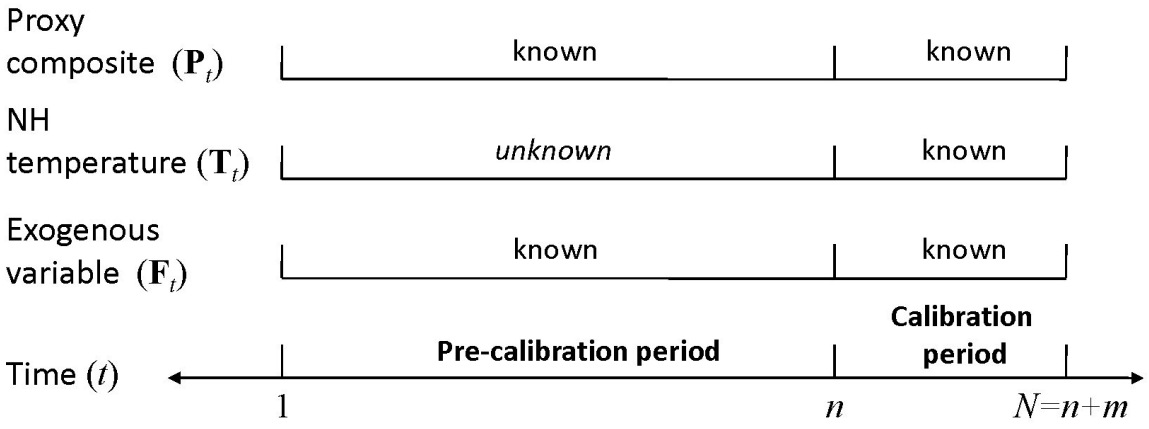

Figure 3.1: Structure of the reconstruction problem.

The remainder of this chapter is organized as follows. An introduction to state-space models and the Kalman filter (Kalman 1960) is given in section 3.1. The connection between the state-space model and the reconstruction problem will also be described. Sections 3.2 and 3.3 provide details on estimating the parameters that specify the state-space model. The existing parameter estimation method is modified to better suit the reconstruction problem, resulting in two new estimation methods. Asymptotic results for the new estimators are derived. The competing estimation methods are then compared through a simulation study in section 3.4. Concluding remarks are given in section 3.5. Section 3.6 provides proofs and derivations for the results presented in this chapter. It also provides theorems that are required to prove various results in this chapter. The application of the new reconstruction method will be deferred to Chapter 4.

3.1 State-space model and the Kalman filter

Before explaining how the temperature reconstruction problem can be cast in the state-space model framework, some details on what defines a state-space model will be given. In terms of notation, $f_\Theta(\mathbf{X})$ denotes a density function of the random vector $\mathbf{X}$ that is parameterized by the parameter Θ. To conserve space, $\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_N$ and $\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_N$ will be sometimes written as $\mathbf{P}_{1:N}$ and $\mathbf{T}_{1:N}$ respectively. This notation is used throughout this chapter, unless otherwise specified.

Let $\{\mathbf{T}_t\}$ be a sequence of random variables and let $g_\Theta(\mathbf{T}_s|\mathbf{T}_{s-1})$ denote the density of $\mathbf{T}_s$ conditional on $\mathbf{T}_{s-1}$, which is parameterized by the parameter Θ. Let $\{\mathbf{P}_t\}$ denote a sequence of random variables such that given $\mathbf{T}_t$, the $\mathbf{P}_t$'s are independent and the distribution of $\mathbf{P}_s$ depends only on $\mathbf{T}_s$ and has density $h_\Theta(\mathbf{P}_s|\mathbf{T}_s)$. The ***state-space model*** representation of $\mathbf{T}_t$ and $\mathbf{P}_t$ is then formulated as

$$\begin{aligned} \mathbf{P}_t|\mathbf{T}_t &\sim h_\Theta(\mathbf{P}_t|\mathbf{T}_t) \\ \mathbf{T}_t|\mathbf{T}_{t-1} &\sim g_\Theta(\mathbf{T}_t|\mathbf{T}_{t-1}). \end{aligned} \tag{3.1.1}$$

The $\{\mathbf{P}_t\}$ process has observed values $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N$ while the $\{\mathbf{T}_t\}$ process remains unobserved. The $\{\mathbf{P}_t\}$ process is called the ***observation process*** and the $\{\mathbf{T}_t\}$ process is called the ***state process***.

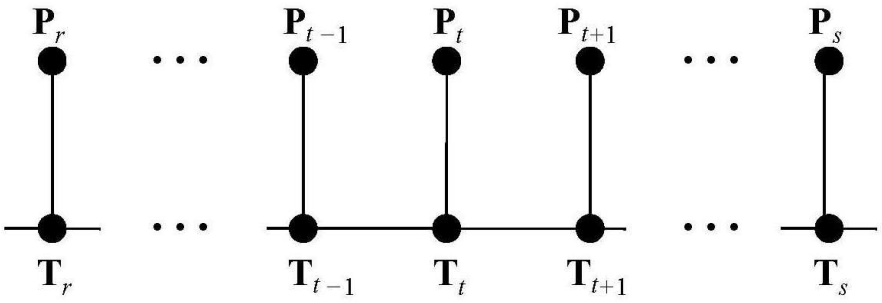

Figure 3.2: Conditional independence structure of a state-space model.

A graphical visualization of the conditional structure of a state-space model is given in Figure 3.2. From this figure, the following can be seen.

1. For $r < t < s$, conditioning on $\mathbf{T}_t$, $\mathbf{P}_r$ is independent of $\mathbf{P}_s$.

2. For $r \leq t \leq s$, $\mathbf{T}_r|(\mathbf{T}_t, \mathbf{T}_s)$ is equivalent to $\mathbf{T}_r|\mathbf{T}_t$. Similarly, $\mathbf{T}_s|(\mathbf{T}_t, \mathbf{T}_r)$ is equivalent to $\mathbf{T}_s|\mathbf{T}_t$

Readers are referred to Harvey (1989), Shumway and Stoffer (2000) and Durbin and Koopman (2001) for more details and examples of state-space models. When the conditional density functions in (3.1.1) are all normal (i.e., Gaussian), the state-space model is called a ***Gaussian state-space model***. Throughout the rest of this thesis, all work presented will be based on the Gaussian state-space model. Departure from the normality assumption has been considered in the literature. Readers are referred to Durbin and Koopman (2001) for a review of available techniques. The multivariate form of a Gaussian state-space model is given below.

A Gaussian state-space model

Let $\mathbf{P}_t$ be the $q \times 1$ observation vector, $\mathbf{T}_t$ be the $p \times 1$ unobserved state vector and $\mathbf{F}_t = [\mathbf{F}_{1t}\ \mathbf{F}_{2t} \ldots \mathbf{F}_{kt}]^{\mathrm{T}}$ is a $k \times 1$ vector of known non-random exogenous variables; then a Gaussian state-space model can be defined as a system that consists of the following two equations.

$$\begin{aligned} \mathbf{P}_t &= \mathcal{A}\mathbf{T}_t + e_t, & e_t &\sim \mathrm{N}(0, \mathcal{R}), \\ \mathbf{T}_t &= \phi\mathbf{T}_{t-1} + \delta\mathbf{F}_t + w_t, & w_t &\sim \mathrm{N}(0, \mathcal{Q}), \end{aligned} \tag{3.1.2}$$

for $t = 1, 2, ..., N$, where $\mathcal{A}$ is a $q \times p$ observation matrix, ϕ is a $p \times p$ transition matrix, δ is a $p \times k$ matrix, $\mathcal{R}$ and $\mathcal{Q}$ are $q \times q$ and $p \times p$ covariance matrices respectively. The first equation in (3.1.2) is called the ***observation equation***, which defines the relationship between the observed and the state processes. The second equation is called the ***state equation*** which describes the dynamics of the state process. The error terms e_t and w_t are assumed to be serially independent and independent of each other at all time points. Furthermore, the initial state vector $\mathbf{T}_0$ is assumed to be $\mathrm{N}(\mu_0, \Sigma_0)$ and independent of e_t and w_t.

In the context of temperature reconstruction, one can view the hemispheric mean temperature as the unobserved state process $\{\mathbf{T}_t\}$ and the composite proxy record as the observation process $\{\mathbf{P}_t\}$. However, this ignores the fact that hemispheric mean temperature is observed over the calibration period. Modifications will be proposed in the latter part of this chapter as required. To describe the state-space representation of the hemispheric mean temperature, one must first introduce some notation and make certain assumptions. Thus, let $\mathbf{GS}_t$, $\mathbf{VOL}_t$ and $\mathbf{SOL}_t$ be the climate model estimated responses to greenhouse gas and sulphate aerosol forcing combined (GS), volcanic forcing (VOL) and solar forcing (SOL) at time t respectively. Such estimated responses are obtained using forced runs of a climate model. These forced runs are conducted by prescribing an estimated historical level of the specified external forcings during the analysis period. More details of this can be found in the latter part of this chapter.

One can then think of the hemispheric mean temperature $\mathbf{T}_t$ as being the sum of the response to external forcing plus the effect of internal variability. With these assumptions, a reasonable statistical model for $\mathbf{T}_t$ $(t = 1,\ 2,\ ...,\ n+m)$ might be

$$\begin{aligned} \mathbf{T}_t &= \delta_0 + \delta_{\mathbf{GS}}\mathbf{GS}_t + \delta_{\mathbf{VOL}}\mathbf{VOL}_t + \delta_{\mathbf{SOL}}\mathbf{SOL}_t + \mathbf{Z}_t \\ &= \delta\mathbf{X}_t + \mathbf{Z}_t \end{aligned} \tag{3.1.3}$$

where $\delta = [\delta_0 \quad \delta_{\mathbf{GS}} \quad \delta_{\mathbf{VOL}} \quad \delta_{\mathbf{SOL}}]$, $\mathbf{X}_t = [1 \quad \mathbf{GS}_t \quad \mathbf{VOL}_t \quad \mathbf{SOL}_t]^{\mathrm{T}}$, δ_0 is the mean of $\mathbf{T}_t$ when the climate system is not influenced by forcings, $\delta_{\mathbf{GS}}, \delta_{\mathbf{VOL}}$ and $\delta_{\mathbf{SOL}}$ are scaling factors that account for error in the magnitude of the estimated response to the specified external forcing and $\mathbf{Z}_t$ represents random variations resulting from internal variability. It is further assumed that $\mathbf{Z}_t$ is an AR(1) process with lag-one autocorrelation ϕ. Thus,

$$\mathbf{Z}_t = \sum_{j=0}^{\infty} \phi^j w_{t-j},$$

where the w_t's are defined as in (3.1.2).

The assumption that $\mathbf{Z}_t$ follows an AR(1) structure is justifiable by investigating the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the northern

hemispheric mean temperature series from a control run of a climate model (i.e., $\mathbf{X}_t = 1$). Figure 3.3 displays the ACF and PCAF of a 1000 years northern hemispheric mean temperature series obtained from a control run of the Canadian Centre for Climate Modelling and Analysis (CCCma) CGCM2 climate model (Flato and Boer 2001). The ACF tails off and the PACF cuts off roughly at lag 1. These behavior suggest that an AR(1) model should provide a reasonable approximation for the structure of natural internal variability.

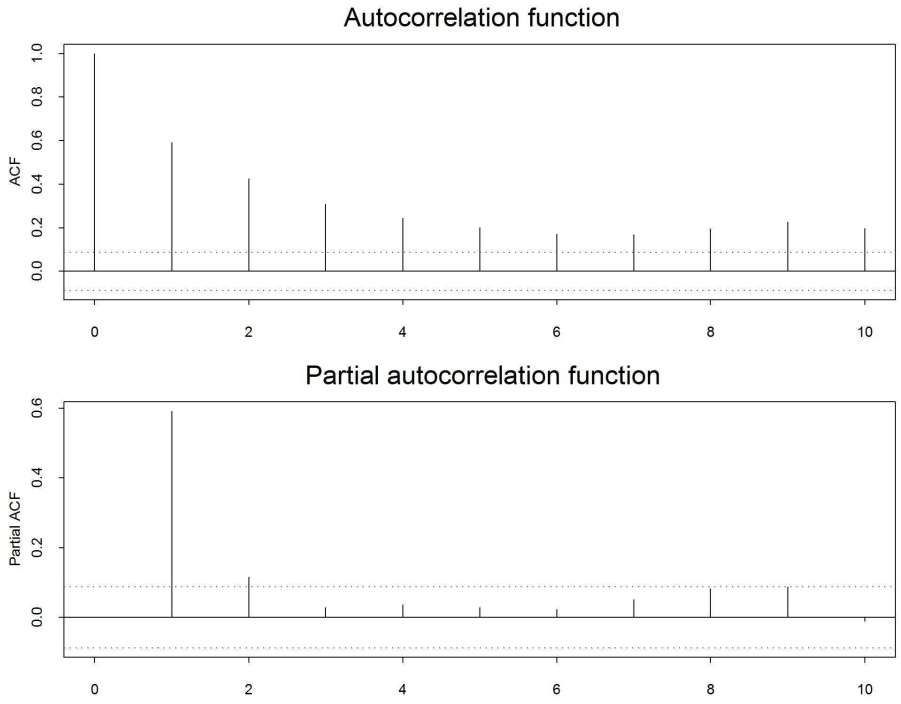

Figure 3.3: Autocorrelation and partial autocorrelation function of a 1000 years northern hemispheric mean temperature series obtained from a control run of CCCma CGCM2 climate model.

With these assumptions, a Gaussian state-space representation of the hemispheric mean temperature $\mathbf{T}_t$ ($t = 1,\ 2,\ ...,\ n + m$) can be given by the model in (3.1.2) with $p = q = 1$ and $\mathbf{F}_t = \mathbf{X}_t - \phi\mathbf{X}_{t-1}$, that is,

$$\begin{aligned} \mathbf{P}_t &= \mathcal{A}\mathbf{T}_t + e_t \\ \mathbf{T}_t &= \phi\mathbf{T}_{t-1} + \delta(\mathbf{X}_t - \phi\mathbf{X}_{t-1}) + w_t. \end{aligned} \tag{3.1.4}$$

The observation equation implies that the composite proxy series is simply a scale version of the unknown hemispheric mean temperature with additive white noise. On the other hand, the state equation models the dynamics of hemispheric mean temperature as an autoregressive process of order 1, i.e. AR(1), with known exogenous variables. It is obtained by considering the difference between $\mathbf{T}_t$ and $\phi\mathbf{T}_{t-1}$ using (3.1.3). The state equation in effect states that the rate of change in temperature depends upon the response to forcing which is governed by the forcing coefficients $\delta_{\mathrm{GS}}, \delta_{\mathrm{VOL}}$ and δ_{SOL}, and upon natural internal variability. Alternatively, one can also define a state equation that does not contain the response to forcings. This can be achieved by setting $\mathbf{X}_t = 1$ and $\delta = \delta_0$.

It should be noted that it is also possible to cast the reconstruction problem in terms of a simpler model, such as a linear regression model. An example of a linear regression model for the reconstruction problem can be defined as

$$\mathbf{P}_t = \mathcal{A}\mathbf{T}_t + e_t.$$

The unknown historical temperature is thus given by $\mathbf{P}_t/\mathcal{A}$, for $t = 1, 2, ...n$. Given that both the proxy series and the temperature record are available over the calibration period, one can estimate the parameter $\mathcal{A}$ by least square regression if it is unknown. Compared to the Gaussian state-space model, such a linear regression model does not contain the state equation, which describes the dynamics of the unknown temperature. The state equation can provide potentially valuable information because it allows the reconstruction to draw on two sources of information; the proxies and the exogenous variables. Also, such an approach permits one to carry out a detection analysis and provide projections for future climate. More details of this are given in Chapter 4.

From a practical point of view, the primary goal of any analysis involving the state-space model is to estimate the unobserved state process $\mathbf{T}_t$ based on $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_s$. When $s < t$, the problem is called ***prediction*** or ***forecasting***. When $s = t$, it is called ***filtering***. When $s > t$, it is called ***smoothing***. The solution to these problems can be obtained by applying the Kalman filter and smoother. The Kalman filter and smoother can be viewed as a set of

sequential updating equations that estimate the mean and variance of the distribution of the state variable $\mathbf{T}_t$ conditional on the observations. The Kalman filter and smoother, in fact, provide best linear estimators in the sense of minimum mean square error. Before stating the Kalman filter and smoother, the following definitions are given:

$$\begin{aligned}\mathbf{T}_{t|s} &\stackrel{def}{=} \mathrm{E}(\mathbf{T}_t|\mathfrak{B}_s)\\ \mathbf{S}_{t_1,t_2|s} &\stackrel{def}{=} \mathrm{E}\left[(\mathbf{T}_{t_1}-\mathbf{T}_{t_1|s})(\mathbf{T}_{t_2}-\mathbf{T}_{t_2|s})^{\mathrm{T}}\big|\mathfrak{B}_s\right],\end{aligned}$$

where $\mathfrak{B}_s$ is the set of past observations $\{\mathbf{P}_1, \mathbf{P}_2, ...\mathbf{P}_s, \mathfrak{B}_0\}$ and with $\mathfrak{B}_0$ defining the initial condition on $\mathbf{T}_0$. Note that $\mathfrak{B}_s = \mathfrak{B}_{s-1} \cup \mathbf{P}_s$. When $t_1 = t_2 = t$, $\mathbf{S}_{t_1,t_2|s}$ will be rewritten as $\mathbf{S}_{t|s}$. The Kalman filter, which was first proposed by Kalman (1960), will now be defined. The result will be stated in its multivariate form. The Gaussian assumption is essential in deriving the Kalman filter and smoother. Detail derivations are given in section 3.1.1.

Property 3.1 The Kalman filter

For the Gaussian state-space model in (3.1.2), with initial conditions $\mathbf{T}_{0|0} = \mu_0$ *and* $\mathbf{S}_{0|0} = \Sigma_0$*, the Kalman prediction estimates* $\mathbf{T}_{t|t-1}$*, for* $t = 1, 2, ..., N$*, are given recursively by,*

$$\begin{aligned}\mathbf{T}_{t|t-1} &= \phi\mathbf{T}_{t-1|t-1} + \delta\mathbf{F}_t\\ \mathbf{S}_{t|t-1} &= \phi\mathbf{S}_{t-1|t-1}\phi^{\mathrm{T}} + \mathcal{Q}\end{aligned} \tag{3.1.5}$$

and the Kalman filter estimates $\mathbf{T}_{t|t}$ *are given by*

$$\begin{aligned}\mathbf{T}_{t|t} &= \mathbf{T}_{t|t-1} + K_t(\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|t-1})\\ \mathbf{S}_{t|t} &= (\mathrm{I} - K_t\mathcal{A})\mathbf{S}_{t|t-1},\end{aligned} \tag{3.1.6}$$

where

$$K_t = \mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}}(\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R})^{-1}.$$

□

In the temperature reconstruction problem, $\mathbf{T}_t$ for $t = n+1, n+2, ..., N$ is actually known and hence one can choose to stop the recursion at $t = n$. However, running the recursion to $t = N$ can provide more information to estimate the unknown $\mathbf{T}_t$ for $t = 1, 2, ..., n$ when using the Kalman smoother which will be described next. The problem of estimating $\mathbf{T}_t$ conditioned on the data set $\mathbf{P}_1, \mathbf{P}_2, ...\mathbf{P}_N$ can be approached by the Kalman smoother.

Property 3.2 The Kalman smoother

For the state space-model in (3.1.2), with initial conditions $\mathbf{T}_{N|N}$ *and* $\mathbf{S}_{N|N}$ *obtained from Kalman filter, the Kalman smoother estimates* $\mathbf{T}_{t|N}$*, for* $t = N-1, N-2, ..., 0$*, are given by,*

$$\begin{aligned} \mathbf{T}_{t|N} &= \mathbf{T}_{t|t} + J_t(\mathbf{T}_{t+1|N} - \mathbf{T}_{t+1|t}) \\ \mathbf{S}_{t|N} &= \mathbf{S}_{t|t} + J_t(\mathbf{S}_{t+1|N} - \mathbf{S}_{t+1|t})J_t^{\mathrm{T}}, \end{aligned} \tag{3.1.7}$$

where

$$J_t = \mathbf{S}_{t|t}\phi^{\mathrm{T}}(\mathbf{S}_{t+1|t})^{-1}.$$

□

The Kalman smoother presented in Property 3.2 gives the recursion for $\mathbf{T}_{t|N} = \mathrm{E}(\mathbf{T}_t|\mathfrak{B}_N)$ and $\mathbf{S}_{t|N} = \mathrm{Var}(\mathbf{T}_t|\mathfrak{B}_N)$. The Kalman smoother assumes that the state process is completely unknown and thus the Kalman smoother estimate of $\mathbf{T}_t$ is based entirely on the observation process. For the reconstruction problem, one can modify the Kalman smoother to take account into the partially known state process. In particular, one can modify the Kalman smoother to provide the moments for $\mathbf{T}_t|(\mathfrak{B}_N, \mathbf{T}_{n+1:N})$. From Figure 3.2, it is clear that these moments are equivalent to the moments of $\mathbf{T}_t|(\mathfrak{B}_N, \mathbf{T}_{n+1})$. Such a modified Kalman smoother will be derived here and is given in the next property. Note that since $\mathbf{T}_t$ for $t = n+1, \cdots, N$ is assumed to be known, no smoother estimate are provided because, for $s > n$, the $\mathrm{E}(\mathbf{T}_s|\mathfrak{B}_N, \mathbf{T}_{n+1:N})$ is simply $\mathbf{T}_s$.

Property 3.3 The modified Kalman smoother for partially known state process *For the state-space model in (3.1.2), with* $\mathbf{T}_{t|N}$, $\mathbf{S}_{t|N}$ *and* J_t *obtained from the Kalman smoother, assume that* $\mathbf{T}_t$ *is observed for* $t = n+1, n+2, \cdots, N$, *the modified Kalman smoother estimates for such partially known state process, denoted by* $\widetilde{\mathbf{T}}_{t|N}$, *for* $t = n, n-1, \cdots, 0$, *are given by,*

$$\begin{aligned}
\widetilde{\mathbf{T}}_{t|N} \overset{def}{=} & \mathrm{E}\left[\mathbf{T}_t \middle| \mathfrak{B}_N, \mathbf{T}_{n+1:N}\right] \\
= & \mathrm{E}\left[\mathbf{T}_t \middle| \mathfrak{B}_N, \mathbf{T}_{n+1}\right] \\
= & \mathbf{T}_{t|N} + (J_t J_{t+1} \cdots J_n)(\mathbf{T}_{n+1} - \mathbf{T}_{n+1|N}), \\
\widetilde{\mathbf{S}}_{t|N} \overset{def}{=} & \mathrm{Var}\left[\mathbf{T}_t \middle| \mathfrak{B}_N, \mathbf{T}_{n+1:N}\right] \\
= & \mathrm{Var}\left[\mathbf{T}_t \middle| \mathfrak{B}_N, \mathbf{T}_{n+1}\right] \\
= & \mathbf{S}_{t|N} - (J_t J_{t+1} \cdots J_n)\mathbf{S}_{n+1|N}(J_t J_{t+1} \cdots J_n)^{\mathrm{T}}.
\end{aligned}$$

□

Proof: See section 3.1.1.

It is worth noting that the impact of $\mathbf{T}_{n+1}$ on $\widetilde{\mathbf{T}}_{t|N}$ depends highly on ϕ. In the case of a univariate state-space model, $J_t J_{t+1} \cdots J_n = \phi^{n-t+1} \prod_{i=t}^{n} (\mathbf{S}_{i|i}/\mathbf{S}_{i+1|i})$. Consider the extreme case where $\phi = 0$, then $J_t J_{t+1} \cdots J_n = 0$ and thus knowledge of the state process at time $n+1$ does not provide any additional information on the estimation of the unknown states. This is expected given that all the $\mathbf{T}_t$'s are independent when $\phi = 0$. On the other hand, since $|\phi| < 1$ for a stationary AR(1) process, $\phi^j \longrightarrow 0$ as $j \longrightarrow \infty$. This implies that the impact of $\mathbf{T}_{n+1}$ on $\widetilde{\mathbf{T}}_{t|N}$ decreases as the time lag between these two points increases. Therefore, if $|\phi|$ is not close to 1, the values obtained from the Kalman smoother and the modified Kalman smoother would not differ substantially, except during the few time steps prior to $t = n+1$.

The lag covariance smoother $\mathbf{S}_{t,r|N}$, for $t \neq r$, will be used in the latter part of this chapter when estimating the unknown state-space model parameters. The recursion for obtaining this smoother is given next.

Property 3.4 The lag covariance smoother

For the state-space model in (3.1.2), with J_t and $\mathbf{S}_{t|N}$ obtained from the Kalman smoother, the lag covariance smoother for $t > r$ is given by

$$\begin{aligned}\mathbf{S}_{t,r|N} &= \mathbf{S}_{t,r+1|N} J_r^{\mathrm{T}} \\ &= \mathbf{S}_{t|N}(J_r J_{r+1} \cdots J_{t-2} J_{t-1})^{\mathrm{T}}.\end{aligned}$$

□

3.1.1 Derivation of the Kalman filter and smoother

Derivation of the Kalman filter (Dethlefsen et al. 1997):

The Kalman filter can be obtained by induction on t. For $t = 0$, $\mathbf{T}_0|\mathfrak{B}_0 \sim \mathrm{N}(\mathbf{T}_{0|0} = \mu_0,\ \mathbf{S}_{0|0} = \Sigma_0)$ by the initial condition of the Gaussian state-space model. Now, assume that

$$\mathbf{T}_{t-1}|\mathfrak{B}_{t-1} \sim \mathrm{N}(\mathbf{T}_{t-1|t-1},\ \mathbf{S}_{t-1|t-1}).$$

Note that $\mathbf{T}_{t-1}$ can be written as

$$\begin{aligned}\mathbf{T}_{t-1} &= \phi \mathbf{T}_{t-2} + \delta \mathbf{F}_{t-1} + w_{t-1} \\ &= \phi(\phi \mathbf{T}_{t-3} + \delta \mathbf{F}_{t-2} + w_{t-2}) + \delta \mathbf{F}_{t-1} + w_{t-1} \\ &= \phi\phi(\phi \mathbf{T}_{t-4} + \delta \mathbf{F}_{t-3} + w_{t-3}) + \phi(\delta \mathbf{F}_{t-2} + w_{t-2}) + \delta \mathbf{F}_{t-1} + w_{t-1} \\ &\vdots\end{aligned}$$

Thus $\mathbf{T}_{t-1}$ is independent of w_t. Next, consider the case at time t. The state equation gives $\mathbf{T}_t = \phi \mathbf{T}_{t-1} + \delta \mathbf{F}_t + w_t$, where $w_t \sim \mathrm{N}(0, \mathcal{Q})$. Since w_t is $\mathrm{N}(0, \mathcal{Q})$, the distribution of $\mathbf{T}_t$ conditional of $\mathfrak{B}_{t-1}$ will be a sum of two independent normally distributed variables and hence

$$\mathbf{T}_t|\mathfrak{B}_{t-1} \sim \mathrm{N}(\phi \mathbf{T}_{t-1|t-1} + \delta \mathbf{F}_t,\ \phi \mathbf{S}_{t-1|t-1} \phi^{\mathrm{T}} + \mathcal{Q}) \stackrel{def}{=} \mathrm{N}(\mathbf{T}_{t|t-1},\ \mathbf{S}_{t|t-1}).$$

The above distribution gives the Kalman prediction estimates.

Now to find the Kalman filter estimate $\mathbf{T}_{t|t}$, one can calculate the joint distribution of

$(\mathbf{P}_t, \mathbf{T}_t)|\mathfrak{B}_{t-1}$ and then apply Theorem 3.6.1 in the appendix (section 3.6.3) to find the distribution of $\mathbf{T}_t|(\mathbf{P}_t, \mathfrak{B}_{t-1}) = \mathbf{T}_t|\mathfrak{B}_t$. To derive the joint distribution of $(\mathbf{P}_t, \mathbf{T}_t)|\mathfrak{B}_{t-1}$, note that $\mathbf{P}_t = \mathcal{A}\mathbf{T}_t + e_t$ and $\mathbf{T}_t$ is independent of e_t. The distribution of $\mathbf{P}_t$ conditional on $\mathfrak{B}_{t-1}$ is therefore

$$\mathbf{P}_t|\mathfrak{B}_{t-1} \sim \mathrm{N}(\mathcal{A}\mathbf{T}_{t|t-1},\ \mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R}). \tag{3.1.8}$$

Also note that any linear combination of $\mathbf{P}_t$ and $\mathbf{T}_t$ is a linear combination of $\mathbf{T}_t$ and e_t. Since $\mathbf{T}_t$ and e_t are mutually independent and are individually normal distributed when conditioning on $\mathfrak{B}_{t-1}$, the distribution of any linear combination of $\mathbf{P}_t$ and $\mathbf{T}_t$ conditional on $\mathfrak{B}_{t-1}$ will also be normal. By Theorem 3.6.2 (section 3.6.3), this implies that the joint distribution of $(\mathbf{P}_t, \mathbf{T}_t)|\mathfrak{B}_{t-1}$ is also a normal distribution. Hence it follows that

$$\begin{bmatrix} \mathbf{T}_t \\ \mathbf{P}_t \end{bmatrix} \Bigg| \mathfrak{B}_{t-1} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{T}_{t|t-1} \\ \mathcal{A}\mathbf{T}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \mathbf{S}_{t|t-1} & \mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} \\ \mathcal{A}\mathbf{S}_{t|t-1} & \mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R} \end{bmatrix} \right)$$

where the cross-covariance matrix between $\mathbf{T}_t$ and $\mathbf{P}_t$ conditional on $\mathfrak{B}_{t-1}$ is given by

$$\begin{aligned} \mathrm{Cov}(\mathbf{T}_t, \mathbf{P}_t|\mathfrak{B}_{t-1}) &= \mathrm{Cov}(\mathbf{T}_t, \mathcal{A}\mathbf{T}_t + e_t|\mathfrak{B}_{t-1}) \\ &= \mathcal{A}\mathrm{Var}(\mathbf{T}_t|\mathfrak{B}_{t-1}) \\ &= \mathcal{A}\mathbf{S}_{t|t-1}. \end{aligned}$$

Thus, by Theorem 3.6.1, one has

$$\mathbf{T}_t|(\mathbf{P}_t, \mathfrak{B}_{t-1}) = \mathbf{T}_t|\mathfrak{B}_t \sim \mathrm{N}(\mathbf{T}_{t|t},\ \mathbf{S}_{t|t})$$

where $\mathbf{T}_{t|t}$ and $\mathbf{S}_{t|t}$ are the Kalman filter estimates given by

$$\begin{aligned} \mathbf{T}_{t|t} &= \mathbf{T}_{t|t-1} + \mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} \left(\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R}\right)^{-1} (\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|t-1}) \\ &= \mathbf{T}_{t|t-1} + K_t(\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|t-1}) \end{aligned}$$

with

$$K_t = \mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}}(\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R})^{-1}$$

and

$$\begin{aligned}\mathbf{S}_{t|t} &= \mathbf{S}_{t|t-1} - \mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}}(\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R})^{-1}\mathcal{A}\mathbf{S}_{t|t-1} \\ &= \mathbf{S}_{t|t-1} - K_t\mathcal{A}\mathbf{S}_{t|t-1} \\ &= (\mathbf{I} - K_t\mathcal{A})\mathbf{S}_{t|t-1}.\end{aligned}$$

■

Derivation of the Kalman smoother (Dethlefsen et al. 1997):

The Kalman smoother can be obtained by backward induction on t. From the Kalman filter, one has $\mathbf{T}_N|\mathfrak{B}_N \sim \mathrm{N}(\mathbf{T}_{N|N},\ \mathbf{S}_{N|N})$. Now, assume that, for $t < N$,

$$\mathbf{T}_{t+1}|\mathfrak{B}_N \sim \mathrm{N}(\mathbf{T}_{t+1|N},\ \mathbf{S}_{t+1|N}).$$

Next, consider the case at time t. From the Kalman filter, one has

$$\begin{aligned}\mathbf{T}_{t+1}|\mathfrak{B}_t &\sim \mathrm{N}(\mathbf{T}_{t+1|t},\ \mathbf{S}_{t+1|t}) \\ \mathbf{T}_t|\mathfrak{B}_t &\sim \mathrm{N}(\mathbf{T}_{t|t},\ \mathbf{S}_{t|t}).\end{aligned}$$

Since $\mathbf{T}_{t+1} = \phi\mathbf{T}_t + \delta\mathbf{F}_{t+1} + w_{t+1}$ is a linear function of $\mathbf{T}_t$ and w_{t+1}, then any linear combination of $\mathbf{T}_{t+1}$ and $\mathbf{T}_t$ will be a function of $\mathbf{T}_t$ and w_{t+1}. Since $\mathbf{T}_t$ and w_{t+1} are independent and are individually normal distributed when conditioning on $\mathfrak{B}_t$, any linear combination of $\mathbf{T}_{t+1}$ and $\mathbf{T}_t$ conditional on B_t will be normally distributed. Hence, by Theorem 3.6.2,

$$\left.\begin{bmatrix}\mathbf{T}_{t+1} \\ \mathbf{T}_t\end{bmatrix}\right|\mathfrak{B}_t \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{T}_{t+1|t} \\ \mathbf{T}_{t|t}\end{bmatrix}, \begin{bmatrix}\mathbf{S}_{t+1|t} & \phi\mathbf{S}_{t|t} \\ \mathbf{S}_{t|t}\phi^{\mathrm{T}} & \mathbf{S}_{t|t}\end{bmatrix}\right),$$

where the cross-covariance matrix between $\mathbf{T}_{t+1}$ and $\mathbf{T}_t$ conditional on $\mathfrak{B}_t$ is given by

$$\begin{aligned}\mathrm{Cov}(\mathbf{T}_{t+1}, \mathbf{T}_t|\mathfrak{B}_t) &= \mathrm{Cov}(\phi\mathbf{T}_t + \delta\mathbf{F}_{t+1} + w_{t+1}, \mathbf{T}_t|\mathfrak{B}_t) \\ &= \mathrm{Cov}(\phi\mathbf{T}_t, \mathbf{T}_t|\mathfrak{B}_t) \\ &= \phi\mathbf{S}_{t|t}.\end{aligned}$$

By Theorem 3.6.1,

$$\begin{aligned}\mathbf{T}_t|(\mathbf{T}_{t+1},\mathfrak{B}_t) &\sim \mathrm{N}\left(\mathbf{T}_{t|t}+\mathbf{S}_{t|t}\phi^{\mathrm{T}}\mathbf{S}^{-1}_{t+1|t}(\mathbf{T}_{t+1}-\mathbf{T}_{t+1|t}),\ \mathbf{S}_{t|t}-\mathbf{S}_{t|t}\phi^{\mathrm{T}}\mathbf{S}^{-1}_{t+1|t}\phi\mathbf{S}_{t|t}\right)\\ &= \mathrm{N}\left(\mathbf{T}_{t|t}+J_t(\mathbf{T}_{t+1}-\mathbf{T}_{t+1|t}),\ \mathbf{S}_{t|t}-J_t\mathbf{S}_{t+1|t}J_t^{\mathrm{T}}\right)\end{aligned} \tag{3.1.9}$$

From Figure 3.2, it follows that this distribution is equivalent to the distribution of $\mathbf{T}_t|(\mathbf{T}_{t+1},\mathfrak{B}_N)$. Since $\mathbf{T}_{t+1}|\mathfrak{B}_N$ is normally distributed by the induction assumption, one has, by Theorem 3.6.3 (section 3.6.3),

$$\mathbf{T}_t|\mathfrak{B}_N \sim \mathrm{N}\left(\mathbf{T}_{t|t}+J_t(\mathbf{T}_{t+1|N}-\mathbf{T}_{t+1|t}),\ \mathbf{S}_{t|N}\right)$$

where

$$\begin{aligned}\mathbf{S}_{t|N} &= \mathbf{S}_{t|t}-J_t\mathbf{S}_{t+1|t}J_t^{\mathrm{T}}+J_t\mathbf{S}_{t+1|N}J_t^{\mathrm{T}}\\ &= \mathbf{S}_{t|t}+J_t(\mathbf{S}_{t+1|N}-\mathbf{S}_{t+1|t})J_t^{\mathrm{T}}.\end{aligned}$$

■

Derivation of the modified Kalman smoother:

Obtaining the modified Kalman smoother that takes account into the fact that $\mathbf{T}_{n+1},...,\mathbf{T}_N$ are known involves deriving the joint distribution of $(\mathbf{T}_t,\mathbf{T}_{n+1})|\mathfrak{B}_N$ and subsequently applying Theorem 3.6.1 to obtain the $\mathrm{E}(\mathbf{T}_t|\mathfrak{B}_N,\mathbf{T}_{n+1})$ and $\mathrm{Var}(\mathbf{T}_t|\mathfrak{B}_N,\mathbf{T}_{n+1})$. As pointed out earlier, the moments of $\mathbf{T}_t|(\mathfrak{B}_N,\mathbf{T}_{n+1})$ are equivalent to that of $\mathbf{T}_t|(\mathfrak{B}_N,\mathbf{T}_{n+1:N})$. For now, pretend that the $\{\mathbf{T}_t\}$ process is completely unknown. From the the Kalman smoother, one has,

$$\begin{aligned}\mathbf{T}_{n+1}|\mathfrak{B}_N &\sim \mathrm{N}(\mathbf{T}_{n+1|N},\mathbf{S}_{n+1|N})\\ \mathbf{T}_t|\mathfrak{B}_N &\sim \mathrm{N}(\mathbf{T}_{t|N},\mathbf{S}_{t|N}).\end{aligned} \tag{3.1.10}$$

Note that $\mathbf{T}_{n+1}$ can be written as

$$\begin{aligned}\mathbf{T}_{n+1} &= \phi\mathbf{T}_n+\delta\mathbf{F}_{n+1}+w_{n+1}\\ &= \phi(\phi\mathbf{T}_{n-1}+\delta\mathbf{F}_n+w_n)+\delta\mathbf{F}_{n+1}+w_{n+1}\\ &= \phi(\phi\mathbf{T}_{n-2}+\delta\mathbf{F}_{n-1}+w_{n-1})+\phi(\delta\mathbf{F}_n+w_n)+\delta\mathbf{F}_{n+1})+\delta\mathbf{F}_{n+1}+w_{n+1}\\ &\vdots\end{aligned}$$

Thus, $\mathbf{T}_{n+1}$ can be expressed as a function of $\mathbf{T}_t$ and $w_{t+1}, w_{t+2}, \cdots, w_{n+1}$. Therefore, any linear combination of $\mathbf{T}_t$ and $\mathbf{T}_{n+1}$ will be functions of $\mathbf{T}_t$ and $w_{t+1}, w_{t+2}, \cdots, w_{n+1}$. Since $\mathbf{T}_t$ is independent of $w_{t+1}, w_{t+2}, \cdots, w_{n+1}$, any linear combination of $\mathbf{T}_t$ and $\mathbf{T}_{n+1}$ conditional on $\mathfrak{B}_N$ will be normal. By Theorem 3.6.2, this implies that ,

$$\left[\begin{array}{c} \mathbf{T}_{n+1} \\ \mathbf{T}_t \end{array}\right] \Bigg| \mathfrak{B}_N \sim \mathrm{N}\left(\left[\begin{array}{c} \mathbf{T}_{n+1|N} \\ \mathbf{T}_{t|N} \end{array}\right], \left[\begin{array}{cc} \mathbf{S}_{n+1|N} & \mathbf{S}_{n+1,t+1|N} J_t^{\mathrm{T}} \\ J_t \mathbf{S}_{n+1,t+1|N}^{\mathrm{T}} & \mathbf{S}_{t|N} \end{array}\right]\right),$$

where the cross-covariance matrix between $\mathbf{T}_{n+1}$ and $\mathbf{T}_t$ conditional on $\mathfrak{B}_N$ is given by Property 3.4, which is

$$\mathrm{Cov}(\mathbf{T}_{n+1}, \mathbf{T}_t | \mathfrak{B}_N) = \mathbf{S}_{n+1,t+1|N} J_t^{\mathrm{T}} = \mathbf{S}_{n+1|N} (J_t J_{t+1} \cdots J_n)^{\mathrm{T}}.$$

If $\mathbf{T}_{n+1}$ is observed, by Theorem 3.6.1, one can write,

$$\mathbf{T}_t \big| (\mathfrak{B}_N, \mathbf{T}_{n+1}) \sim \mathrm{N}(\widetilde{\mathbf{T}}_{t|N}, \widetilde{\mathbf{S}}_{t|N}),$$

where

$$\begin{aligned}
\widetilde{\mathbf{T}}_{t|N} &= \mathbf{T}_{t|N} + (J_t J_{t+1} \cdots J_n) \mathbf{S}_{n+1|N} \mathbf{S}_{n+1|N}^{-1} (\mathbf{T}_{n+1} - \mathbf{T}_{n+1|N}) \\
&= \mathbf{T}_{t|N} + (J_t J_{t+1} \cdots J_n)(\mathbf{T}_{n+1} - \mathbf{T}_{n+1|N}), \\
\widetilde{\mathbf{S}}_{t|N} &= \mathbf{S}_{t|N} - (J_t J_{t+1} \cdots J_n) \mathbf{S}_{n+1|N} \mathbf{S}_{n+1|N}^{-1} \mathbf{S}_{n+1|N} (J_t J_{t+1} \cdots J_n)^{\mathrm{T}} \\
&= \mathbf{S}_{t|N} - (J_t J_{t+1} \cdots J_n) \mathbf{S}_{n+1|N} (J_t J_{t+1} \cdots J_n)^{\mathrm{T}}.
\end{aligned}$$

■

Derivation of the lag covariance smoother (Dethlefsen et al. 1997):

The derivation follows from the well known result that $\mathrm{Var}(\mathbf{X}) = \mathrm{E}[\mathrm{Var}(\mathbf{X}|\mathbf{Y})] + \mathrm{Var}[\mathrm{E}(\mathbf{X}|\mathbf{Y})]$.

$$
\begin{aligned}
\mathbf{S}_{t,r|N} &= \mathrm{E}\left[(\mathbf{T}_t - \mathbf{T}_{t|N})(\mathbf{T}_r - \mathbf{T}_{r|N})^{\mathrm{T}} \middle| \mathfrak{B}_N\right] \\
&= \mathrm{Cov}(\mathbf{T}_t, \mathbf{T}_r | \mathfrak{B}_N) \\
&= \mathrm{E}\left[\mathrm{Cov}(\mathbf{T}_t, \mathbf{T}_r | \mathbf{T}_{r+1}, \mathbf{T}_t, \mathfrak{B}_N) \middle| \mathfrak{B}_N\right] \\
&\quad + \mathrm{Cov}\left[\mathrm{E}(\mathbf{T}_t | \mathbf{T}_{r+1}, \mathbf{T}_t, \mathfrak{B}_N), \mathrm{E}(\mathbf{T}_r | \mathbf{T}_{r+1}, \mathbf{T}_t, \mathfrak{B}_N) \middle| \mathfrak{B}_N\right] \\
&= \mathrm{Cov}\left[\mathrm{E}(\mathbf{T}_t | \mathbf{T}_{r+1}, \mathbf{T}_t, \mathfrak{B}_N), \mathrm{E}(\mathbf{T}_r | \mathbf{T}_{r+1}, \mathbf{T}_t, \mathfrak{B}_N) \middle| \mathfrak{B}_N\right] \\
&= \mathrm{Cov}\left[\mathbf{T}_t, \mathrm{E}(\mathbf{T}_r | \mathbf{T}_{r+1}, \mathfrak{B}_N) \middle| \mathfrak{B}_N\right] \\
&= \mathrm{Cov}\left(\mathbf{T}_t, \mathbf{T}_{r|r} + J_r(\mathbf{T}_{r+1} - \mathbf{T}_{r+1|r}) \middle| \mathfrak{B}_N\right) \\
&= \mathrm{Cov}(\mathbf{T}_t, J_r \mathbf{T}_{r+1} | \mathfrak{B}_N) \\
&= \mathbf{S}_{t,r+1|N} J_r^{\mathrm{T}} \\
&= \mathbf{S}_{t|N} J_{t-1}^{\mathrm{T}} J_{t-2}^{\mathrm{T}} \cdots J_{r+1}^{\mathrm{T}} J_r^{\mathrm{T}}
\end{aligned}
$$

The fifth equality follows because from Figure 3.2, one can see that $\mathrm{E}(\mathbf{T}_r | \mathbf{T}_{r+1}, \mathbf{T}_t, \mathfrak{B}_N) = \mathrm{E}(\mathbf{T}_r | \mathbf{T}_{r+1}, \mathfrak{B}_N)$, for $r < t$. The sixth equality follows from (3.1.9).

■

3.2 Maximum likelihood estimation for a state-space model with an unknown state process

A difficulty in using the Kalman filter and smoother arises when the parameters $\mathcal{A}, \mathcal{R}, \phi, \delta, \mathcal{Q}$, μ_0, Σ_0 that specify the state-space model in (3.1.2) are unknown. Hence, one will need to estimate the unknown parameters before using the Kalman filter and smoother to estimate the state process. Estimation of these parameters is quite involved. Furthermore, the existing estimation methods do not consider the case where the state process is partially known. In this section, the existing estimation approach will first be presented. Extension to the case where the state process is partially known will be considered in the next section. For the rest of this chapter, emphasis will be placed on parameter estimation approaches. Application to the reconstruction problem will be the topic of Chapter 4.

Note that an assumption of the Gaussian state-space model is that the exogenous variable $\mathbf{F}_t$ is known and non-random. If ϕ is unknown, the exogenous variables therefore cannot contain the parameter ϕ. Thus, in the state-space representation of the hemispheric mean temperature in (3.1.4), one needs to fix the parameter ϕ that is involved in $\mathbf{F}_t = \mathbf{X}_t - \phi\mathbf{X}_{t-1}$ before estimating the other parameters. More details on the choice of ϕ are given in the next chapter. For the purpose of this chapter, it is assumed that $\mathbf{F}_t$ is known and is free of unknown parameters.

3.2.1 The $\mathbb{PXY}$ approach

The existing parameter estimation approach (Schweppe 1965) utilizes the method of maximum likelihood. Under the assumption that the initial state $\mathbf{T}_0$ is $\mathrm{N}(\mu_0, \Sigma_0)$, one can write the likelihood function for the data $\{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_N\}$ as,

$$\begin{aligned} \mathcal{L}_{\mathbf{P}}(\Theta) &= f_\Theta(\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_N) \\ &= \prod_{t=1}^{N} f_\Theta(\mathbf{P}_t | \mathbf{P}_{1:t-1}). \end{aligned} \tag{3.2.1}$$

where $\Theta = \{\mathcal{R}, \phi, \delta, \mathcal{Q}, \mu_0, \Sigma_0\}$. The ***likelihood function*** describes how likely the observed data is as a function of the parameter vector Θ. Maximizing the likelihood function with

respect to Θ gives the Θ value that agrees most closely with the observed data.

It is important to point out that the parameter $\mathcal{A}$ is not included in Θ. In existing applications, the parameter $\mathcal{A}$ is usually considered to be a known design matrix and thus no estimation is required. In fact, if the state process $\{\mathbf{T}_t\}$ is completely unknown, the parameter $\mathcal{Q}$ is confounded with $\mathcal{A}$, ϕ and δ. Thus the parameters $\mathcal{A}$, ϕ, δ and $\mathcal{Q}$ cannot be estimated simultaneously. It turns out that fixing either $\mathcal{A}$, δ or $\mathcal{Q}$ resolves the ill-posed estimation problem. To explain this, recall that our Gaussian state-space model is given by

$$\begin{aligned} \mathbf{P}_t &= \mathcal{A}\mathbf{T}_t + e_t, & e_t &\sim \mathrm{N}(0, \mathcal{R}), \\ \mathbf{T}_t &= \phi\mathbf{T}_{t-1} + \delta\mathbf{F}_t + w_t, & w_t &\sim \mathrm{N}(0, \mathcal{Q}). \end{aligned}$$

Note that there is degeneracy in the model in the sense that one can represent the above model using fewer number of parameters by moving all the structure of matrix $\mathcal{Q}$ into the matrices $\mathcal{A}$, ϕ and δ. This is done by standardizing the unknown state process to obtain a model in which the matrix $\mathcal{Q}$ is identity. This corresponds to rewriting the state equation as,

$$\mathcal{Q}^{-1/2}\mathbf{T}_t = \mathcal{Q}^{-1/2}\phi\mathbf{T}_{t-1} + \mathcal{Q}^{-1/2}\delta\mathbf{F}_t + z_t, \quad z_t \sim \mathrm{N}(0, \mathbf{I}),$$

where $z_t = \mathcal{Q}^{-1/2}w_t$, $\mathbf{I}$ is an identity matrix, $\mathcal{Q}^{1/2} = \mathcal{C}\mathcal{D}^{1/2}\mathcal{C}^{\mathrm{T}}$, matrix $\mathcal{C}$ contains the normalized eigenvectors of $\mathcal{Q}$ and $\mathcal{D}$ is a diagonal matrix of the eigenvalues of $\mathcal{Q}$. By defining a new state vector $\ddot{\mathbf{T}}_t = \mathcal{Q}^{-1/2}\mathbf{T}_t$, one can then rewrite the Gaussian state space model above as,

$$\begin{aligned} \mathbf{P}_t &= \ddot{\mathcal{A}}\ddot{\mathbf{T}}_t + e_t, & e_t &\sim \mathrm{N}(0, \mathcal{R}), \\ \ddot{\mathbf{T}}_t &= \ddot{\phi}\ddot{\mathbf{T}}_{t-1} + \ddot{\delta}\mathbf{F}_t + z_t, & z_t &\sim \mathrm{N}(0, \mathbf{I}), \end{aligned} \tag{3.2.2}$$

where

$$\ddot{\mathcal{A}} = \mathcal{A}\mathcal{Q}^{1/2}, \quad \ddot{\phi} = \mathcal{Q}^{-1/2}\phi\mathcal{Q}^{1/2} \quad \text{and} \quad \ddot{\delta} = \mathcal{Q}^{-1/2}\delta.$$

The above model is in fact exactly equivalent to the original state-space model because the $\{\mathbf{T}_t\}$ process is unknown and thus one is free to standardize it. So it is now clear that the parameter $\mathcal{Q}$ is confounded with $\mathcal{A}$, ϕ and δ and therefore these parameters cannot be

estimated simultaneously. On the other hand, the observation process cannot be standardized in the same way when the parameter $\mathcal{R}$ is unknown. This is because such standardized model violates the state-space model assumption in that the observation process $\{R^{-1/2}\mathbf{P}_t\}$ is unknown.

Fixing either the parameter $\mathcal{A}$, δ or $\mathcal{Q}$ resolves the confounding problem. Note that fixing the parameter ϕ does not resolve the problem because one cannot solve for $\mathcal{Q}$ from the expression for $\ddot{\phi}$ even when ϕ is known. The parameter $\mathcal{A}$ in existing applications of state-space models is usually a known design matrix which resolves the ill-posed estimation problem. See Shumway and Stoffer (2000, Ch. 4) for examples of applications. In the temperature reconstruction problem, none of the parameters are known. Following the existing parameter estimation approach, one can fix the parameter $\mathcal{A}$ in the reconstruction problem and the other parameters can then be estimated through maximum likelihood estimation. A reasonable choice for the parameter $\mathcal{A}$ can be obtained by least squares regression using the calibration data set in which the regression model is the observation equation in (3.1.2).

Now, let's return to the likelihood function in (3.2.1). Note that maximizing $\ln \mathcal{L}_{\mathbf{P}}(\Theta)$ with respect to Θ is equivalent to maximizing $\mathcal{L}_{\mathbf{P}}(\Theta)$. From the derivation of the Kalman filter (equation (3.1.8) in section 3.1.1),

$$\mathbf{P}_t|\mathbf{P}_{1:t-1} \sim \mathrm{N}(\mathcal{A}\mathbf{T}_{t|t-1}, \mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R}).$$

Hence, ignoring a constant, the log of the likelihood, $\ell_{\mathbf{P}}(\Theta)$, can be written as

$$\begin{aligned} \ell_{\mathbf{P}}(\Theta) = &- 0.5 \sum_{t=1}^{N} \ln\left|\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R}\right| \\ &- 0.5 \sum_{t=1}^{N} (\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|t-1})^{\mathrm{T}} (\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}} + \mathcal{R})^{-1} (\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|t-1}), \end{aligned} \tag{3.2.3}$$

where $\mathbf{T}_{t|t-1}$ and $\mathbf{S}_{t|t-1}$ are functions of Θ and are defined as in Property 3.1. This log likelihood function will be denoted as the proxy only ($\mathbb{PXY}$) log likelihood and the approach of parameter estimation through maximizing this function will be termed as the $\mathbb{PXY}$ approach. The log likelihood function in (3.2.3) is a highly nonlinear function of the unknown parameters

and hence numerical procedures are needed to find the optimal Θ value. The details are given in the next section.

The asymptotic distribution of the maximum likelihood estimator for the parameters of a state-space model has been considered by Caines (1988) and Jensen and Petersen (1999). Caines (1988) gave a proof of the asymptotic result for the Gaussian state-space model in (3.1.2). Jensen and Petersen (1999) provided the asymptotic properties for estimators under the assumption that both the observation and state processes are stationary but not necessary normally distributed. The result from Caines (1988) will be stated next. The proof is rather long and cumbersome. Readers are referred to Caines (1988, Ch. 7) for details.

Theorem 3.2.1 *Let $\hat{\Theta}$ be the estimator of the true $\Theta = \Theta_0$ obtained by maximizing $\mathcal{L}_{\mathbf{P}}(\Theta)$ in (3.2.1). Under general assumptions about the stability of the state-space model over time, as $N \longrightarrow \infty$,*

$$\sqrt{N}(\hat{\Theta} - \Theta_0) \longrightarrow N(0, \mathfrak{I}(\Theta_0)^{-1})$$

where $\mathfrak{I}(\Theta)$ is given by

$$\mathfrak{I}(\Theta) = N^{-1}\mathrm{E}\left[-\partial^2 \ell_{\mathbf{P}}(\Theta)/(\partial\Theta\ \partial\Theta^{\mathrm{T}})\right].$$

□

In practice, $\mathfrak{I}(\Theta_0)$ is unknown and thus it is estimated by the negative of the Hessian matrix of the log likelihood function $\ell_{\mathbf{P}}(\hat{\Theta})$ divided by N. An approximate confidence region for the estimator can then be obtained using such an estimate.

3.2.2 Numerical estimation procedure

Newton-Raphson algorithm and expectation maximization (EM) algorithm are two procedures that have been proposed for finding the optimal Θ value of the likelihood function of a state-space model. In the case of Newton-Raphson, the usual procedure is to fix μ_0 and Σ_0 and then estimate the other parameters. The steps involved in using the Newton-Raphson method to estimate the parameter $\Theta = \{\mathcal{R}, \phi, \delta, \mathcal{Q}\}$ are given below.

Newton-Raphson estimation procedure

1. Fix μ_0 and Σ_0. Select the starting value for the parameter, say $\Theta^{(0)}$. On iteration j (j=1, 2, ...):
2. Use Newton-Raphson method to obtain new estimate of the parameters, say $\Theta^{(j)}$.
3. Repeat step 2 until the parameters converge or the log likelihood function stabilizes.

There are several disadvantages with the Newton Raphson estimation procedure. First, this method requires the calculation of the derivatives of the likelihood function at each iteration. Given the complexity of the likelihood function, numerical estimation is required to estimate the derivatives. Furthermore, the successive steps involved in a Newton-Raphson procedure may not necessarily increase the value of the likelihood. In particular, since the likelihood function is highly non-linear, there is a danger that the Newton-Raphson procedure will converge to a local minimum.

Shumway and Stoffer (1982) presented an alternative estimation procedure that is based on the EM algorithm (Dempster et al. 1977). The EM algorithm is an iterative algorithm to obtain maximum likelihood estimate. In order to describe the estimation procedure in the case of state-space model, let's first consider the fundamental ideas behind the EM algorithm. Let $\mathbf{Y}$ denoted a vector of observed data with density function $f_\Theta(\mathbf{Y})$, where Θ is a parameter vector which is unknown and to be estimated. Let $\mathbf{X}$ be a vector of unobserved data, so the complete data $\{\mathbf{Y}, \mathbf{X}\}$ has a joint density function $f_\Theta(\mathbf{Y}, \mathbf{X})$. Now consider rewriting the density function, or equivalently the likelihood function of $\mathbf{Y}$, as

$$\begin{aligned}\mathcal{L}_\mathbf{Y}(\Theta) &= f_\Theta(\mathbf{Y}) \\ &= \frac{f_\Theta(\mathbf{Y}) f_\Theta(\mathbf{Y}, \mathbf{X})}{f_\Theta(\mathbf{Y}, \mathbf{X})} \\ &= \frac{f_\Theta(\mathbf{Y}, \mathbf{X})}{f_\Theta(\mathbf{X} \mid \mathbf{Y})}.\end{aligned}$$

Taking logarithms of the above expression, one has

$$\ell_\mathbf{Y}(\Theta) = \ln f_\Theta(\mathbf{Y}, \mathbf{X}) - \ln f_\Theta(\mathbf{X} \mid \mathbf{Y}).$$

Since the EM algorithm is iterative, let's assume one has a temporary value for the parameter Θ, denoted by Θ^*. Now, multiply both sides of the above equation by $f_{\Theta^*}(\mathbf{X} \mid \mathbf{Y})$ and integrate with respect to $\mathbf{X}$ to obtain

$$\begin{aligned} \ell_{\mathbf{Y}}(\Theta) &= \int \ln f_{\Theta}(\mathbf{Y}, \mathbf{X}) f_{\Theta^*}(\mathbf{X} \mid \mathbf{Y}) d\mathbf{X} - \int \ln f_{\Theta}(\mathbf{X} \mid \mathbf{Y}) f_{\Theta^*}(\mathbf{X} \mid \mathbf{Y}) d\mathbf{X} \\ &= H(\Theta \mid \Theta^*) - Z(\Theta \mid \Theta^*) \end{aligned} \tag{3.2.4}$$

where

$$H(\Theta \mid \Theta^*) = \mathrm{E}_{\mathbf{X}}[\ln f_{\Theta}(\mathbf{Y}, \mathbf{X}) | \mathbf{Y}, \Theta^*]$$

$$Z(\Theta \mid \Theta^*) = \mathrm{E}_{\mathbf{X}}[\ln f_{\Theta}(\mathbf{X} | \mathbf{Y}) | \mathbf{Y}, \Theta^*].$$

The goal here is to find the value of the parameter Θ which maximizes the log likelihood function $\ell_{\mathbf{Y}}(\Theta)$. The key result of the EM algorithm is that in order to maximize $\ell_{\mathbf{Y}}(\Theta)$, one only needs to maximize $H(\Theta \mid \Theta^*)$ iteratively. It turns out that the log likelihood function is non-decreasing at each step of the maximization of $H(\Theta \mid \Theta^*)$. Hence, one can ignore $Z(\Theta \mid \Theta^*)$ when maximizing the log likelihood function. See Theorem 3.6.5 in section 3.6.3 for details. The steps involved in using the EM algorithm to estimate the parameter Θ are as follows.

EM algorithm estimation procedure

1. Select an initial value for the parameter Θ, say $\Theta^{(0)}$. On iteration j (j=1, 2, ...):
2. Maximize $H(\Theta \mid \Theta^{(j-1)})$ with respect to Θ and let such Θ be $\Theta^{(j)}$, i.e., $H(\Theta^{(j)} \mid \Theta^{(j-1)}) \geq H(\Theta \mid \Theta^{(j-1)})$.
3. Repeat step 2 until the log likelihood function stabilizes.

Under mild regularity conditions, one is guaranteed convergence to a global or local maximum, or a saddle point (Wu 1983). If there is more than one maximum or saddle point for the likelihood function, then the choice of initial value $\Theta^{(0)}$ will be influential on the point that the algorithm converges to. For estimating the parameters in the state-space model, use of the EM algorithm implies that in order to maximize the log likelihood function $\ell_{\mathbf{P}}(\Theta)$ in

(3.2.3), one can successively maximize

$$H\left(\Theta \mid \Theta^{(j-1)}\right) = \mathrm{E}_{\mathbf{T}}\left[\ln f_{\Theta}(\mathbf{P}_{1:N}, \mathbf{T}_{0:N}) \middle| \mathbf{P}_{1:N}, \Theta^{(j-1)}\right],$$

where $f_{\Theta}(\mathbf{P}_{1:N}, \mathbf{T}_{0:N})$ is the joint density function of the complete data $(\mathbf{P}_{1:N}, \mathbf{T}_{0:N})$. After some manipulation, it can be shown that the Θ value that maximizes $H\left(\Theta \mid \Theta^{(j-1)}\right)$ at iteration j of step 2 of the EM algorithm estimation procedure is given by (detail derivations are given in section 3.6.1)

$$\begin{aligned}
\mathcal{R}^{(j)} &= N^{-1} \sum_{t=1}^{N} \left[\mathcal{A}\mathbf{S}_{t|N}\mathcal{A}^{\mathrm{T}} + (\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|N})(\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|N})^{\mathrm{T}}\right] \\
\delta^{(j)} &= \left(F_{11}^{\mathrm{T}} - S_{10}S_{00}^{-1}F_{10}^{\mathrm{T}}\right)\left(F_{00} - F_{10}S_{00}^{-1}F_{10}^{\mathrm{T}}\right)^{-1} \\
\phi^{(j)} &= \left(S_{10} - \delta^{(j)}F_{10}\right)S_{00}^{-1} \\
\mathcal{Q}^{(j)} &= N^{-1}\left(S_{11} - \phi^{(j)}S_{10}^{\mathrm{T}} - \delta^{(j)}F_{11}\right) \\
\mu_0^{(j)} &= \mathbf{T}_{0|N},
\end{aligned} \tag{3.2.5}$$

where

$$\begin{aligned}
S_{00} &= \sum_{t=1}^{N}\left(\mathbf{T}_{t-1|N}\mathbf{T}_{t-1|N}^{\mathrm{T}} + \mathbf{S}_{t-1|N}\right) \\
S_{10} &= \sum_{t=1}^{N}\left(\mathbf{T}_{t|N}\mathbf{T}_{t-1|N}^{\mathrm{T}} + \mathbf{S}_{t,t-1|N}\right) \\
S_{11} &= \sum_{t=1}^{N}\left(\mathbf{T}_{t|N}\mathbf{T}_{t|N}^{\mathrm{T}} + \mathbf{S}_{t|N}\right) \\
F_{00} &= \sum_{t=1}^{N}\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \\
F_{10} &= \sum_{t=1}^{N}\mathbf{F}_t\mathbf{T}_{t-1|N}^{\mathrm{T}} \\
F_{11} &= \sum_{t=1}^{N}\mathbf{F}_t\mathbf{T}_{t|N}^{\mathrm{T}}.
\end{aligned}$$

Unlike the Newton-Raphson method, no calculation of derivatives is required at each iteration of the EM algorithm and thus it is conceptually simpler to implement. In the above, the Kalman smoother and lag-one covariance smoother are calculated under the present value of the parameter $\Theta^{(j-1)}$ using Property 3.2 and 3.4. The initial mean μ_0 and covariance Σ_0 cannot be estimated simultaneously (see section 3.6.1) and so the covariance matrix is fixed and μ_0 is estimated by the Kalman smoother at $t = 0$.

3.3 Maximum likelihood estimation for a state-space model with a partially known state process

As pointed out earlier, the existing estimation approach only uses information from the observation process $\{\mathbf{P}_t\}$ and does not take into account the possibility of a partially known state process. Here, two approaches are proposed for estimating the parameters when the state process is partially observed. Both approaches assume that the state process is known at $t = n+1, n+2, \cdots, N$.

3.3.1 The $\mathbb{ALL}$ approach

In the case where the state process is partially known, the maximum likelihood method can continue to be used to estimate the parameter $\Theta = \{\mathcal{A}, \mathcal{R}, \phi, \delta, \mathcal{Q}, \mu_0, \Sigma_0\}$. The likelihood function for the observed data set $\{\mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}\}$ in these circumstances is given by

$$\begin{aligned}
\mathcal{L}_{\mathbf{P},\mathbf{T}}(\Theta) &= f_\Theta(\mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}) \\
&= f_\Theta(\mathbf{P}_{1:n}) f_\Theta(\mathbf{P}_{n+1:N}, \mathbf{T}_{n+1:N} | \mathbf{P}_{1:n}) \\
&= \left[\prod_{t=1}^{n} f_\Theta(\mathbf{P}_t | \mathbf{P}_{1:t-1})\right] f_\Theta(\mathbf{P}_{n+1:N} | \mathbf{T}_{n+1:N}, \mathbf{P}_{1:n}) f_\Theta(\mathbf{T}_{n+1:N} | \mathbf{P}_{1:n}) \\
&= \left[\prod_{t=1}^{n} f_\Theta(\mathbf{P}_t | \mathbf{P}_{1:t-1})\right] \left[\prod_{t=n+1}^{N} f_\Theta(\mathbf{P}_t | \mathbf{T}_t)\right] \left[\prod_{t=n+2}^{N} f_\Theta(\mathbf{T}_t | \mathbf{T}_{t-1})\right] f_\Theta(\mathbf{T}_{n+1} | \mathbf{P}_{1:n}).
\end{aligned} \tag{3.3.1}$$

Under the normality assumption and ignoring constants, the log likelihood function can be written as

$$\begin{aligned}
-2\ \ell_{\mathbf{P},\mathbf{T}}(\Theta) = &\sum_{t=1}^{n}\Big[\ln|\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}}+\mathcal{R}| + (\mathbf{P}_t-\mathcal{A}\mathbf{T}_{t|t-1})^{\mathrm{T}}(\mathcal{A}\mathbf{S}_{t|t-1}\mathcal{A}^{\mathrm{T}}+\mathcal{R})^{-1}(\mathbf{P}_t-\mathcal{A}\mathbf{T}_{t|t-1})\Big] \\
&+ m\ln|\mathcal{R}| + \sum_{t=n+1}^{N}(\mathbf{P}_t-\mathcal{A}\mathbf{T}_t)\mathcal{R}^{-1}(\mathbf{P}_t-\mathcal{A}\mathbf{T}_t)^{\mathrm{T}} \\
&+ (m-1)\ln|\mathcal{Q}| + \sum_{t=n+2}^{N}(\mathbf{T}_t-\phi\mathbf{T}_{t-1}-\delta\mathbf{F}_t)\mathcal{Q}^{-1}(\mathbf{T}_t-\phi\mathbf{T}_{t-1}-\delta\mathbf{F}_t)^{\mathrm{T}} \\
&+ \ln|\phi\mathbf{S}_{n|n}\phi^{\mathrm{T}}+\mathcal{Q}| \\
&+ (\mathbf{T}_{n+1}-\phi\mathbf{T}_{n|n}-\delta\mathbf{F}_{n+1})(\phi\mathbf{S}_{n|n}\phi^{\mathrm{T}}+\mathcal{Q})^{-1}(\mathbf{T}_{n+1}-\phi\mathbf{T}_{n|n}-\delta\mathbf{F}_{n+1})^{\mathrm{T}}.
\end{aligned} \tag{3.3.2}$$

This estimation approach will be called the all data ($\mathbb{ALL}$) approach. In this scenario, the parameter $\mathcal{A}$ is estimable within the maximum likelihood estimation process because $\mathbf{T}_{n+1:N}$ is known and thus one is not allowed to standardize the state process freely as in the previous section. This implies that all the unknown parameters can be estimated simultaneously. Unlike the $\mathbb{PXY}$ approach, this estimation approach includes the known state process in the construction of the likelihood function, which can potentially provide useful additional information for estimating the parameter Θ. See subsequent sections for comparison between the efficiency of estimators from different estimation approaches.

The Newton-Raphson or EM algorithms can again be used to obtain the maximum likelihood estimate of Θ for the log likelihood function in (3.3.2). For the EM algorithm, the maximum likelihood estimates is obtained by successively maximizing

$$H\left(\Theta \mid \Theta^{(j-1)}\right) = \mathrm{E}_{\mathbf{T}}\left[\ln f_{\Theta}(\mathbf{P}_{1:N},\mathbf{T}_{0:N})\big|\mathbf{P}_{1:N},\mathbf{T}_{n+1:N},\Theta^{(j-1)}\right].$$

After some manipulation, the value of $\Theta=\{\mathcal{A},\mathcal{R},\phi,\delta,\mathcal{Q},\mu_0\}$ that maximizes $H\left(\Theta \mid \Theta^{(j-1)}\right)$ at iteration j of step 2 of the EM algorithm estimation procedure is given as follows (detail

derivations are given in section 3.6.2):

$$
\begin{aligned}
\mathcal{A}^{(j)} &= \left[\sum_{t=1}^{n}(\widetilde{\mathbf{T}}_{t|N}^{\mathrm{T}}\widetilde{\mathbf{T}}_{t|N}+\widetilde{\mathbf{S}}_{t|N})+\sum_{t=n+1}^{N}\mathbf{T}_t^{\mathrm{T}}\mathbf{T}_t\right]^{-1}\left[\sum_{t=1}^{n}\widetilde{\mathbf{T}}_{t|N}^{\mathrm{T}}\mathbf{P}_t+\sum_{t=n+1}^{N}\mathbf{T}_t^{\mathrm{T}}\mathbf{P}_t\right]\\
\mathcal{R}^{(j)} &= N^{-1}\sum_{t=1}^{n}\left[\mathcal{A}^{(j)}\widetilde{\mathbf{S}}_{t|N}\mathcal{A}^{(j)\mathrm{T}}+(\mathbf{P}_t-\mathcal{A}^{(j)}\widetilde{\mathbf{T}}_{t|N})(\mathbf{P}_t-\mathcal{A}^{(j)}\widetilde{\mathbf{T}}_{t|N})^{\mathrm{T}}\right]+\\
&\quad N^{-1}\sum_{t=n+1}^{N}\left[(\mathbf{P}_t-\mathcal{A}^{(j)}\mathbf{T}_t)(\mathbf{P}_t-\mathcal{A}^{(j)}\mathbf{T}_t)^{\mathrm{T}}\right]\\
\delta^{(j)} &= \left(F_{11}^{\mathrm{T}}-S_{10}S_{00}^{-1}F_{10}^{\mathrm{T}}\right)\left(F_{00}-F_{10}S_{00}^{-1}F_{10}^{\mathrm{T}}\right)^{-1}\\
\phi^{(j)} &= \left(S_{10}-\delta^{(j)}F_{10}\right)S_{00}^{-1}\\
\mathcal{Q}^{(j)} &= N^{-1}\left(S_{11}-\phi^{(j)}S_{10}^{\mathrm{T}}-\delta^{(j)}F_{11}\right)\\
\mu_0^{(j)} &= \widetilde{\mathbf{T}}_{0|N}
\end{aligned}
\tag{3.3.3}
$$

where

$$
\begin{aligned}
S_{00} &= \sum_{t=1}^{n+1}\left(\widetilde{\mathbf{T}}_{t-1|N}\widetilde{\mathbf{T}}_{t-1|N}^{\mathrm{T}}+\widetilde{\mathbf{S}}_{t-1|N}\right)+\sum_{t=n+2}^{N}\mathbf{T}_{t-1}\mathbf{T}_{t-1}^{\mathrm{T}}\\
S_{10} &= \sum_{t=1}^{n}\left(\widetilde{\mathbf{T}}_{t|N}\widetilde{\mathbf{T}}_{t-1|N}^{\mathrm{T}}+\mathbf{S}_{t,t-1|N}\right)+\mathbf{T}_{n+1}\widetilde{\mathbf{T}}_{n|N}^{\mathrm{T}}+\sum_{t=n+2}^{N}\mathbf{T}_t\mathbf{T}_{t-1}^{\mathrm{T}}\\
S_{11} &= \sum_{t=1}^{n}\left(\widetilde{\mathbf{T}}_{t|N}\widetilde{\mathbf{T}}_{t|N}^{\mathrm{T}}+\widetilde{\mathbf{S}}_{t|N}\right)+\sum_{t=n+1}^{N}\mathbf{T}_t\mathbf{T}_t^{\mathrm{T}}\\
F_{00} &= \sum_{t=1}^{N}\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}\\
F_{10} &= \sum_{t=1}^{n+1}\mathbf{F}_t\widetilde{\mathbf{T}}_{t-1|N}^{\mathrm{T}}+\sum_{t=n+2}^{N}\mathbf{F}_t\mathbf{T}_{t-1}^{\mathrm{T}},\quad F_{11}=\sum_{t=1}^{n}\mathbf{F}_t\widetilde{\mathbf{T}}_{t|N}^{\mathrm{T}}+\sum_{t=n+1}^{N}\mathbf{F}_t\mathbf{T}_t^{\mathrm{T}}.
\end{aligned}
$$

Again, the modified Kalman smoother and the lag-one covariance smoother are calculated under the present value of the parameter $\Theta^{(j-1)}$ using Property 3.3 and 3.4. The initial mean μ_0 and covariance Σ_0 again cannot be estimated simultaneously and so the covariance matrix is fixed and μ_0 is estimated by the modified Kalman smoother at $t = 0$.

Due to the complexity of the likelihood function, the theoretical asymptotic distribution (when both n and $m = N - n$ go to infinity) for the estimator obtained by maximizing the

$\mathbb{ALL}$ likelihood in (3.3.1) was not derived. Recall that $\ell_{\mathbf{P},\mathbf{T}}(\Theta)$ is

$$\ell_{\mathbf{P},\mathbf{T}}(\Theta) = \underbrace{\sum_{t=1}^{n} \ln f_\Theta(\mathbf{P}_t|\mathbf{P}_{1:t-1})}_{f_1} + \underbrace{\sum_{t=n+2}^{N} \ln f_\Theta(\mathbf{T}_t|\mathbf{T}_{t-1})}_{f_2} + \underbrace{\sum_{t=n+1}^{N} \ln f_\Theta(\mathbf{P}_t|\mathbf{T}_t)}_{f_3} + \ln f_\Theta(\mathbf{T}_{n+1}|\mathbf{P}_{1:n}).$$

In general, one can derive the asymptotic distribution for an estimator in possibly two ways. The first approach would be to write out the estimator as a function of the data and then derive the asymptotic distribution of the estimator directly. That is, one would take the derivative of the log likelihood function and solve for the optimal value of the parameter. The optimal value, which is also the estimator, will be a function of the data and hence one may directly derive the asymptotic distribution of the estimator based on the distributional property of the data. However, this approach is not applicable for estimator obtained from $\ell_{\mathbf{P},\mathbf{T}}(\Theta)$ because one cannot write out the estimator in terms of the data due to the recursive nature of the Kalman filter estimate that is involved when constructing $\sum_{t=1}^{n} f_\Theta(\mathbf{P}_t|\mathbf{P}_{1:t-1})$.

The second approach, introduced by Cramèr (1946), would involve the following three general steps: 1) derive the limiting distribution of the first and second derivatives of the log likelihood function; 2) show that the estimator is consistent, i.e. $\hat{\Theta} \xrightarrow{p} \Theta_0$, where Θ_0 is the true parameter value and 3) apply a Taylor series expansion on the log likelihood function to derive the asymptotic distribution for the estimator. Details of this approach can be found in many statistical books, such as Serfling (1980, p. 144-149) and Lehmann and Casella (1998, Ch. 6).

For the log likelihood function $\ell_{\mathbf{P},\mathbf{T}}(\Theta)$, there are existing results that describe the limiting distribution of two of its components. The first part of the log likelihood, f_1, is equivalent to the $\mathbb{PXY}$ log likelihood, $\ell_{\mathbf{P}}(\Theta)$, with N replaced by n. From the proof of Theorem 3.2.1 (Caines 1988), one has

$$n^{-1/2}\frac{\partial}{\partial\Theta} f_1(\Theta_0) \xrightarrow{d} \mathrm{N}(0, \mathfrak{I}_0^{(1)})$$
$$n^{-1}\frac{\partial^2}{\partial\Theta\partial\Theta^{\mathrm{T}}} f_1(\Theta^*) \xrightarrow{p} -\mathfrak{I}_0^{(1)},$$

where $\Theta^* \xrightarrow{p} \Theta_0$ and $\mathfrak{I}_0^{(1)}$ is the Fisher information matrix for $\{\mathbf{P}_t\}$. From the literature on maximum likelihood estimation for dependent observations, for example, Bhat (1974) and

Crowder (1976), one also has the result for the second part of the log likelihood function that

$$(m-1)^{-1/2}\frac{\partial}{\partial\Theta}f_2(\Theta_0) \xrightarrow{d} \mathrm{N}(0, \mathfrak{I}_0^{(2)})$$
$$(m-1)^{-1}\frac{\partial^2}{\partial\Theta\partial\Theta^{\mathrm{T}}}f_2(\Theta^*) \xrightarrow{p} -\mathfrak{I}_0^{(2)},$$

where $\mathfrak{I}_0^{(2)}$ is the Fisher information matrix for $\{\mathbf{T}_t\}$. The third part of the log likelihood function, f_3, contains two dependent stochastic processes, namely $\{\mathbf{P}_t\}$ and $\{\mathbf{T}_t\}$. Theories have been developed for the case where the likelihood consists of one stochastic process, such as those by Billingsley (1961), Bhat (1974) and Crowder (1976). However, there does not appear to be existing theory that deals with the situation where the likelihood consists of two dependent stochastic processes. Deriving the asymptotic distribution for the derivatives of f_3 from scratch would be a difficult task given that the derivation for the case of a single stochastic process is not straightforward. Even if one can derive the asymptotic distribution for the derivatives of f_3, one is still faced with the problem of combining the individual results. Since f_1, f_2 and f_3 are clearly dependent, some work would be required to derive the limiting distribution of the sum of the derivatives of f_1, f_2 and f_3. The last term in the log likelihood consists of only one single quantity and it would likely vanish in the limit and thus should have no effect on the asymptotic results. Alternatively, one can try to derive the asymptotic distribution of $\ell_{\mathbf{P},\mathbf{T}}(\Theta)$ directly instead of working with each component separately. This is also hard to accomplish given the complexity of the likelihood function.

Even though the theoretical asymptotic result (when both n and m go to infinity) is not available, simulation results from a later section of this chapter do provide evidence that the estimator obtained from the $\mathbb{ALL}$ likelihood possesses the same asymptotic property as that stated in Theorem 3.2.1 with $\mathcal{L}_{\mathbf{P}}(\Theta)$ replaced by $\mathcal{L}_{\mathbf{P},\mathbf{T}}(\Theta)$. On the other hand, if the size of m is fixed but n goes to infinity, the second and third part of the log likelihood function, f_2 and f_3, will likely vanish in the limit and thus the limiting results should be like that from the $\mathbb{PXY}$ approach. Similarly, if n is fixed but m goes to infinity, f_1 and the last term in the log likelihood will likely vanish and thus the limiting results should be like that in Theorem 3.3.1 which will be discussed in the next section.

3.3.2 The $\mathbb{CAL}$ approach

A conceptually simpler method for estimating the parameters that specify a state-space model is to consider constructing a likelihood function that uses only the calibration period data. For simplicity, the conditional likelihood function of the calibration period data conditioning on $\mathbf{T}_{n+1}$ is considered and it can be written as

$$\begin{aligned}\mathcal{L}_{\mathbf{c}}(\Theta) &= f_\Theta(\mathbf{P}_{n+1:N}, \mathbf{T}_{n+2:N}|\mathbf{T}_{n+1}) \\ &= f_\Theta(\mathbf{P}_{n+1:N}|\mathbf{T}_{n+1:N}) f_\Theta(\mathbf{T}_{n+2:N}|\mathbf{T}_{n+1}) \\ &= \left[\prod_{t=n+1}^{N} f_\Theta(\mathbf{P}_t|\mathbf{T}_t)\right]\left[\prod_{t=n+2}^{N} f_\Theta(\mathbf{T}_t|\mathbf{T}_{t-1})\right],\end{aligned} \tag{3.3.4}$$

where $\Theta = \{\mathcal{A}, \mathcal{R}, \phi, \delta, \mathcal{Q}\}$. This conditional likelihood function differs from the full (unconditional) likelihood function $f_\Theta(\mathbf{P}_{n+1:N}, \mathbf{T}_{n+1:N})$ only slightly with the latter being the former multiplied by the density function of $\mathbf{T}_{n+1}$. However, the use of the conditional likelihood has several advantages. First, for the Gaussian state-space model (3.1.2) with $q = 1$, $\mathbf{T}_{n+1}$ can be written as,

$$\mathbf{T}_{n+1} = \phi^{n+1}\mathbf{T}_0 + \sum_{j=0}^{n} \phi^j(\delta\mathbf{F}_{n+1-j} + w_{n+1-j}).$$

So, the density function of $\mathbf{T}_{n+1}$ would involve the parameters of the initial state (μ_0 and Σ_0) and a total of $n+1$ unique power terms of ϕ. Thus, solving the maximum likelihood estimates of ϕ from the unconditional likelihood becomes more complicated (see later part of this section for the maximum likelihood estimates of the conditional likelihood). Second, the conditional likelihood depends only on the data and the model assumptions for the calibration period, and hence is insensitive to misspecifications of the model for the pre-calibration period. On the other hand, the full unconditional likelihood would depend on the correct specification of the model for the pre-calibration period in order to compute the correct density function for $\mathbf{T}_{n+1}$. Also, from an asymptotic view point, as the length of the calibration period goes to infinity, the loss of information by the conditional likelihood, which ignores the likelihood based on the single point $\mathbf{T}_{n+1}$, should be negligible.

Unlike the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches, numerical procedures are not required to find the

maximum likelihood estimates of the parameter Θ in (3.3.4) in the case of a Gaussian state-space model. However, this estimation approach does not allow the estimation of the parameters μ_0 and Σ_0 because these parameters are not involved in the likelihood function. The likelihood function also excludes information from the pre-calibration period observations, which might be useful for estimating the parameter Θ. However, from a robustness standpoint, this approach turns out to be superior. In particular, when the structure of the observation noise e_t is different from that assumed in the state-space model, the parameter estimates from this approach remain asymptotically unbiased. More details will be given in subsequent sections.

Throughout this section, $\mathbf{P}_t$ and $\mathbf{T}_t$ are assumed to be univariate (i.e. p=q=1). Taking logarithms and ignoring a constant, the log likelihood function for the univariate temperature reconstruction problem is given by

$$\begin{aligned} -2\ \ell_{\mathbf{c}}(\Theta) &= \sum_{t=n+1}^{N} \left[\ln \mathcal{R} + (\mathbf{P}_t - \mathcal{A}\mathbf{T}_t)^2/\mathcal{R}\right] + \sum_{t=n+2}^{N} \left[\ln \mathcal{Q} + (\mathbf{T}_t - \phi\mathbf{T}_{t-1} - \delta\mathbf{F}_t)^2/\mathcal{Q}\right] \\ &= \sum_{t=n+1}^{N} \left[\ln \mathcal{R} + (\mathbf{P}_t - \mathcal{A}\mathbf{T}_t)^2/\mathcal{R}\right] + \sum_{t=n+2}^{N} \left[\ln \mathcal{Q} + (\mathbf{T}_t - \Upsilon^{\mathrm{T}} D_t)^2/\mathcal{Q}\right], \end{aligned} \tag{3.3.5}$$

where

$$\Upsilon = \begin{bmatrix} \phi \\ \delta^{\mathrm{T}} \end{bmatrix}, \quad D_t = \begin{bmatrix} \mathbf{T}_{t-1} \\ \mathbf{F}_t \end{bmatrix}.$$

This estimation approach will be called the calibration data ($\mathbb{CAL}$) approach. Taking derivative of the log likelihood with respect to the parameters, one has

$$\begin{aligned} \frac{\partial \ell_{\mathbf{c}}(\Theta)}{\partial \mathcal{A}} &= \frac{1}{\mathcal{R}} \sum_{t=n+1}^{N} \mathbf{T}_t(\mathbf{P}_t - \mathcal{A}\mathbf{T}_t) \\ \frac{\partial \ell_{\mathbf{c}}(\Theta)}{\partial \mathcal{R}} &= \frac{-m}{2\mathcal{R}} + \frac{1}{2\mathcal{R}^2} \sum_{t=n+1}^{N} (\mathbf{P}_t - \mathcal{A}\mathbf{T}_t)^2 \\ \frac{\partial \ell_{\mathbf{c}}(\Theta)}{\partial \Upsilon} &= \frac{1}{\mathcal{Q}} \sum_{t=n+2}^{N} D_t(\mathbf{T}_t - \Upsilon^{\mathrm{T}} D_t) \\ \frac{\partial \ell_{\mathbf{c}}(\Theta)}{\partial \mathcal{Q}} &= -\frac{m-1}{2\mathcal{Q}} + \frac{1}{2\mathcal{Q}^2} \sum_{t=n+2}^{N} (\mathbf{T}_t - \Upsilon^{\mathrm{T}} D_t)^2. \end{aligned} \tag{3.3.6}$$

Setting the derivatives equal to zero, the maximum likelihood estimate of $\Theta = \{\mathcal{A}, \mathcal{R}, \phi, \delta, \mathcal{Q}\}$ for (3.3.5) can be solved analytically, and is given by,

$$\begin{aligned}
\hat{\mathcal{A}} &= \left(\sum_{t=n+1}^{N} \mathbf{T}_t^2 \right)^{-1} \sum_{t=n+1}^{N} \mathbf{T}_t \mathbf{P}_t \\
\hat{\mathcal{R}} &= m^{-1} \sum_{t=n+1}^{N} (\mathbf{P}_t - \hat{\mathcal{A}} \mathbf{T}_t)^2 \\
\hat{\Upsilon} &= \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}} \right)^{-1} \sum_{t=n+2}^{N} D_t \mathbf{T}_t \\
\hat{\mathcal{Q}} &= (m-1)^{-1} \sum_{t=n+2}^{N} (\mathbf{T}_t - \hat{\Upsilon}^{\mathrm{T}} D_t)^2,
\end{aligned} \tag{3.3.7}$$

The asymptotic property of the maximum likelihood estimator of Θ obtained from the $\mathbb{CAL}$ likelihood in (3.3.5) will be provided next. Since the estimator can be written in terms of the data in this case, the first approach for deriving the asymptotic distribution as stated earlier was used.

Theorem 3.3.1 *For the Gaussian state-space model (3.1.2) with $p = q = 1$, suppose that $|\phi| < 1$, $\mathbf{F}_t$ is non-random and $|\mathbf{F}_{jt}| \leq c$ for all j and $t \geq n+2$, where c is a constant and $\sum_{t=n+2}^{N} \mathbf{F}_{jt}^2 \to \infty$ as $m \to \infty$ for all j. Then, as $m \to \infty$, the maximum likelihood estimates of $\mathcal{A}, \mathcal{Q}, \Upsilon, \mathcal{R}$ in (3.3.7) satisfy:*

$$\begin{aligned}
\sqrt{\mathfrak{I}_{\mathcal{A}}}(\hat{\mathcal{A}} - \mathcal{A}) &\xrightarrow{d} \mathrm{N}(0,1) \\
\sqrt{\mathfrak{I}_{\mathcal{R}}}(\hat{\mathcal{R}} - \mathcal{R}) &\xrightarrow{d} \mathrm{N}(0,1) \\
\sqrt{\mathfrak{I}_{\mathcal{Q}}}(\hat{\mathcal{Q}} - \mathcal{Q}) &\xrightarrow{d} \mathrm{N}(0,1) \\
\sqrt{\mathfrak{I}_{\Upsilon}}(\hat{\Upsilon} - \Upsilon) &\xrightarrow{d} \mathrm{MVN}\left(\vec{0}, \mathbf{I}\right)
\end{aligned}$$

where $\mathbf{I}$ is an identity matrix and

$$\begin{aligned}
\mathfrak{I}_{\mathcal{A}} &= \mathrm{E}[-\partial^2 \ell_{\mathbf{c}}(\Theta)/\partial \mathcal{A}^2] = \mathcal{R}^{-1} \sum_{t=n+1}^{N} [\mu_t^2 + \gamma_t] \\
\mathfrak{I}_{\mathcal{R}} &= \mathrm{E}[-\partial^2 \ell_{\mathbf{c}}(\Theta)/\partial \mathcal{R}^2] = m/2\mathcal{R}^2 \\
\mathfrak{I}_{\mathcal{Q}} &= \mathrm{E}[-\partial^2 \ell_{\mathbf{c}}(\Theta)/\partial \mathcal{Q}^2] = (m-1)/2\mathcal{Q}^2 \\
\mathfrak{I}_{\Upsilon} &= \mathrm{E}[-\partial^2 \ell_{\mathbf{c}}(\Theta)/\partial \Upsilon \; \partial \Upsilon^{\mathrm{T}}] = \mathcal{Q}^{-1} \sum_{t=n+2}^{N} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1}\mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1}\mathbf{F}_t & \mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \\
\mu_t &= \begin{cases} \mathbf{T}_{n+1}, & \text{if } t = n+1 \\ \phi^{t-(n+1)}\mathbf{T}_{n+1} + \sum_{j=0}^{t-(n+2)} \phi^j \delta \mathbf{F}_{t-j}, & \text{if } t \geq n+2 \end{cases} \\
\gamma_t &= \begin{cases} 0, & \text{if } t = n+1 \\ \mathcal{Q} \sum_{j=0}^{t-(n+2)} \phi^{2j}, & \text{if } t \geq n+2 \end{cases} \\
\sqrt{\mathfrak{I}_{\Upsilon}} &= \mathbf{C}\mathbf{D}^{1/2}\mathbf{C}^{\mathrm{T}},
\end{aligned}$$

where $\mathbf{C}$ is a matrix whose columns contain the normalized eigenvectors of $\mathfrak{I}_{\Upsilon}$ and $\mathbf{D}^{1/2}$ is a diagonal matrix of the square root of the eigenvalues of $\mathfrak{I}_{\Upsilon}$.

□

Proof: *See next section.*

3.3.3 Proof of asymptotic distribution

In this section, the following conditions hold:

(C.1) $|\phi| < 1$,

(C.2) $|\mathbf{F}_{jt}| \leq c$ for all j and $t \geq n+2$, where c is a constant,

(C.3) $\sum_{t=n+2}^{N} \mathbf{F}_{jt}^2 \to \infty$ as $m \to \infty$ for all j.

Condition (C.1) is necessary to ensure stationarity of the internal variability in (3.1.3). Conditions (C.2) and (C.3) are both technical conditions imposed on the climate model estimated response to forcings, which are needed to establish various convergence results. Condition (C.2) roughly means that the impact of any external forcing variable on the temperature at any given year has an uniform bound. On the other hand, condition (C.3) roughly means that the cumulative impact of any such variable, as measured by the sum in the condition, is not bounded as time goes to infinity. An examination of the estimated response to forcings from an EBM simulation (Hegerl et al. 2007a) for the last 1000 years suggests that these conditions are well supported.

The variables and parameters used in this section are defined in (3.3.5), (3.3.7) and Theorem 3.3.1, unless otherwise specified. In particular, $m = N - n$ is the length of the calibration period,

$$\Upsilon = \begin{bmatrix} \phi \\ \delta^{\mathrm{T}} \end{bmatrix}, \quad D_t = \begin{bmatrix} \mathbf{T}_{t-1} \\ \mathbf{F}_t \end{bmatrix},$$

$$\mu_t = \begin{cases} \mathbf{T}_{n+1}, & \text{if } t = n+1 \\ \phi^{t-(n+1)}\mathbf{T}_{n+1} + \sum_{j=0}^{t-(n+2)} \phi^j \delta \mathbf{F}_{t-j}, & \text{if } t \geq n+2, \end{cases}$$

$$\gamma_t = \begin{cases} 0, & \text{if } t = n+1 \\ \mathcal{Q} \sum_{j=0}^{t-(n+2)} \phi^{2j}, & \text{if } t \geq n+2. \end{cases}$$

Note that for $t \geq n+2$, $\gamma_t \geq \mathcal{Q}$. Also, $\mathbf{I}$ is an identity matrix. Furthermore, if $\mathbf{W}$ is a symmetric matrix, then the square root matrix $\sqrt{\mathbf{W}} = \mathbf{W}^{1/2}$ is defined as,

$$\sqrt{\mathbf{W}} = \mathbf{G}\mathbf{H}^{1/2}\mathbf{G}^{\mathrm{T}},$$

where $\mathbf{G}$ is a matrix whose columns contain the normalized eigenvectors of $\mathbf{W}$ and $\mathbf{H}^{1/2}$ is a diagonal matrix of the square root of the eigenvalues of $\mathbf{W}$. Note that

$$\begin{aligned}\mathbf{W}^{1/2}\mathbf{W}^{1/2} &= \mathbf{G}\mathbf{H}^{1/2}\mathbf{G}^{\mathrm{T}}\mathbf{G}\mathbf{H}^{1/2}\mathbf{G}^{\mathrm{T}} \\ &= \mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}} \\ &= \mathbf{W}\end{aligned}$$

because $\mathbf{G}^{\mathrm{T}}\mathbf{G} = \mathbf{I}$.

Before proving Theorem 3.3.1, a series of lemmas will first be presented. To begin, let

$$\widetilde{\mathbf{T}}_t = \begin{cases} 0, & \text{if } t = n+1 \\ \sum_{j=0}^{t-(n+2)} \phi^j w_{t-j}, & \text{if } t \geq n+2, \end{cases}$$

For $t \geq n+1$, one can rewrite $\mathbf{T}_t$ as,

$$\begin{aligned}\mathbf{T}_t &= \begin{cases} \mathbf{T}_{n+1}, & \text{if } t = n+1 \\ \phi\mathbf{T}_{t-1} + \delta\mathbf{F}_t + w_t, & \text{if } t \geq n+2, \end{cases} \\ &= \begin{cases} \mathbf{T}_{n+1}, & \text{if } t = n+1 \\ \phi^{t-(n+1)}\mathbf{T}_{n+1} + \sum_{j=0}^{t-(n+2)} \phi^j(\delta\mathbf{F}_{t-j} + w_{t-j}), & \text{if } t \geq n+2, \end{cases} \\ &= \mu_t + \widetilde{\mathbf{T}}_t.\end{aligned}$$

Note that the w_t's are i.i.d. with mean zero. Thus,

$$\mathrm{E}(\widetilde{\mathbf{T}}_t) = \mathrm{E}\left(\sum_{j=0}^{t-(n+2)} \phi^j w_{t-j}\right) = 0.$$

and

$$\mathrm{E}(\mathbf{T}_t) = \mu_t.$$

Lemma 1 *The entries in the matrix* $m^{-1}\sum_{t=n+1}^{N}\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}$ *are uniformly bounded by* c^2 *for all* m*. Also,* μ_t *is uniformly bounded in* t *and consequently,* $m^{-1}\sum_{t=n+1}^{N}\mu_t^2$ *and the elements in the vector* $(m-1)^{-1}\sum_{t=n+2}^{N}\mu_{t-1}\mathbf{F}_t$ *are also uniformly bounded for all* m*.*

□

Proof: Consider the ij^{th} entry of $m^{-1}\sum_{t=n+1}^{N}\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}$,

$$\begin{aligned} m^{-1}\Big|\sum_{t=n+1}^{N}\mathbf{F}_{it}\mathbf{F}_{jt}\Big| &\leq m^{-1}\sum_{t=n+1}^{N}c^2 \\ &= c^2 \quad < \infty. \end{aligned}$$

Similarly, to show that the elements in $m^{-1}\sum_{t=n+1}^{N}\mu_t^2$ and $(m-1)^{-1}\sum_{t=n+2}^{N}\mu_{t-1}\mathbf{F}_t$ are uniformly bounded, it is sufficient to show that μ_t is uniformly bounded. The desire results can then be obtained using the same argument as above. Let $\delta_\ell,\ \ell = 1,2,..,k$ be the ℓ^{th} elements in δ and note that

$$\begin{aligned} |\mu_t| &= \Big|\phi^{t-(n+1)}\mathbf{T}_{n+1} + \sum_{j=0}^{t-(n+2)}\phi^j\delta\mathbf{F}_{t-j}\Big| \\ &\leq |\phi^{t-(n+1)}\mathbf{T}_{n+1}| + \sum_{j=0}^{t-(n+2)}|\phi^j\delta\mathbf{F}_{t-j}| \\ &= |\phi^{t-(n+1)}\mathbf{T}_{n+1}| + \sum_{j=0}^{t-(n+2)}|\phi^j||\delta\mathbf{F}_{t-j}| \\ &\leq |\phi^{t-(n+1)}\mathbf{T}_{n+1}| + \sum_{j=0}^{t-(n+2)}|\phi|^j c\sum_{l=1}^{k}|\delta_\ell| \\ &\leq |\phi^{t-(n+1)}\mathbf{T}_{n+1}| + \frac{c\sum_{\ell=1}^{k}|\delta_\ell|}{1-|\phi|}. \end{aligned}$$

Thus, $|\mu_t|$ is bounded by a constant. This completes the proof of Lemma 1.

■

Lemma 2 *For $t \geq n+1$, the following properties hold for $\widetilde{\mathbf{T}}_t$ and $\mathbf{T}_t$.*

1) $\mathrm{E}(\widetilde{\mathbf{T}}_t^2) = \gamma_t$ *and is uniformly bounded in* t,

2) $\mathrm{E}(\widetilde{\mathbf{T}}_t^3) = 0$,

3) $\mathrm{E}(\widetilde{\mathbf{T}}_t^4)$ *is uniformly bounded in* t,

4) $\mathrm{E}(\mathbf{T}_t^2) = \mu_t^2 + \gamma_t$ *and is uniformly bounded in* t,

□

Proof: For $t = n+1$, it is clear that the equations in 1) to 4) all hold and the four expectations are all finite since $\mathbf{T}_{n+1}$ is given (see (3.3.4)). For $t \geq n+2$, note that $\widetilde{\mathbf{T}}_t = \sum_{j=0}^{t-(n+2)} \phi^j w_{t-j}$ and the w_t's are i.i.d. with mean zero and variance $\mathcal{Q}$. Thus,

$$\begin{aligned}\mathrm{E}(\widetilde{\mathbf{T}}_t^2) &= \mathcal{Q} \sum_{j=0}^{t-(n+2)} \phi^{2j} \\ &= \gamma_t.\end{aligned}$$

Note that γ_t is uniformly bounded in t because

$$\begin{aligned}|\gamma_t| &= \Big|\mathcal{Q} \sum_{j=0}^{t-(n+2)} \phi^{2j}\Big| \\ &= \mathcal{Q} \sum_{j=0}^{t-(n+2)} \phi^{2j} \\ &\leq \mathcal{Q} \sum_{j=0}^{\infty} \phi^{2j} \\ &= \frac{\mathcal{Q}}{1-\phi^2}.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathrm{E}(\widetilde{\mathbf{T}}_t^3) &= \mathrm{E}\left(\sum_{i=0}^{t-(n+2)} \sum_{j=0}^{t-(n+2)} \sum_{k=0}^{t-(n+2)} \phi^i \phi^j \phi^k w_{t-i} w_{t-j} w_{t-k} \right) \\ &= 0\end{aligned}$$

since $\mathrm{E}(w_{t-i}w_{t-j}w_{t-k}) = 0$ for all i, j and k because the w_t's are i.i.d. normal random variables with mean zero and $\mathrm{E}(w_t^3)$=0.

Now, consider the fourth moment of $\widetilde{\mathbf{T}}_t$.

$$\mathrm{E}(\widetilde{\mathbf{T}}_t^4) = \mathrm{E}\left(\sum_{i=0}^{t-(n+2)} \sum_{j=0}^{t-(n+2)} \sum_{k=0}^{t-(n+2)} \sum_{\ell=0}^{t-(n+2)} \phi^i \phi^j \phi^k \phi^\ell w_{t-i} w_{t-j} w_{t-k} w_{t-\ell} \right).$$

The expectation of the product of w_t's is non-zero if one of the following is true:

$$i = j \neq k = \ell, \quad i = k \neq j = \ell, \quad i = \ell \neq j = k \quad \text{or} \quad i = j = k = \ell.$$

Thus,

$$\begin{aligned} \mathrm{E}(\widetilde{\mathbf{T}}_t^4) &= 3 \sum_{j \neq k} \phi^{2j} \phi^{2k} \mathrm{E}(w_{t-j}^2) \mathrm{E}(w_{t-k}^2) + \sum_{i=0}^{t-(n+2)} \phi^{4i} \mathrm{E}(w_{t-i}^4) \\ &= 3\mathcal{Q}^2 \sum_{j \neq k} \phi^{2j} \phi^{2k} + 3\mathcal{Q}^2 \sum_{i=0}^{t-(n+2)} \phi^{4i} \\ &= 3\mathcal{Q}^2 \sum_{j=0}^{t-(n+2)} \sum_{k=0}^{t-(n+2)} \phi^{2j} \phi^{2k}. \end{aligned}$$

Since

$$\sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \phi^{2j} \phi^{2k} = (1 - \phi^2)^{-2},$$

it is now clear that $\mathrm{E}(\widetilde{\mathbf{T}}_t^4)$ is uniformly bounded in t. Next

$$\mathrm{E}(\mathbf{T}_t^2) = \mathrm{E}(\mu_t^2 + 2\mu_t \widetilde{\mathbf{T}}_t + \widetilde{\mathbf{T}}_t^2) = \mu_t^2 + \gamma_t.$$

By Lemma 1 and the first property of this lemma, one can see that $\mathrm{E}(\mathbf{T}_t^2)$ is uniformly bounded in t. This completes the proof of Lemma 2.

■

Theorem 3.3.2 ***Weak laws of large numbers for uncorrelated random variables***

Let $X_1, X_2, \ldots$ be uncorrelated with mean $u_1, u_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. If

$$\sum_{i=1}^{n} \sigma_i^2 = o(n^2),$$

then, as $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \frac{1}{n} \sum_{i=1}^{n} u_i.$$

□

Proof: See, for example, Rao (1973, 112-114).

Lemma 3 *As $m \to \infty$, the following convergence results hold.*

$$m^{-1} \sum_{t=n+1}^{N} \mu_t \widetilde{\mathbf{T}}_t \xrightarrow{p} 0$$

and

$$m^{-1} \sum_{t=n+1}^{N} \mathbf{T}_t e_t \xrightarrow{p} 0.$$

Also

$$\frac{\sum_{t=n+1}^{N} \widetilde{\mathbf{T}}_t^2}{\sum_{t=n+1}^{N} \gamma_t} \xrightarrow{p} 1$$

and

$$\frac{\sum_{t=n+1}^{N} \mathbf{T}_t^2}{\sum_{t=n+1}^{N} (\mu_t^2 + \gamma_t)} \xrightarrow{p} 1.$$

Furthermore,

$$\left(\sum_{t=n+2}^{N} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1} \mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1} \mathbf{F}_t & \mathbf{F}_t \mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \right)^{-1} \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}} \right) \xrightarrow{p} \mathbf{I}$$

with respect to the usual entrywise norm.

□

Proof of the first result:

Recall that

$$\widetilde{\mathbf{T}}_t = \begin{cases} 0, & \text{if } t = n+1 \\ \sum_{j=0}^{t-(n+2)} \phi^j w_{t-j}, & \text{if } t \geq n+2. \end{cases}.$$

So,

$$\begin{aligned}
\mu_{n+1}\widetilde{\mathbf{T}}_{n+1} &= 0 \\
\mu_{n+2}\widetilde{\mathbf{T}}_{n+2} &= \mu_{n+2}w_{n+2} \\
\mu_{n+3}\widetilde{\mathbf{T}}_{n+3} &= \mu_{n+3}\phi w_{n+2} + \mu_{n+3}w_{n+3} \\
\mu_{n+4}\widetilde{\mathbf{T}}_{n+4} &= \mu_{n+4}\phi^2 w_{n+2} + \mu_{n+4}\phi w_{n+3} + \mu_{n+4}w_{n+4} \\
&\vdots \\
\mu_N\widetilde{\mathbf{T}}_N &= \mu_N\phi^{N-(n+2)}w_{n+2} + \mu_N\phi^{N-(n+3)}w_{n+3} + \cdots + \mu_N w_N
\end{aligned}$$

and thus

$$\sum_{t=n+1}^{N} \mu_t\widetilde{\mathbf{T}}_t = \sum_{t=n+2}^{N} a_t w_t,$$

where

$$a_t = \mu_t + \mu_{t+1}\phi + \mu_{t+2}\phi^2 + \cdots + \mu_N\phi^{N-t}.$$

From Lemma 1, $|\mu_j| \leq c_1$ for all j, where c_1 is a constant. Thus,

$$\begin{aligned}
|a_t| \leq & c_1 + c_1|\phi| + c_1|\phi|^2 + \cdots + c_1|\phi|^{N-t} \\
\leq & \frac{c_1}{1+|\phi|}.
\end{aligned}$$

That is, a_t is uniformly bounded in t. Then to show the first convergence result, the weak laws of large numbers for uncorrelated random variables (Theorem 3.3.2) is used. To do so, the mean and variance of $a_t w_t$ are first derived and they are

$$\begin{aligned}
\mathrm{E}(a_t w_t) &= a_t\mathrm{E}(w_t) = 0 \\
\mathrm{Var}(a_t w_t) &= a_t^2\mathrm{Var}(w_t) = a_t^2\mathcal{Q}.
\end{aligned}$$

In order to use Theorem 3.3.2, one needs to show that

$$\frac{1}{m^2}\sum_{t=n+1}^{N}\mathrm{Var}(a_t w_t) = \frac{\mathcal{Q}}{m}\sum_{t=n+1}^{N}\frac{a_t^2}{m}$$
$$= o(1).$$

This follows by noting that $\mathcal{Q}/m \longrightarrow 0$ as $m \longrightarrow \infty$ and $m^{-1}\sum_{t=n+1}^{N} a_t^2$ is bounded. Therefore, by Theorem 3.3.2,

$$m^{-1}\sum_{t=n+1}^{N}\mu_t\widetilde{\mathbf{T}}_t = m^{-1}\sum_{t=n+1}^{N}a_t w_t \xrightarrow{p} 0.$$

Proof of the second result:

$$m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_t e_t = m^{-1}\sum_{t=n+1}^{N}(\mu_t + \widetilde{\mathbf{T}}_t)e_t$$
$$= m^{-1}\sum_{t=n+1}^{N}\mu_t e_t + m^{-1}\sum_{t=n+1}^{N}\widetilde{\mathbf{T}}_t e_t$$

By Theorem 3.3.2,

$$m^{-1}\sum_{t=n+1}^{N}\mu_t e_t \xrightarrow{p} 0, \quad m^{-1}\sum_{t=n+1}^{N}\widetilde{\mathbf{T}}_t e_t \xrightarrow{p} 0.$$

The above follows because the e_t's are i.i.d. and for $t \neq s$,

$$\mathrm{Cov}(\widetilde{\mathbf{T}}_t e_t, \widetilde{\mathbf{T}}_s e_s) = \mathrm{E}(e_t e_s \widetilde{\mathbf{T}}_t \widetilde{\mathbf{T}}_s) = 0.$$

which implies the $\widetilde{\mathbf{T}}_t e_t$'s are uncorrelated. Also $\mathrm{E}(\mu_t e_t) = 0$, $\mathrm{E}(\widetilde{\mathbf{T}}_t e_t) = 0$ and

$$m^{-2}\sum_{t=n+1}^{N}\mathrm{Var}(\mu_t e_t) = m^{-2}\mathcal{R}\sum_{t=n+1}^{N}\mu_t^2 = o(1)$$

since $m^{-1}\sum_{t=n+1}^{N}\mu_t^2$ is bounded by Lemma 1. Furthermore, $\widetilde{\mathbf{T}}_{n+1}e_{n+1} = 0$ and for $t \geq n+2$,

$$\widetilde{\mathbf{T}}_t e_t = e_t\sum_{j=0}^{t-(n+2)}\phi^j w_{t-j}.$$

Thus

$$m^{-2}\sum_{t=n+1}^{N}\mathrm{Var}(\widetilde{\mathbf{T}}_t e_t) = m^{-2}\mathcal{R}\mathcal{Q}\sum_{t=n+2}^{N}\left(\sum_{j=0}^{t-(n+2)}\phi^{2j}\right) = o(1)$$

since $\sum_{j=0}^{t-(n+2)}\phi^{2j} \le (1-\phi^2)^{-1}$. Therefore,

$$m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_t e_t \xrightarrow{p} 0.$$

Proof of the third result:

Note that $\widetilde{\mathbf{T}}_{n+1}^2 = 0$ and

$$\begin{aligned}
\widetilde{\mathbf{T}}_{n+2}^2 =& w_{n+2}^2\\
\widetilde{\mathbf{T}}_{n+3}^2 =& (\phi w_{n+2} + w_{n+3})^2\\
\widetilde{\mathbf{T}}_{n+4}^2 =& (\phi^2 w_{n+2} + \phi w_{n+3} + w_{n+4})^2\\
\widetilde{\mathbf{T}}_{n+5}^2 =& (\phi^3 w_{n+2} + \phi^2 w_{n+3} + \phi w_{n+4} + w_{n+5})^2\\
&\vdots\\
\widetilde{\mathbf{T}}_{N}^2 =& (\phi^{N-(n+2)} w_{n+2} + \phi^{N-(n+3)} w_{n+3} + \cdots + w_N)^2.
\end{aligned} \tag{3.3.8}$$

Thus,

$$m^{-1}\sum_{t=n+1}^{N}\widetilde{\mathbf{T}}_t^2 = m^{-1}\sum_{t=n+2}^{N} b_t w_t^2 + 2m^{-1}\sum_{t=n+2}^{N-1}\sum_{s=t+1}^{N} d_{t,s} w_t w_s,$$

where

$$b_t = 1 + \phi^2 + \phi^4 + \cdots + \phi^{2(N-t)} = \sum_{j=0}^{N-t}\phi^{2j},$$

$$d_{t,s} = \sum_{j=0}^{N-s}\phi^{(s-t)+2j}.$$

By careful examination of (3.3.8), the expression of b_t can easily be verified. Also, one can see that the cross product terms of w's in $\sum_{t=n+1}^{N}\widetilde{\mathbf{T}}_t^2$, for $n+2 \le t < s \le N$, are

$$(\phi^{s-t}w_t)w_s,\ (\phi^{s-t+1}w_t)(\phi w_s),\ (\phi^{s-t+2}w_t)(\phi^2 w_s),\ \cdots,\ (\phi^{N-t}w_t)(\phi^{N-s}w_s).$$

Thus,

$$d_{t,s} = \phi^{s-t} + \phi^{s-t+2} + \phi^{s-t+4} + \cdots + \phi^{2N-t-s}.$$

The above gives the desired expression for $d_{t,s}$ by noting that $2N - t - s = (s-t) + 2(N-s)$.

Next, by Theorem 3.3.2,

$$\frac{\sum_{t=n+2}^{N} b_t w_t^2}{Q \sum_{t=n+2}^{N} b_t} \xrightarrow{p} 1.$$

This follows because $\mathrm{E}(b_t w_t^2) = b_t \mathcal{Q}$ and

$$\begin{aligned}\frac{1}{(m-1)^2} \sum_{t=n+2}^{N} \mathrm{Var}(b_t w_t^2) =& \frac{1}{(m-1)^2} \sum_{t=n+2}^{N} b_t^2 \mathrm{Var}(w_t^2) \\ =& o(1),\end{aligned}$$

since b_t is uniformly bounded by $(1-\phi^2)^{-1}$ and

$$\mathrm{Var}(w_t^2) = \mathrm{E}(w_t^4) - [\mathrm{E}(w_t^2)]^2 = 3\mathcal{Q}^2 - \mathcal{Q}^2 = 2\mathcal{Q}^2.$$

Also, note that

$$\begin{aligned}\mathrm{E}\left(\sum_{t=n+1}^{N} \widetilde{\mathbf{T}}_t^2\right) &= \sum_{t=n+2}^{N} b_t \mathrm{E}(w_t^2) + 2 \sum_{t=n+2}^{N-1} \sum_{s=t+1}^{N} d_{t,s} \mathrm{E}(w_t w_s) \\ &= Q \sum_{t=n+2}^{N} b_t.\end{aligned}$$

Since $\mathrm{E}(\sum_{t=n+1}^{N} \widetilde{\mathbf{T}}_t^2) = \sum_{t=n+1}^{N} \gamma_t$ by Lemma 2,

$$Q \sum_{t=n+2}^{N} b_t = \sum_{t=n+1}^{N} \gamma_t$$

and thus

$$\frac{\sum_{t=n+2}^{N} b_t w_t^2}{\sum_{t=n+1}^{N} \gamma_t} \xrightarrow{p} 1.$$

To complete the proof of the third convergence result, one still needs to show that

$$m^{-1}\sum_{t=n+2}^{N-1}\sum_{s=t+1}^{N} d_{t,s}w_t w_s \xrightarrow{p} 0.$$

The above result can be obtained using Theorem 3.3.2 by first noting that

$$m^{-1}\sum_{t=n+2}^{N-1}\sum_{s=t+1}^{N} d_{t,s}w_t w_s = \left(\frac{m-2}{m}\right)(m-2)^{-1}\sum_{t=n+2}^{N-1}\left(w_t\sum_{s=t+1}^{N} d_{t,s}w_s\right).$$

Also, since the w_t's are i.i.d. with mean 0, for $t \neq k$,

$$\begin{aligned}
&\mathrm{Cov}\left(w_t\sum_{i=t+1}^{N} d_{t,i}w_i,\ w_k\sum_{j=k+1}^{N} d_{k,j}w_j\right)\\
=&\mathrm{E}\left[w_t w_k\left(\sum_{i=t+1}^{N} d_{t,i}w_i\right)\left(\sum_{j=k+1}^{N} d_{k,j}w_j\right)\right] - \mathrm{E}\left(w_t\sum_{i=t+1}^{N} d_{t,i}w_i\right)\mathrm{E}\left(w_k\sum_{j=k+1}^{N} d_{k,j}w_j\right)\\
=&\mathrm{E}\left[w_t w_k\left(\sum_{i=t+1}^{N} d_{t,i}w_i\right)\left(\sum_{j=k+1}^{N} d_{k,j}w_j\right)\right]\\
=&\begin{cases}\mathrm{E}(w_t)\mathrm{E}\left[w_k\left(\sum_{i=t+1}^{N} d_{t,i}w_i\right)\left(\sum_{j=k+1}^{N} d_{k,j}w_j\right)\right], & \text{if } t<k\\ \mathrm{E}(w_k)\mathrm{E}\left[w_t\left(\sum_{i=t+1}^{N} d_{t,i}w_i\right)\left(\sum_{j=k+1}^{N} d_{k,j}w_j\right)\right], & \text{if } k<t\end{cases}\\
=&0.
\end{aligned}$$

This means that the variables $w_t\sum_{s=t+1}^{N} d_{t,s}w_s$ for $t = n+2, \cdots, N-1$ are uncorrelated and so Theorem 3.3.2 can be used on this set of variables. Now,

$$\mathrm{E}\left(w_t\sum_{s=t+1}^{N} d_{t,s}w_s\right) = 0$$

and

$$\begin{aligned}
\mathrm{Var}\left(w_t\sum_{s=t+1}^{N} d_{t,s}w_s\right) &= \mathrm{Var}(w_t)\sum_{s=t+1}^{N}\mathrm{Var}(d_{t,s}w_s)\\
&= \mathcal{Q}^2\sum_{s=t+1}^{N} d_{t,s}^2.
\end{aligned}$$

Note that for $t < s \leq N$,

$$\begin{aligned} d_{t,s}^2 = \left(\sum_{j=0}^{N-s} \phi^{(s-t)+2j} \right)^2 &= \left(\left| \sum_{j=0}^{N-s} \phi^{(s-t)+2j} \right| \right)^2 \\ &\leq \left(\sum_{j=0}^{N-s} |\phi^{(s-t)+2j}| \right)^2 \\ &= \left(\sum_{j=0}^{N-s} |\phi|^{(s-t)+2j} \right)^2 \\ &\leq \left(\sum_{j=0}^{\infty} |\phi|^{(s-t)+2j} \right)^2 . \end{aligned}$$

Hence, under (C.1), where $|\phi| < 1$,

$$\begin{aligned} \sum_{s=t+1}^{N} d_{t,s}^2 \leq \sum_{s=t+1}^{\infty} d_{t,s}^2 &\leq \sum_{s=t+1}^{\infty} \left(\sum_{j=0}^{\infty} |\phi|^{(s-t)+2j} \right)^2 \\ &= \sum_{k=1}^{\infty} \left(\sum_{j=0}^{\infty} |\phi|^{k+2j} \right)^2 \\ &= \sum_{k=1}^{\infty} \left(|\phi|^k \sum_{j=0}^{\infty} |\phi|^{2j} \right)^2 \\ &= \sum_{k=1}^{\infty} \left(\frac{|\phi|^k}{1-\phi^2} \right)^2 \\ &= (1-\phi^2)^{-2} \sum_{k=1}^{\infty} |\phi|^{2k} \\ &= (1-\phi^2)^{-2} \left(\frac{1}{1-\phi^2} - 1 \right) \\ &= \frac{\phi^2}{(1-\phi^2)^3}. \end{aligned}$$

The above implies that $\sum_{s=t+1}^{N} d_{t,s}^2$ is bounded, even as $N \to \infty$.

Hence,

$$\begin{aligned}
&(m-2)^{-2}\sum_{t=n+2}^{N-1}\mathrm{Var}\left(w_t\sum_{s=t+1}^{N}d_{t,s}w_s\right)\\
=&\frac{\mathcal{Q}^2}{m-2}(m-2)^{-1}\sum_{t=n+2}^{N-1}\left(\sum_{s=t+1}^{N}d_{t,s}^2\right)\\
=&o(1)
\end{aligned}$$

and thus

$$(m-2)^{-1}\sum_{t=n+2}^{N-1}\left(w_t\sum_{s=t+1}^{N}d_{t,s}w_s\right)\xrightarrow{p}0.$$

Furthermore, $(m-2)/m\to 1$ as $m\to\infty$, thus

$$m^{-1}\sum_{t=n+2}^{N-1}\sum_{s=t+1}^{N}d_{t,s}w_tw_s=\left(\frac{m-2}{m}\right)(m-2)^{-1}\sum_{t=n+2}^{N-1}\left(w_t\sum_{s=t+1}^{N}d_{t,s}w_s\right)\xrightarrow{p}0.$$

Therefore,

$$\frac{\sum_{t=n+1}^{N}\widetilde{\mathbf{T}}_t^2}{\sum_{t=n+1}^{N}\gamma_t}\xrightarrow{p}1.$$

Proof of the fourth result:

$$\begin{aligned}
&m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_t^2\\
=&m^{-1}\sum_{t=n+1}^{N}(\mu_t+\widetilde{\mathbf{T}}_t)^2\\
=&m^{-1}\sum_{t=n+1}^{N}\mu_t^2+2m^{-1}\sum_{t=n+1}^{N}\mu_t\widetilde{\mathbf{T}}_t+m^{-1}\sum_{t=n+1}^{N}\widetilde{\mathbf{T}}_t^2.
\end{aligned}$$

By Lemma 1, the first term above is bounded and by the first convergence result, the second term

$$m^{-1}\sum_{t=n+1}^{N}\mu_t\widetilde{\mathbf{T}}_t\xrightarrow{p}0.$$

By the third convergence result, as $m \to \infty$

$$\frac{\sum_{t=n+1}^{N} \widetilde{\mathbf{T}}_t^2}{\sum_{t=n+1}^{N} \gamma_t} \xrightarrow{p} 1.$$

Thus,

$$\frac{\sum_{t=n+1}^{N} \mathbf{T}_t^2}{\sum_{t=n+1}^{N} (\mu_t^2 + \gamma_t)} \xrightarrow{p} 1.$$

Proof of the fifth result:

$$(m-1)^{-1} \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}} \right) = (m-1)^{-1} \begin{bmatrix} \sum_{t=n+2}^{N} \mathbf{T}_{t-1}^2 & \sum_{t=n+2}^{N} \mathbf{T}_{t-1} \mathbf{F}_t^{\mathrm{T}} \\ \sum_{t=n+2}^{N} \mathbf{T}_{t-1} \mathbf{F}_t & \sum_{t=n+2}^{N} \mathbf{F}_t \mathbf{F}_t^{\mathrm{T}} \end{bmatrix}.$$

By the fourth convergence result,

$$\frac{\sum_{t=n+2}^{N} \mathbf{T}_{t-1}^2}{\sum_{t=n+2}^{N} (\mu_{t-1}^2 + \gamma_{t-1})} \xrightarrow{p} 1.$$

Next, $(m-1)^{-1} \sum_{t=n+2}^{N} \mathbf{F}_t \mathbf{F}_t^{\mathrm{T}}$ is finite by Lemma 1. Also, note that

$$\begin{aligned} &(m-1)^{-1} \sum_{t=n+2}^{N} \mathbf{T}_{t-1} \mathbf{F}_t \\ =&(m-1)^{-1} \sum_{t=n+2}^{N} \mu_{t-1} \mathbf{F}_t + (m-1)^{-1} \sum_{t=n+2}^{N} \mathbf{F}_t \widetilde{\mathbf{T}}_{t-1}. \end{aligned}$$

The first term is bounded by Lemma 1. The second term goes to a zero vector as $m \to \infty$ by the same arguments used to prove the first convergence result. Hence, as $m \to \infty$,

$$\left(\sum_{t=n+2}^{N} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1} \mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1} \mathbf{F}_t & \mathbf{F}_t \mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \right)^{-1} \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}} \right) \xrightarrow{p} \mathbf{I}.$$

This completes the proof of Lemma 3.

■

Lemma 4 *Let* $\lambda = [\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_{k+1}]^T \in \mathbb{R}^{k+1}$. *Then, for* $t \geq n+2$,

$$\mathrm{E}(\lambda^T D_t w_t) = 0$$

and

$$\mathrm{Var}(\lambda^{\mathrm{T}} D_t w_t) = \lambda^{\mathrm{T}} \mathcal{Q} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1}\mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1}\mathbf{F}_t & \mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \lambda.$$

□

Proof: Since $\mathbf{T}_{t-1}$ is independent of w_t,

$$\mathrm{E}(\lambda^{\mathrm{T}} D_t w_t) = \lambda^{\mathrm{T}} \mathrm{E}(D_t)\mathrm{E}(w_t) = 0$$

and

$$\begin{aligned} \mathrm{Var}(D_t w_t) &= \mathrm{E}(w_t^2 D_t D_t^{\mathrm{T}}) \\ &= \mathcal{Q}\mathrm{E}(D_t D_t^{\mathrm{T}}) \\ &= \mathcal{Q} \begin{bmatrix} \mathrm{E}(\mathbf{T}_{t-1}^2) & \mathrm{E}(\mathbf{T}_{t-1}\mathbf{F}_t^{\mathrm{T}}) \\ \mathrm{E}(\mathbf{T}_{t-1}\mathbf{F}_t) & \mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} &\mathrm{E}(\mathbf{T}_{t-1}^2) = \mu_{t-1}^2 + \gamma_{t-1}, \\ &\mathrm{E}(\mathbf{T}_{t-1}\mathbf{F}_t) = \mu_{t-1}\mathbf{F}_t. \end{aligned}$$

Therefore,

$$\mathrm{Var}(\lambda^{\mathrm{T}} D_t w_t) = \lambda^{\mathrm{T}} \mathcal{Q} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1}\mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1}\mathbf{F}_t & \mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \lambda.$$

This completes the proof of Lemma 4.

■

Theorem 3.3.3 ***Martingale central limit theorem***

Consider a sequence of variables $\{\mathcal{Y}_t, t \geq 1\}$ *such that,*

$$\mathrm{E}(\mathcal{Y}_t) = 0, \quad \mathrm{E}(\mathcal{Y}_t^2) < \infty \quad \textit{and} \quad \mathrm{E}(\mathcal{Y}_t|\mathcal{Y}_{t-1}, \mathcal{Y}_{t-2}, \cdots, \mathcal{Y}_1) = 0.$$

Also let $s_v^2 = \sum_{t=1}^{v} \mathrm{E}(\mathcal{Y}_t^2)$. *Then if, as* $v \longrightarrow \infty$,

$$\frac{\sum_{t=1}^{v} \mathrm{E}(\mathcal{Y}_t^2|\mathcal{Y}_{t-1}, \mathcal{Y}_{t-2}, \cdots, \mathcal{Y}_1)}{s_v^2} \xrightarrow{p} 1,$$

and if $\{\mathcal{Y}_t, t \geq 1\}$ *satisfy the Lindeberg condition, i.e., for every* $\varepsilon > 0$, *as* $v \longrightarrow \infty$,

$$\frac{\sum_{t=1}^{v} \int_{|\mathcal{Y}|>\varepsilon s_v} \mathcal{Y}^2 dF_t(\mathcal{Y})}{s_v^2} \longrightarrow 0,$$

where $s_v = \sqrt{s_v^2}$, *then*

$$v^{-1/2} \sum_{t=1}^{v} \mathcal{Y}_t \xrightarrow{d} \mathrm{N}(0, s_v^2/v).$$

□

Proof: See Brown (1971).

The Lindeberg condition in the above theorem can be implied from the Lyapunov's condition, which impose the condition that, as $v \longrightarrow \infty$,

$$s_v^{-(2+\kappa)} \sum_{t=1}^{v} \mathrm{E}(|\mathcal{Y}_t|^{2+\kappa}) \longrightarrow 0$$

for some $\kappa > 0$. In practice, this condition is more tractable than the Lindeberg condition. If $\mathrm{E}(|\mathcal{Y}_t|^{2+\kappa})$ is uniformly bounded in t, the Lyapunov's condition is satisfied provided that $v^{-1} s_v^{2+\kappa} \longrightarrow \infty$ as $v \longrightarrow \infty$.

Lemma 5 *The sequence* $\{\lambda^{\mathrm{T}} D_t w_t, t \geq n+2\}$ *satisfies*

$$\left(\lambda^{\mathrm{T}} \mathcal{Q} \sum_{t=n+2}^{N} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1}\mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1}\mathbf{F}_t & \mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \lambda \right)^{-1/2} \left(\lambda^{\mathrm{T}} \sum_{t=n+2}^{N} D_t w_t \right) \xrightarrow{d} \mathrm{N}(0,1).$$

□

Proof: Consider partitioning the vector λ into λ_1 and $\tilde{\lambda}$, such that $\tilde{\lambda} = [\lambda_2 \quad \lambda_3 \quad \ldots \quad \lambda_{k+1}]^{\mathrm{T}}$. First, consider the trivial case where $\lambda_1 = 0$. In this case, $\lambda^{\mathrm{T}} D_t w_t = \tilde{\lambda}^{\mathrm{T}} \mathbf{F}_t w_t$. Since the w_t's are i.i.d. normal random variables, the $\lambda^{\mathrm{T}} D_t w_t$'s are, in fact, independent and normally distributed. Hence,

$$\lambda^{\mathrm{T}} \sum_{t=n+2}^{N} D_t w_t \sim \mathrm{N}\left(0, \mathcal{Q} \sum_{t=n+2}^{N} (\tilde{\lambda}^{\mathrm{T}} \mathbf{F}_t)^2\right).$$

Now, consider the case where $\lambda_1 \neq 0$. In this case, the martingale central limit theorem (Theorem 3.3.3) is used. To apply Theorem 3.3.3, one needs to show that the sequence $\{\lambda^{\mathrm{T}} D_t w_t, t \geq n+2\}$ satisfies the conditions in the theorem. To begin, by Lemma 4,

$$\mathrm{E}(\lambda^{\mathrm{T}} D_t w_t) = 0.$$

Also, one has,

$$\begin{aligned}
&\mathrm{Var}(\lambda^{\mathrm{T}} D_t w_t) \\
=&\mathrm{E}\left[(\lambda^{\mathrm{T}} D_t w_t)^2\right] \\
=&\lambda^{\mathrm{T}} \mathcal{Q} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1}\mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1}\mathbf{F}_t & \mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \lambda \\
=&\mathcal{Q}\left\{\lambda_1^2[\mu_{t-1}^2 + \gamma_{t-1}] + 2\lambda_1\mu_{t-1}\tilde{\lambda}^{\mathrm{T}}\mathbf{F}_t + \tilde{\lambda}^{\mathrm{T}}\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}\tilde{\lambda}\right\} \\
=&\mathcal{Q}\left[\lambda_1^2\gamma_{t-1} + (\lambda_1\mu_{t-1} + \tilde{\lambda}^{\mathrm{T}}\mathbf{F}_t)^2\right] \\
<&\infty.
\end{aligned}$$

Now, let $\mathfrak{K}_{t-1}$ be a set of random variable such that $\mathfrak{K}_{t-1} = \{\lambda D_{t-1} w_{t-1},\ \lambda D_{t-2} w_{t-2},\ \cdots,\ \lambda D_{n+2} w_{n+2}\}$. Since $D_t = [\mathbf{T}_{t-1} \quad \mathbf{F}_t]^{\mathrm{T}}$, the set $\mathfrak{K}_{t-1}$ can be expressed as:

$$\begin{aligned}
\lambda^{\mathrm{T}} D_{t-1} w_{t-1} &= (\lambda_1 \mathbf{T}_{t-2} + \tilde{\lambda}^{\mathrm{T}} \mathbf{F}_{t-1}) w_{t-1} \\
\lambda^{\mathrm{T}} D_{t-2} w_{t-2} &= (\lambda_1 \mathbf{T}_{t-3} + \tilde{\lambda}^{\mathrm{T}} \mathbf{F}_{t-2}) w_{t-2} \\
&\vdots \\
\lambda^{\mathrm{T}} D_{n+3} w_{n+3} &= (\lambda_1 \mathbf{T}_{n+2} + \tilde{\lambda}^{\mathrm{T}} \mathbf{F}_{n+3}) w_{n+3} \\
\lambda^{\mathrm{T}} D_{n+2} w_{n+2} &= (\lambda_1 \mathbf{T}_{n+1} + \tilde{\lambda}^{\mathrm{T}} \mathbf{F}_{n+2}) w_{n+2}.
\end{aligned}$$

Since $\mathbf{T}_{n+1}$ is given (see (3.3.4)), w_{n+2} is fixed when given $\lambda^{\mathrm{T}} D_{n+2} w_{n+2}$. Since $\mathbf{T}_{n+2} = \phi \mathbf{T}_{n+1} + \delta \mathbf{F}_{n+2} + w_{n+2}$, w_{n+3} is fixed when given the set $\{\lambda^{\mathrm{T}} D_{n+2} w_{n+2}, \lambda^{\mathrm{T}} D_{n+3} w_{n+3}\}$. In general, when given $\mathfrak{K}_{t-1}$, the set $\{w_{t-1}, w_{t-2}, \cdots, w_{n+2}\}$ is fixed. So,

$$\mathrm{E}(D_t | \mathfrak{K}_{t-1}) = D_t$$

because $D_t = [\mathbf{T}_{t-1} \quad \mathbf{F}_t]^{\mathrm{T}}$ and

$$\mathbf{T}_{t-1} = \begin{cases} \mathbf{T}_{n+1}, & \text{if } t-1 = n+1 \\ \phi^{t-1-(n+1)} \mathbf{T}_{n+1} + \sum_{j=0}^{t-1-(n+2)} \phi^j (\delta \mathbf{F}_{t-1-j} + w_{t-1-j}), & \text{if } t-1 \geq n+2. \end{cases}$$

Thus

$$\begin{aligned}
\mathrm{E}(\lambda^{\mathrm{T}} D_t w_t | \mathfrak{K}_{t-1}) &= \lambda^{\mathrm{T}} \mathrm{E}(D_t | \mathfrak{K}_{t-1}) \mathrm{E}(w_t | \mathfrak{K}_{t-1}) \\
&= \lambda^{\mathrm{T}} D_t \mathrm{E}(w_t) \\
&= 0
\end{aligned}$$

and similarly

$$\begin{aligned}\sum_{t=n+2}^{v} \mathrm{E}\left[(\lambda^{\mathrm{T}} D_t w_t)^2 \middle| \mathfrak{K}_{t-1}\right] =& \lambda^{\mathrm{T}} \sum_{t=n+2}^{v} \mathrm{E}(D_t D_t^{\mathrm{T}} w_t^2 | \mathfrak{K}_{t-1}) \lambda \\ =& \lambda^{\mathrm{T}} \sum_{t=n+2}^{v} \mathrm{E}(D_t D_t^{\mathrm{T}} | \mathfrak{K}_{t-1}) \mathrm{E}(w_t^2 | \mathfrak{K}_{t-1}) \lambda \\ =& \lambda^{\mathrm{T}} \sum_{t=n+2}^{v} D_t D_t^{\mathrm{T}} \mathrm{E}(w_t^2) \lambda \\ =& \lambda^{\mathrm{T}} \mathcal{Q} \sum_{t=n+2}^{v} D_t D_t^{\mathrm{T}} \lambda.\end{aligned}$$

Therefore, by Lemma 3, as $v \longrightarrow \infty$,

$$\left(\lambda^{\mathrm{T}} \mathcal{Q} \sum_{t=n+2}^{v} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1} \mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1} \mathbf{F}_t & \mathbf{F}_t \mathbf{F}_t^{\mathrm{T}} \end{bmatrix} \lambda \right)^{-1} \sum_{t=n+2}^{v} \mathrm{E}[(\lambda^{\mathrm{T}} D_t w_t)^2 | \mathfrak{K}_{t-1}] \xrightarrow{p} 1.$$

Compare this with the result from Lemma 4, it is now apparent that $\{\lambda^{\mathrm{T}} D_t w_t, t \geq n+2\}$ satisfy the first convergence requirement in the martingale CLT.

Next, the sequence $\{\lambda^{\mathrm{T}} D_t w_t, t \geq n+2\}$ will be shown to satisfy the Lyapunov's condition with $\kappa = 2$. Consider

$$\begin{aligned}&\mathrm{E}(|\lambda^{\mathrm{T}} D_t w_t|^4) \\ =&\mathrm{E}\left[(\lambda_1 \mathbf{T}_{t-1} w_t + \tilde{\lambda}^{\mathrm{T}} \mathbf{F}_t w_t)^4\right] \\ =&\mathrm{E}\left[\lambda_1^4 \mathbf{T}_{t-1}^4 w_t^4 + (\tilde{\lambda}^{\mathrm{T}} \mathbf{F}_t)^4 w_t^4 + 4\lambda_1^3 \tilde{\lambda}^{\mathrm{T}} \mathbf{F}_t \mathbf{T}_{t-1}^3 w_t^4 + 4\lambda_1 (\tilde{\lambda}^{\mathrm{T}} \mathbf{F}_t)^3 \mathbf{T}_{t-1} w_t^4 + 6\lambda_1^2 (\tilde{\lambda}^{\mathrm{T}} \mathbf{F}_t)^2 \mathbf{T}_{t-1}^2 w_t^4\right].\end{aligned}$$

Note that $\mathbf{T}_{t-1}$ and w_t are independent. Also, $\mathrm{E}(w_t^4) = 3\mathcal{Q}^2$, $\mathrm{E}(\mathbf{T}_{t-1}) = \mu_{t-1}$, $\mathrm{E}(\mathbf{T}_{t-1}^2) = \mu_{t-1}^2 + \gamma_{t-1}$ and by Lemma 2,

$$\begin{aligned}\mathrm{E}(\mathbf{T}_{t-1}^3) =& \mathrm{E}\left[(\mu_{t-1} + \widetilde{\mathbf{T}}_{t-1})^3\right] \\ =& \mathrm{E}(\mu_{t-1}^3 + \widetilde{\mathbf{T}}_{t-1}^3 + 3\mu_{t-1}^2 \widetilde{\mathbf{T}}_{t-1} + 3\mu_{t-1} \widetilde{\mathbf{T}}_{t-1}^2) \\ =& \mu_{t-1}^3 + 3\mu_{t-1}\gamma_{t-1}\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(\mathbf{T}_{t-1}^4) =& \mathrm{E}\left[(\mu_{t-1} + \widetilde{\mathbf{T}}_{t-1})^4\right] \\ =& \mathrm{E}(\mu_{t-1}^4 + \widetilde{\mathbf{T}}_{t-1}^4 + 4\mu_{t-1}^3\widetilde{\mathbf{T}}_{t-1} + 4\mu_{t-1}\widetilde{\mathbf{T}}_{t-1}^3 + 6\mu_{t-1}^2\widetilde{\mathbf{T}}_{t-1}^2) \\ =& \mu_{t-1}^4 + \mathrm{E}(\widetilde{\mathbf{T}}_{t-1}^4) + 6\mu_{t-1}^2\gamma_{t-1}.\end{aligned}$$

Since μ_t, γ_t and $\mathrm{E}(\widetilde{\mathbf{T}}_{t-1})^4$ are uniformly bounded in t, it is clear that $\mathrm{E}(|\lambda^{\mathrm{T}} D_t w_t|^4)$ is uniformly bounded in t. Let $s_v^2 = \sum_{t=n+2}^{v} \mathrm{Var}(\lambda^{\mathrm{T}} D_t w_t)$, then

$$\begin{aligned}s_v^{2+2} &= \left[\sum_{t=n+2}^{v} \mathrm{Var}(\lambda^{\mathrm{T}} D_t w_t)\right]^2 \\ &= \left[\mathcal{Q} \sum_{t=n+2}^{v} \left(\lambda_1^2\gamma_{t-1} + (\lambda_1\mu_{t-1} + \tilde{\lambda}^{\mathrm{T}}\mathbf{F}_t)^2\right)\right]^2.\end{aligned}$$

Therefore,

$$\begin{aligned}(v-n-1)^{-1}s_v^{2+2} &= \left[(v-n-1)^{-1/2}\mathcal{Q} \sum_{t=n+2}^{v} \left(\lambda_1^2\gamma_{t-1} + (\lambda_1\mu_{t-1} + \tilde{\lambda}^{\mathrm{T}}\mathbf{F}_t)^2\right)\right]^2 \\ &\geq \left[(v-n-1)^{-1/2}\mathcal{Q} \sum_{t=n+2}^{v} \lambda_1^2\gamma_{t-1}\right]^2 \\ &\geq \left[(v-n-1)^{-1/2}\mathcal{Q}\lambda_1^2(v-n-2)\mathcal{Q}\right]^2 \\ &= \left[(v-n-1)^{1/2}\left(\frac{v-n-2}{v-n-1}\right)\mathcal{Q}^2\lambda_1^2\right]^2\end{aligned}$$

It is now clear that $(v-n-1)^{-1}s_v^{2+2} \longrightarrow \infty$ as $v \longrightarrow \infty$. Hence, the sequence $\{\lambda^{\mathrm{T}} D_t w_t, t \geq n+2\}$ satisfies the Lyapunov's condition with $\kappa = 2$, which implies the Lindeberg condition is also satisfied.

■

Proof of Theorem 3.3.1:

Theorem 3.3.1, the asymptotic distribution of the maximum likelihood estimators obtained using the $\mathbb{CAL}$ likelihood $\ell_{\mathbf{c}}(\Theta)$, will be proved next using the lemmas given before.

Asymptotic distribution of $\hat{\mathcal{A}}$:

To begin, note that

$$\begin{aligned}\hat{\mathcal{A}} &= \frac{\sum_{t=n+1}^{N} \mathbf{T}_t \mathbf{P}_t}{\sum_{t=n+1}^{N} \mathbf{T}_t^2} \\ &= \frac{\sum_{t=n+1}^{N} \mathbf{T}_t (\mathcal{A}\mathbf{T}_t + e_t)}{\sum_{t=n+1}^{N} \mathbf{T}_t^2} \\ &= \mathcal{A} + \frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sum_{t=n+1}^{N} \mathbf{T}_t^2}.\end{aligned}$$

The e_t's are i.i.d. normal random variables and are independent of the $\mathbf{T}_t$'s. If one conditions on $\mathbf{T}_{n+1}, \mathbf{T}_{n+2}, \cdots, \mathbf{T}_N$,

$$\sum_{t=n+1}^{N} \mathbf{T}_t e_t \sim \mathrm{N}(0, \mathcal{R} \sum_{t=n+1}^{N} \mathbf{T}_t^2).$$

Thus, conditioning on $\mathbf{T}_{n+1}, \mathbf{T}_{n+2}, \cdots, \mathbf{T}_N$,

$$\frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sqrt{\mathcal{R} \sum_{t=n+1}^{N} \mathbf{T}_t^2}} \sim \mathrm{N}(0,1),$$

where the distribution does not depend on the $\mathbf{T}$'s. This implies that the unconditional distribution of the left hand side of the above is also N(0,1). Next, by Lemma 3,

$$\frac{\sum_{t=n+1}^{N} \mathbf{T}_t^2}{\sum_{t=n+1}^{N} (\mu_t^2 + \gamma_t)} \xrightarrow{p} 1.$$

Now,

$$\begin{aligned}\hat{\mathcal{A}} - \mathcal{A} &= \frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sum_{t=n+1}^{N} \mathbf{T}_t^2} \\ &= \left(\frac{\mathcal{R}}{\sum_{t=n+1}^{N} (\mu_t^2 + \gamma_t)} \right)^{1/2} \left(\frac{\sum_{t=n+1}^{N} (\mu_t^2 + \gamma_t)}{\sum_{t=n+1}^{N} \mathbf{T}_t^2} \right)^{1/2} \frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sqrt{\mathcal{R} \sum_{t=n+1}^{N} \mathbf{T}_t^2}}.\end{aligned}$$

Thus, it is clear that

$$\left(\frac{\mathcal{R}}{\sum_{t=n+1}^{N}(\mu_t^2+\gamma_t)}\right)^{-1/2}(\hat{\mathcal{A}}-\mathcal{A}) \xrightarrow{d} \mathrm{N}(0,1). \tag{3.3.9}$$

Also note that, from (3.3.6),

$$\frac{\partial \ell_{\mathbf{c}}(\Theta)}{\partial \mathcal{A}} = \frac{1}{\mathcal{R}}\sum_{t=n+1}^{N}\mathbf{T}_t(\mathbf{P}_t-\mathcal{A}\mathbf{T}_t).$$

Thus,

$$\begin{aligned}\mathfrak{I}_{\mathcal{A}} &= \mathrm{E}[-\partial^2\ell_{\mathbf{c}}(\Theta)/\partial\mathcal{A}^2]\\ &= \mathcal{R}^{-1}\sum_{t=n+1}^{N}\mathrm{E}(\mathbf{T}_t^2)\\ &= \mathcal{R}^{-1}\sum_{t=n+1}^{N}(\mu_t^2+\gamma_t).\end{aligned}$$

Therefore

$$\sqrt{\mathfrak{I}_{\mathcal{A}}}(\hat{\mathcal{A}}-\mathcal{A}) \xrightarrow{d} \mathrm{N}(0,1).$$

Asymptotic distribution of $\hat{\mathcal{R}}$:

$$\begin{aligned}\hat{\mathcal{R}} &= m^{-1}\sum_{t=n+1}^{N}(\mathbf{P}_t-\hat{\mathcal{A}}\mathbf{T}_t)^2\\ &= m^{-1}\sum_{t=n+1}^{N}(\mathbf{P}_t^2-2\hat{\mathcal{A}}\mathbf{T}_t\mathbf{P}_t+\hat{\mathcal{A}}^2\mathbf{T}_t^2)\\ &= m^{-1}\sum_{t=n+1}^{N}\left[(\mathcal{A}\mathbf{T}_t+e_t)^2-2\hat{\mathcal{A}}\mathbf{T}_t(\mathcal{A}\mathbf{T}_t+e_t)+\hat{\mathcal{A}}^2\mathbf{T}_t^2\right]\\ &= m^{-1}\sum_{t=n+1}^{N}\left[(\mathcal{A}^2+\hat{\mathcal{A}}^2-2\mathcal{A}\hat{\mathcal{A}})\mathbf{T}_t^2+2(A-\hat{\mathcal{A}})\mathbf{T}_te_t+e_t^2\right].\end{aligned}$$

By Lemma 3,

$$\begin{aligned}m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_t^2 &\xrightarrow{p} m^{-1}\sum_{t=n+1}^{N}(\mu_t^2+\gamma_t)<\infty,\\ m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_te_t &\xrightarrow{p} 0.\end{aligned}$$

Note that,

$$\sum_{t=n+1}^{N} (\mu_t^2 + \gamma_t) \geq \sum_{t=n+1}^{N} \gamma_t \geq (m-1)\mathcal{Q} \longrightarrow \infty \text{ as } m \longrightarrow \infty.$$

Thus, (3.3.9) implies that $\hat{\mathcal{A}} \xrightarrow{p} \mathcal{A}$. Hence, $\hat{\mathcal{R}} \xrightarrow{p} m^{-1} \sum_{t=n+1}^{N} e_t^2$. Note that the e_t's are i.i.d. random variables. Therefore, the central limit theorem can be applied to derive the asymptotic distribution of $m^{-1} \sum_{t=n+1}^{N} e_t^2$. Since

$$\begin{aligned}
&\mathrm{E}(e_t^2) = \mathrm{Var}(e_t) = \mathcal{R} \\
&\mathrm{Var}(e_t^2) = \mathrm{E}(e_t^4) - [\mathrm{E}(e_t^2)]^2 = 3\mathcal{R}^2 - \mathcal{R}^2 = 2\mathcal{R}^2,
\end{aligned}$$

it follows that

$$m^{-1} \sum_{t=n+1}^{N} e_t^2 \xrightarrow{d} \mathrm{N}(\mathcal{R}, 2\mathcal{R}^2/m)$$

and thus

$$\frac{\hat{\mathcal{R}} - R}{\sqrt{2\mathcal{R}^2/m}} \xrightarrow{d} \mathrm{N}(0,1).$$

Also note that

$$\begin{aligned}
\mathfrak{I}_{\mathcal{R}} &= \mathrm{E}[-\partial^2 \ell_{\mathbf{c}}(\Theta)/\partial \mathcal{R}^2] \\
&= \left(-\frac{m}{2\mathcal{R}^2} + \frac{1}{\mathcal{R}^3} \sum_{t=n+1}^{N} \mathrm{E}[(\mathbf{P}_t - \mathcal{A}\mathbf{T}_t)^2] \right) \\
&= \left(-\frac{m}{2\mathcal{R}^2} + \frac{m}{\mathcal{R}^2} \right) \\
&= \frac{m}{2\mathcal{R}^2}.
\end{aligned}$$

Thus, as $m \to \infty$,

$$\sqrt{\mathfrak{I}_{\mathcal{R}}}(\hat{\mathcal{R}} - \mathcal{R}) \xrightarrow{d} \mathrm{N}(0,1).$$

Asymptotic distribution of $\hat{\Upsilon}$***:***

$$\begin{aligned}\hat{\Upsilon} &= \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}}\right)^{-1} \sum_{t=n+2}^{N} D_t \mathbf{T}_t \\ &= \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}}\right)^{-1} \sum_{t=n+2}^{N} D_t (D_t^{\mathrm{T}} \Upsilon + w_t) \\ &= \Upsilon + \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}}\right)^{-1} \sum_{t=n+2}^{N} D_t w_t.\end{aligned}$$

Now let

$$\mathcal{U} = \sum_{t=n+2}^{N} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1}\mathbf{F}_t^{\mathrm{T}} \\ \mu_{t-1}\mathbf{F}_t & \mathbf{F}_t\mathbf{F}_t^{\mathrm{T}} \end{bmatrix}.$$

By Lemma 3,

$$\mathcal{U}^{-1}\left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}}\right) \xrightarrow{p} \mathbf{I}.$$

By Lemma 5,

$$(\lambda^{\mathrm{T}} \mathcal{Q}\mathcal{U}\lambda)^{-1/2}\left(\lambda^{\mathrm{T}} \sum_{t=n+2}^{N} D_t w_t\right) \xrightarrow{d} \mathrm{N}(0,1).$$

By the Cramer-Wold device (Theorem 3.6.6 in section 3.6.3),

$$(\mathcal{Q}\mathcal{U})^{-1/2}\left(\sum_{t=n+2}^{N} D_t w_t\right) \xrightarrow{d} \mathrm{MVN}(\vec{0}, \mathbf{I}). \tag{3.3.10}$$

Now consider

$$\begin{aligned}\hat{\Upsilon} - \Upsilon &= \left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}}\right)^{-1} \sum_{t=n+2}^{N} D_t w_t \\ &= \left[\left(\sum_{t=n+2}^{N} D_t D_t^{\mathrm{T}}\right)^{-1} \mathcal{U}\right] \mathcal{U}^{-1} (\mathcal{Q}\mathcal{U})^{1/2} \left[(\mathcal{Q}\mathcal{U})^{-1/2} \sum_{t=n+2}^{N} D_t w_t\right].\end{aligned}$$

Thus, it is clear that

$$\mathcal{Q}^{-1/2}\mathcal{U}^{1/2}(\hat{\Upsilon} - \Upsilon) \xrightarrow{d} \mathrm{MVN}(\vec{0}, \mathbf{I}). \tag{3.3.11}$$

Equivalently, as $m \to \infty$, $\sqrt{\mathfrak{I}_\Upsilon}(\hat{\Upsilon} - \Upsilon) \xrightarrow{d} \text{MVN}(\vec{0}, \mathbf{I})$ because

$$\begin{aligned}
\mathfrak{I}_\Upsilon &= \text{E}[-\partial^2 \ell_{\mathbf{c}}(\Theta)/\partial\Upsilon\ \partial\Upsilon^{\text{T}}] \\
&= \mathcal{Q}^{-1} \sum_{t=n+2}^{N} \text{E}(D_t D_t^{\text{T}}) \\
&= \mathcal{Q}^{-1} \sum_{t=n+2}^{N} \begin{bmatrix} \mu_{t-1}^2 + \gamma_{t-1} & \mu_{t-1}\mathbf{F}_t^{\text{T}} \\ \mu_{t-1}\mathbf{F}_t & \mathbf{F}_t\mathbf{F}_t^{\text{T}} \end{bmatrix} \\
&= \mathcal{Q}^{-1}\mathcal{U}.
\end{aligned}$$

Asymptotic distribution of $\hat{\mathcal{Q}}$

$$\begin{aligned}
\hat{\mathcal{Q}} &= (m-1)^{-1} \sum_{t=n+2}^{N} (\mathbf{T}_t - \hat{\Upsilon}^{\text{T}} D_t)^2 \\
&= (m-1)^{-1} \sum_{t=n+2}^{N} (\mathbf{T}_t^2 - 2\hat{\Upsilon}^{\text{T}} D_t \mathbf{T}_t + \hat{\Upsilon}^{\text{T}} D_t D_t^{\text{T}} \hat{\Upsilon}) \\
&= (m-1)^{-1} \sum_{t=n+2}^{N} \left[(\Upsilon^{\text{T}} D_t + w_t)^2 - 2\hat{\Upsilon}^{\text{T}} D_t (\Upsilon^{\text{T}} D_t + w_t) + \hat{\Upsilon}^{\text{T}} D_t D_t^{\text{T}} \hat{\Upsilon} \right] \\
&= (m-1)^{-1} \sum_{t=n+2}^{N} \left[(\Upsilon - \hat{\Upsilon})^{\text{T}} D_t D_t^{\text{T}} (\Upsilon - \hat{\Upsilon}) + 2(\Upsilon - \hat{\Upsilon})^{\text{T}} D_t w_t + w_t^2 \right]
\end{aligned}$$

By Lemma 3,

$$(m-1)^{-1} \sum_{t=n+2}^{N} D_t D_t^{\text{T}} \xrightarrow{p} (m-1)^{-1}\mathcal{U},$$

where the entries of $(m-1)^{-1}\mathcal{U}$ are bounded. Since the entries of $(m-1)^{-1}\mathcal{U}$ are bounded, (3.3.10) implies

$$(m-1)^{-1} \sum_{t=n+2}^{N} D_t w_t \xrightarrow{p} \vec{0}.$$

Also, under condition (C.3), the diagonal elements of $\mathcal{U}$, namely, $\sum_{t=n+2}^{N}(\mu_{t-1}^2 + \gamma_{t-1})$ and $\text{diag}(\sum_{t=n+2}^{N} \mathbf{F}_t\mathbf{F}_t^{\text{T}})$, all approach infinity as $m \longrightarrow \infty$. Thus, (3.3.11) implies $\hat{\Upsilon} \xrightarrow{p} \Upsilon$.

Therefore $\hat{\mathcal{Q}} \xrightarrow{p} (m-1)^{-1} \sum_{t=n+2}^{N} w_t^2$. Since

$$\mathrm{E}(w_t^2) = \mathcal{Q},$$
$$\mathrm{Var}(w_t^2) = 2\mathcal{Q}^2,$$

it follows that

$$(m-1)^{-1} \sum_{t=n+2}^{N} w_t^2 \xrightarrow{d} \mathrm{N}\left(\mathcal{Q}, 2\mathcal{Q}^2/(m-1)\right)$$

and thus

$$\frac{\hat{\mathcal{Q}} - Q}{\sqrt{2\mathcal{Q}^2/(m-1)}} \xrightarrow{d} \mathrm{N}(0,1). \tag{3.3.12}$$

Equivalently, as $m \to \infty, \sqrt{\mathfrak{I}_{\mathcal{Q}}}(\hat{\mathcal{Q}} - \mathcal{Q}) \xrightarrow{d} \mathrm{N}(0,1)$ because

$$\begin{aligned}
\mathfrak{I}_{\mathcal{Q}} &= \mathrm{E}[-\partial^2 \ell_{\mathbf{c}}(\Theta)/\partial \mathcal{Q}^2] \\
&= \left(-\frac{m-1}{2\mathcal{Q}^2} + \frac{1}{\mathcal{Q}^3} \sum_{t=n+2}^{N} \mathrm{E}[(\mathbf{T}_t - \Upsilon^{\mathrm{T}} D_t)^2] \right) \\
&= \left(-\frac{m-1}{2\mathcal{Q}^2} + \frac{m-1}{\mathcal{Q}^2} \right) \\
&= \frac{m-1}{2\mathcal{Q}^2}.
\end{aligned}$$

This completes the proof of Theorem 3.3.1. ■

3.4 Comparing the different estimators

Three different approaches have been presented in the earlier sections for estimating the parameters that specify a state-space model. A natural question that arises is which of these approaches provides the best estimator. Such a question can be answered from two different perspectives, namely from an asymptotic perspective (i.e., sample size goes to infinity) and from a finite sample size perspective.

From Theorem 3.2.1 and 3.3.1, one can see that estimators from the $\mathbb{PXY}$ and the $\mathbb{CAL}$ approaches are both asymptotically unbiased. Also, asymptotically, the estimator behaves like a normally distributed random variable. This is a very useful result because one can now easily construct confidence regions for the estimators to make inferences. In general, an unbiased estimator would not be considered to be useful if there exists another unbiased estimator that has smaller variance. The asymptotic variances for the estimators from the $\mathbb{PXY}$ and the $\mathbb{CAL}$ approaches are given in the earlier sections. However, it is not easy to compare these variances directly from a theoretical standpoint because the likelihood function $\ell_{\mathbf{P}}(\Theta)$ is a highly non-linear function of Θ. Thus, the estimator variance from the $\mathbb{PXY}$ approach $(\mathfrak{I}(\Theta_0)^{-1})$ cannot be written out explicitly. Therefore, one is restricted to comparing these variances empirically through simulation studies from a finite sample size perspective. Since the asymptotic distribution for the estimator from the $\mathbb{ALL}$ approach is not available, simulation studies can then be used to also provide important information on the quality of such estimator.

The rest of this section provides results from three simulation studies. Each simulation study is carried out by first simulating data under some specific structure for the observation and state processes. The different estimation approaches are used to provide estimates of the model parameters. This process is repeated a large number of times, each time using a new set of simulated data. The resulting set of estimated parameter values is then used to assess the quality of the different estimation approaches. The three simulation studies differ in how the observation and state processes are specified. The first study simulates data under the univariate Gaussian state-space model in (3.1.2). The other two studies investigate the robustness of the estimator with respect to departures from model assumptions. The first

of these two studies examines the robustness of each estimator with respect to observational noise structure. In particular, the effect of replacing white observational noise with red noise is considered. The second study investigates the robustness of each estimator with respect to state equation specification. In this study, the AR(1) state equation that is used to generate the simulated data is replaced by an AR(2) state equation.

3.4.1 Performance of estimator under the assumed state-space model

This section provides results from a simulation study under the assumed Gaussian state-space model. To be more realistic, the true parameter value of the state-space model used in the simulation is chosen to mimic the reconstruction problem. From Figure 3.3, the PACF at lag 1 is about 0.6, suggesting a reasonable choice for the parameter ϕ would be 0.6. Thus, the exogenous variable $\mathbf{F}_t$ equals to $\mathbf{X}_t - 0.6\mathbf{X}_{t-1}$. More detail on the choice of ϕ in real-world application is given in Chapter 4. The variable $\mathbf{X}_t$ is chosen to be the estimated combined annual response to greenhouse gas forcing, sulphate aerosol forcing, volcanic forcing and solar forcing. In another words, $\mathbf{X}_t = \mathbf{GS}_t + \mathbf{SOL}_t + \mathbf{VOL}_t$. The estimated response is obtained from an energy balanced model (EBM) driven with reconstructed solar, volcanic and anthropogenic forcings (Hegerl et al. 2003, 2007a) . The EBM simulation used here is the same as that used in Hegerl et al. (2007a), from which the 30N-90N average response to greenhouse gas, sulfate aerosol, volcanic and solar forcing are available from 1000-1997. It is further assumed that the amplitude of the response is correct and hence $\delta = 1$. From experiments conducted in the next chapter with climate model data (not shown), a realistic choice of $\mathcal{Q}$ would be around 0.02 and for $\mathcal{R}$ would be between 0.02 and 0.1. So, $\mathcal{Q} = 0.02$ and $\mathcal{R} = 0.1$ are used here in the simulation. For simplicity, the value of $\mathcal{A}$ is chosen to be 1. Thus, the state-space model used to generate simulated data is,

$$\begin{aligned} \mathbf{P}_t &= \mathbf{T}_t + e_t, & e_t &\sim \mathrm{N}(0, 0.1), \\ \mathbf{T}_t &= 0.6\mathbf{T}_{t-1} + \mathbf{F}_t + w_t, & w_t &\sim \mathrm{N}(0, 0.02). \end{aligned}$$

A suite of simulations has been conducted to investigate the effect of sample size on the performance of the estimators. In particular, the choices for the sample size are $N = 200$, 500

and 998. These sample sizes correspond to the year 1798-1997, 1498-1997 and 1000-1997 for the exogenous variable $\mathbf{F}_t$. The maximum sample size is 998 because the exogenous variable is only available for years 1000-1997. On the other hand, the length of the calibration period is fixed at $m = 100$ for all simulations to reflect the length of the calibration period in real-world analyses. One thousand runs are conducted for each choice of sample size. In each run, the parameters of the state-space model are estimated through maximum likelihood estimation using the three different likelihood functions, namely, $\mathbb{PXY}$, $\mathbb{ALL}$ and $\mathbb{CAL}$. For the $\mathbb{PXY}$ and $\mathbb{ALL}$ likelihood, the parameters are solved numerically using EM algorithm and the algorithm is stopped when the successive change in the log likelihood function is less than 0.0005. In the EM algorithm, the true parameter value is used as the initial value for the parameters. This is rather an idealized situation since one do not know the true parameter values in real-world application and an improper initial value can result in convergence to a non-global maximum because both the $\mathbb{PXY}$ and $\mathbb{ALL}$ likelihood functions are highly non-linear. So, results obtained here for the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches are biased in the sense that information which is not available in real-world application is used. Suggestions on the initial value choice in real-world analysis will be given in the concluding section. Note that for the $\mathbb{CAL}$ likelihood, regardless of the value of N, only data from the last $m = 100$ years are used in the estimation process.

Two measures are considered to assess the finite sample performance of the estimators: biasedness and efficiency. In the following results, the bias of an estimator $\hat{\theta}$ is expressed as the percentage of absolute deviation from the true parameter value and is defined as

$$\text{bias} = \left| \frac{\theta_0 - \frac{1}{M}\sum_{i=1}^{M} \hat{\theta}_i}{\theta_0} \right| \times 100,$$

where M is the number of simulation runs, θ_0 is the true value of θ and $\hat{\theta}_i$ is the estimated value of θ at the i^{th} run. The hat notation ($\hat{}$) will be used throughout this section to denote an estimated value of a particular parameter. Let $\theta^{(j)}$ be an unbiased estimator of θ from the j^{th} approach, then the efficiency of $\theta^{(j)}$ relative to an unbiased baseline estimator $\theta^{(B)}$ is

$$\text{efficiency} = \frac{\text{Var}(\theta^{(j)})}{\text{Var}(\theta^{(B)})}.$$

In this simulation study, $\text{Var}(\theta^{(j)})$ is obtained by calculating the sample variance from $\hat{\theta}_1^{(j)}, \cdots, \hat{\theta}_M^{(j)}$, where $\hat{\theta}_i^{(j)}$ is the estimate obtained at the i^{th} run of a simulation from the j^{th} approach. The baseline value is chosen to be the sample variance obtained from the $\mathbb{ALL}$ approach. As pointed out earlier, an unbiased estimator would not be consider to be useful if there exists another unbiased estimator that has smaller variance. Hence, if efficiency is greater than 1, the estimator obtained from the j^{th} approach is considered to be less efficient than that obtained from the $\mathbb{ALL}$ approach and vice versa. However, such a statement is only valid if the estimator from the $\mathbb{ALL}$ approach is found to be unbiased.

Since the primary goal of using the state-space model is to estimate the state process using the Kalman filter (or smoother), it is necessary to assess the reliability of the Kalman filter estimates obtained by using the parameters estimated from the different approaches. The mean square error (MSE) between the Kalman filter estimates and the simulated state process can be used for such purpose. The MSE for the j^{th} approach over a simulation with M runs is defined as

$$\text{MSE}^{(j)} = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{N}\sum_{t=1}^{N}\left(\mathbf{T}_{t|t-1}^{(ij)} - \mathbf{T}_t^{(i)}\right)^2\right],$$

where $\mathbf{T}_{t|t-1}^{(ij)}$ is the Kalman filter estimate obtained by using parameters estimated from the j^{th} approach at the i^{th} run of the simulation and $\mathbf{T}_t^{(i)}$ is the simulated state process at the i^{th} run. A smaller MSE means better Kalman filter estimates.

Figure 3.4 displays the bias of the estimators obtained from the different likelihood functions with varying sample size. The bias of $\hat{\mathcal{A}}$ is not shown for the $\mathbb{PXY}$ approach because this approach does not allow the estimation of $\mathcal{A}$. It is estimated with least square regression using the calibration period data. Thus the estimate will be identical to that obtained from the $\mathbb{CAL}$ approach. From this figure, one can see that the bias of all estimators from the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches became smaller as the sample size increases, providing empirical evidence that the estimators are asymptotically unbiased. An exception is the estimate of $\mathcal{A}$ from the $\mathbb{ALL}$ approach, for which the bias is somewhat similar across the different sample sizes. This is because the estimation of $\mathcal{A}$ relies mainly on the calibration period data. Given

that the length of the calibration period is the same across the simulations, it is not surprising to see no improvement in the biasedness of $\hat{\mathcal{A}}$ when the sample size N increases. It should be emphasized that the sample size for the $\mathbb{CAL}$ approach is fixed at m=100 regardless of N and so the bias of the estimates from this approach is very similar across the different N. The small difference in bias across different N is merely a result of sampling variability.

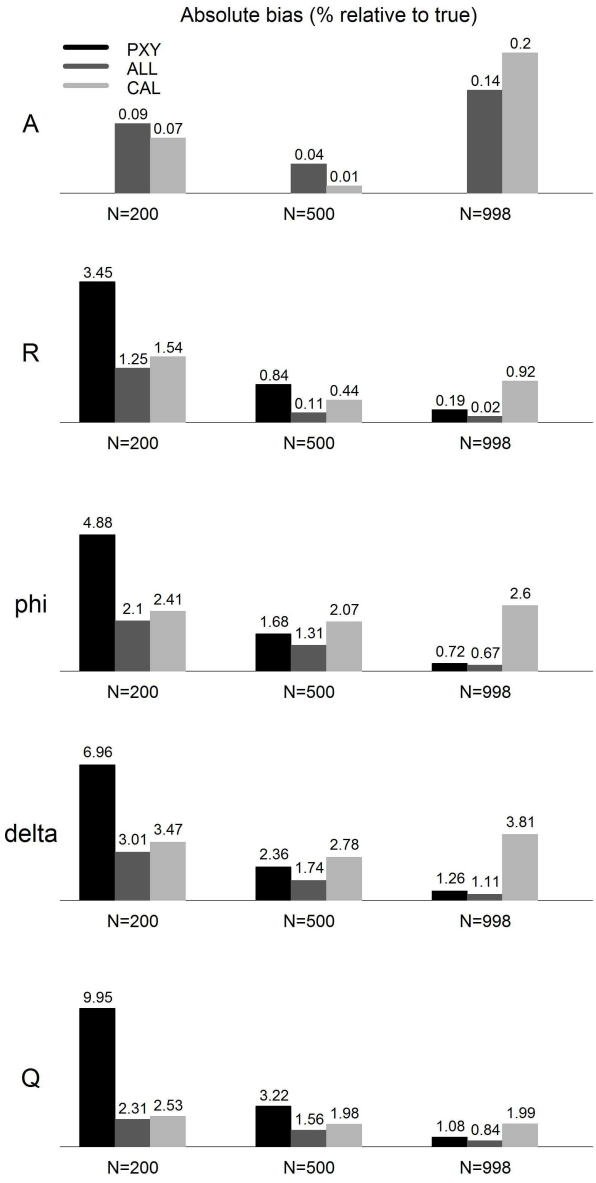

Figure 3.4: Absolute bias of estimators obtained using the three different likelihood functions for three sample sizes. Data is simulated under the Gaussian state-space model in (3.1.2). Absolute bias is expressed in terms of percentage relative to the true parameter value.

Inter-comparison between the three approaches reveals that the $\mathbb{ALL}$ approach has the smallest bias regardless of sample size. When sample size is 998, the bias of the estimates from the $\mathbb{PXY}$ approach becomes comparable with that obtained from the $\mathbb{ALL}$ approach. Even though the $\mathbb{CAL}$ approach only uses 200 data points no matter what the value of N is (100 from $\mathbf{T}_t$ and 100 from $\mathbf{P}_t$), it still produces comparable or better results than the $\mathbb{PXY}$ approach that uses 200 or 500 data points from the $\{\mathbf{P}_t\}$ process.

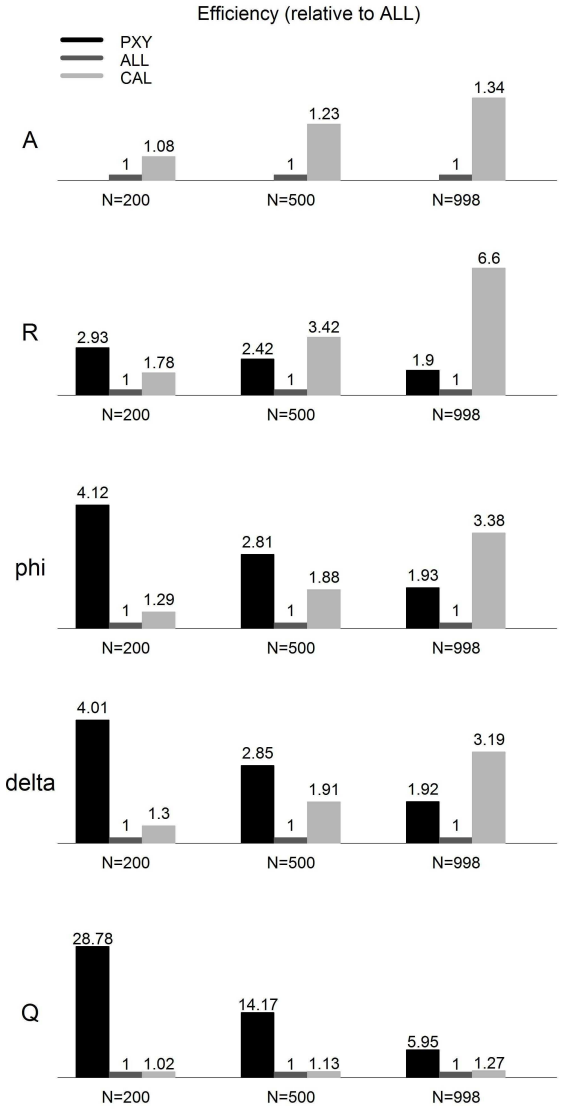

Figure 3.5: Estimated efficiency of estimators obtained using the three different likelihood functions for three sample sizes. Data is simulated under the Gaussian state-space model in (3.1.2). Efficiency is expressed relative to the estimates sample variance obtained from the $\mathbb{ALL}$ likelihood.

The estimated efficiency of the estimators is displayed in Figure 3.5. Since the estimator variance from the $\mathbb{ALL}$ approach is used as the baseline, the efficiency for the estimators from $\mathbb{ALL}$ is 1. Among the three approaches that are being considered, the $\mathbb{ALL}$ approach clearly produces the most efficient estimates regardless of sample size. The estimate of $\mathcal{Q}$ from the $\mathbb{PXY}$ approach is extremely inefficient, compared to that of the $\mathbb{ALL}$ approach. In particular, $\hat{\mathcal{Q}}$ from the $\mathbb{ALL}$ approach is 29 times more efficient than that obtained using the $\mathbb{PXY}$ approach when the sample size is 200. This number is reduced to 14 when $N = 500$ and 6 when $N = 998$. In general, the $\mathbb{CAL}$ approach produces more efficient estimators than the $\mathbb{PXY}$ approach when N=200 and 500. Table 3.1 gives the sample standard error of $\hat{\mathcal{R}}$, $\hat{\phi}$ and $\hat{\mathcal{Q}}$ from the simulations for the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches. It is clear that the standard errors decrease as the sample size increases. The behavior for $\hat{\delta}$ is similar and thus not shown. The findings for the $\mathbb{PXY}$ approach agree with the asymptotic properties of the estimators given in Theorem 3.2.1.

	$\hat{\mathcal{R}}$		$\hat{\phi}$		$\hat{\mathcal{Q}}$	
Sample size (N)	200	998	200	998	200	998
$\mathbb{PXY}$	0.0184	0.0077	0.1187	0.0519	0.0149	0.0060
$\mathbb{ALL}$	0.0108	0.0056	0.0585	0.0374	0.0028	0.0024

Table 3.1: Standard error of the estimators obtained from simulations with 1000 runs with sample size 200 or 998. Results from the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches are shown.

The MSE from the three approaches relative to that obtained from the $\mathbb{ALL}$ approach is displayed in Figure 3.6. Even though the sample size for the $\mathbb{CAL}$ approach is fixed at $m = 100$ and is smaller than that for the $\mathbb{ALL}$ approach, the MSEs from these two approaches are fairly comparable regardless of N . On the other hand, the $\mathbb{PXY}$ approach tends to have larger MSE than the other two approaches. In particular, when $N = 200$, the MSE from the $\mathbb{PXY}$ approach is 11% larger than that from the $\mathbb{ALL}$ approach. However, when sample size increases to 998, the MSEs from all three approaches are fairly similar.

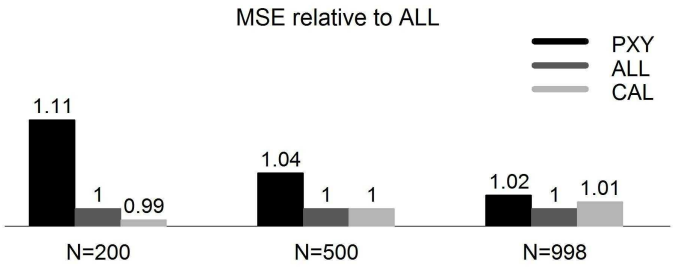

Figure 3.6: Mean square error (MSE) between the simulated state process and the Kalman filter estimated state process over 1000 simulation runs for the different estimation approaches. The Kalman filter estimates are calculated with parameters estimated from the corresponding estimation approach. MSEs are expressed relative to the $\mathbb{ALL}$ approach.

	$\hat{\mathcal{R}}$		$\hat{\phi}$		$\hat{\mathcal{Q}}$	
Sample size (N)	998	∞	998	∞	998	∞
$\mathbb{PXY}$	0.0077	0.0085	0.0519	0.0546	0.0060	0.0068
$\mathbb{ALL}$	0.0056	0.0058	0.0374	0.0376	0.0024	0.0026
$\mathbb{CAL}$	0.0144	0.0141	0.0687	0.0655	0.0028	0.0028

Table 3.2: Comparison between the sample standard error from simulations with $N = 998$ and the theoretical asymptotic standard error for the three estimation approaches.

A reasonable hypothesis is that the theoretical asymptotic variance for estimator obtained using the $\mathbb{ALL}$ approach can be expressed as that shown in Theorem 3.2.1 with $\mathcal{L}_{\mathbf{P}}(\Theta)$ replaced by $\mathcal{L}_{\mathbf{P},\mathbf{T}}(\Theta)$. Table 3.2 shows the sample standard error for the estimates from the simulations with $N = 998$ and the theoretical asymptotic standard error for estimators obtained from the three approaches. Let $\mathfrak{H}_{\mathbf{P}}$ and $\mathfrak{H}_{\mathbf{P},\mathbf{T}}$ be the hessian matrix of the log likelihood function $\ell_{\mathbf{P}}(\hat{\Theta})$ and $\ell_{\mathbf{P},\mathbf{T}}(\hat{\Theta})$ respectively. Then, for the $\mathbb{PXY}$ approach, an estimate of the estimator asymptotic standard error is given by the square root of the diagonal of $-\mathfrak{H}_{\mathbf{P}}^{-1}/N$. The mean of this value over the 1000 simulation runs with $N = 998$ is listed under $N = \infty$ in the table. For the $\mathbb{ALL}$ approach, the same values obtained from $\mathfrak{H}_{\mathbf{P},\mathbf{T}}$ are shown. For the $\mathbb{CAL}$ approach, the values in the $N = \infty$ column are obtained using Theorem 3.3.1. For the $\mathbb{PXY}$ and $\mathbb{CAL}$ approaches, the sample standard errors from the simulation agree reasonably well with the theoretical asymptotic standard errors, while for the $\mathbb{ALL}$ approach, the value listed under

the $N = 998$ column is almost the same as that listed in the $N = \infty$ column. This provides evidence that the theoretical asymptotic variance for the estimator obtained using the $\mathbb{ALL}$ approach is very close to the inverse of the $\mathfrak{I}(\Theta_0)$ matrix given in Theorem 3.2.1 with $\mathcal{L}_{\mathbf{P}}(\Theta)$ replaced by $\mathcal{L}_{\mathbf{P},\mathbf{T}}(\Theta)$. Such evidence is of great practical importance because it means that one can now easily estimate the variance of the estimator in real-world applications using the hessian matrix. The results for $\hat{\mathcal{A}}$ and $\hat{\delta}$ are similar and thus not shown.

Figure 3.7 displays normal probability plots for the estimates obtained from the $\mathbb{PXY}$ approach with varying sample sizes. If the distribution of the estimates follows a normal distribution, one would expect the points in a normal probability plot to follow a straight line. The p-values from the Shapiro-Wilk test are also shown in the figure. The p-values are for the null hypothesis of normality, i.e., a small p-value indicates evidence against the null hypothesis that the distribution of the estimates is normal. For the $\mathbb{PXY}$ approach, departures from normality are prevalent across the different estimators for all sample sizes. Nevertheless, such departure tends to become smaller as the sample size increases. Regardless of the sample size, the Shapiro-Wilk test suggests that there is strong evidence against the null hypothesis of normality for almost all the estimates obtained from this approach. From Theorem 3.2.1, the distribution of the estimators from the $\mathbb{PXY}$ approach should converge to a normal distribution when the sample size becomes large. However, simulation results reveal that the convergence is quite slow, which may be due to the fact that each observed data point from the observation process is dependent of each other when the state process is unknown (Figure 3.2). The convergence rate for a dependent sequence will depend on the problem under investigation and is in general no better than $N^{-1/4}$ and can be as slow as $N^{-1/4} \log N$ (See Hall and Heyde (1980) for details). These convergence rates are much slower than the rate of convergence of order $N^{-1/2}$ for an independent sequence. To reduce the error encountered when approximating the finite sample distribution by the limiting distribution to an order of 0.1, an independent sequence would require 100 observations, whereas, depending on the application, a dependent sequence may require 10000 to 2×10^9 observations to achieve the same level of performance.

Figures 3.8 and 3.9 display normal probability plots for the parameter estimates obtained from the $\mathbb{ALL}$ and $\mathbb{CAL}$ approaches. For comparison, the $\mathbb{CAL}$ approach is also run with $m = 500$ and 998. The results are displayed in Figure 3.9. Departures from normality for the estimates from both approaches are smaller than the $\mathbb{PXY}$ approach. The probability plots indicate mild evidence against normality in the tail values. However, based on the Shapiro-Wilk test, there is strong evidence against the null hypothesis of normality for most of the estimates from the $\mathbb{ALL}$ approach. This is not surprising given that when sample size is large, the p-values from the Shapiro-Wilk test can be small even if the null hypothesis is only slightly violated.

It should be noted that observations from the observation process are mutually dependent during the pre-calibration period, but are independent over the calibration period when conditioned on the known state process (Figure 3.2). If indeed the distribution of the estimates from the $\mathbb{ALL}$ approach converges to a normal distribution, the rate of convergence for the $\mathbb{ALL}$ approach should be faster than the $\mathbb{PXY}$ approach. This is because the calibration period state process is used in the $\mathbb{ALL}$ approach. Given that the departure from normality is not severe, it should still be reasonable to use the normality assumption to make inferences for estimators from the $\mathbb{ALL}$ approach in applications when sample size is large. On the other hand, there is little evidence against the null hypothesis of normality for the estimates from the $\mathbb{CAL}$ approach when m=500 and 998. The stronger evidence of normality in the $\mathbb{CAL}$ approach is due to the use of only the conditionally independent calibration period data. However, it should be noted that a value of $m > 150$ is currently not feasible in real-world reconstruction problem because systematic temperature records were not widely kept until the latter half of the 19^{th} century.

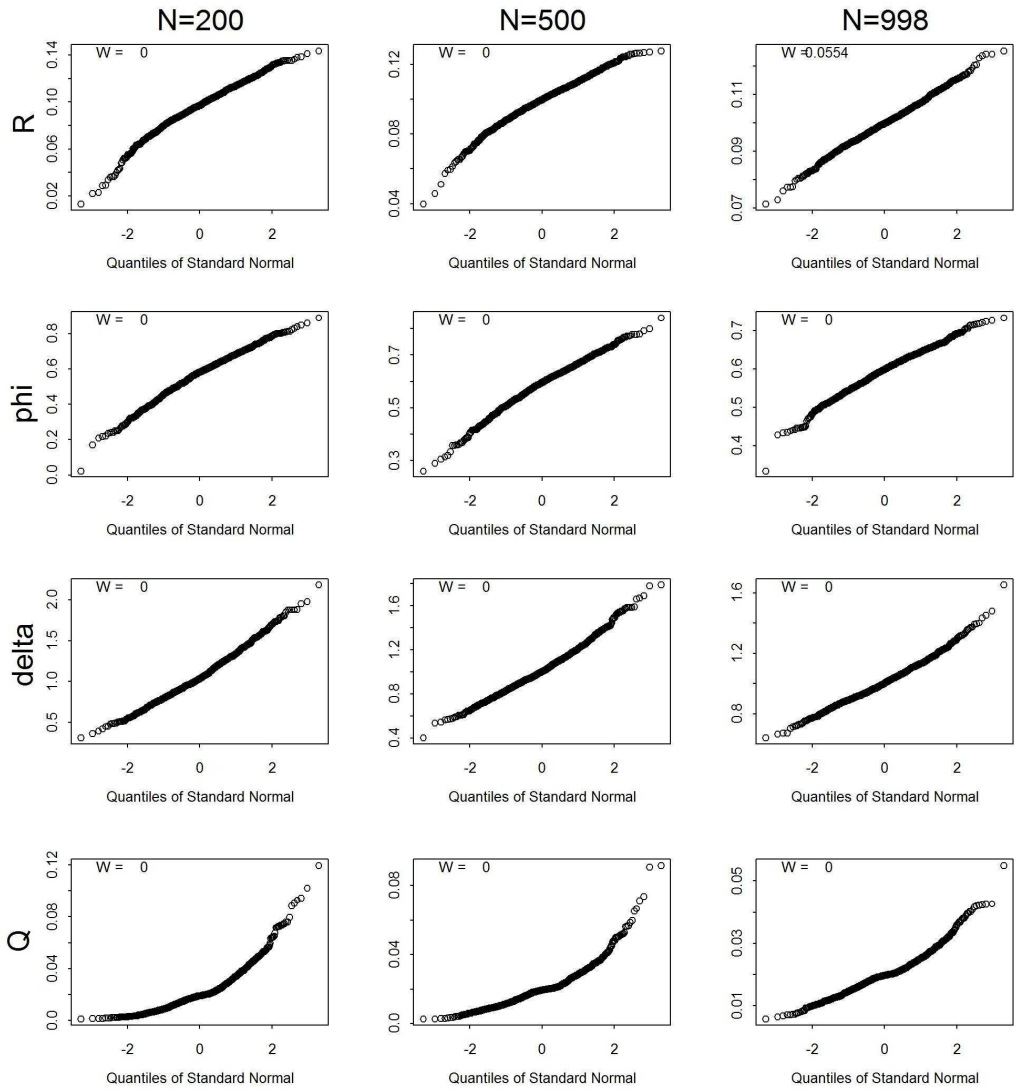

Figure 3.7: Normal probability plots for estimates obtained from the $\mathbb{PXY}$ approach with varying sample sizes. The p-values from the Shapiro-Wilk test are shown in the top left corner of each plot. Small p-value indicates evidence against the null hypothesis of normality.

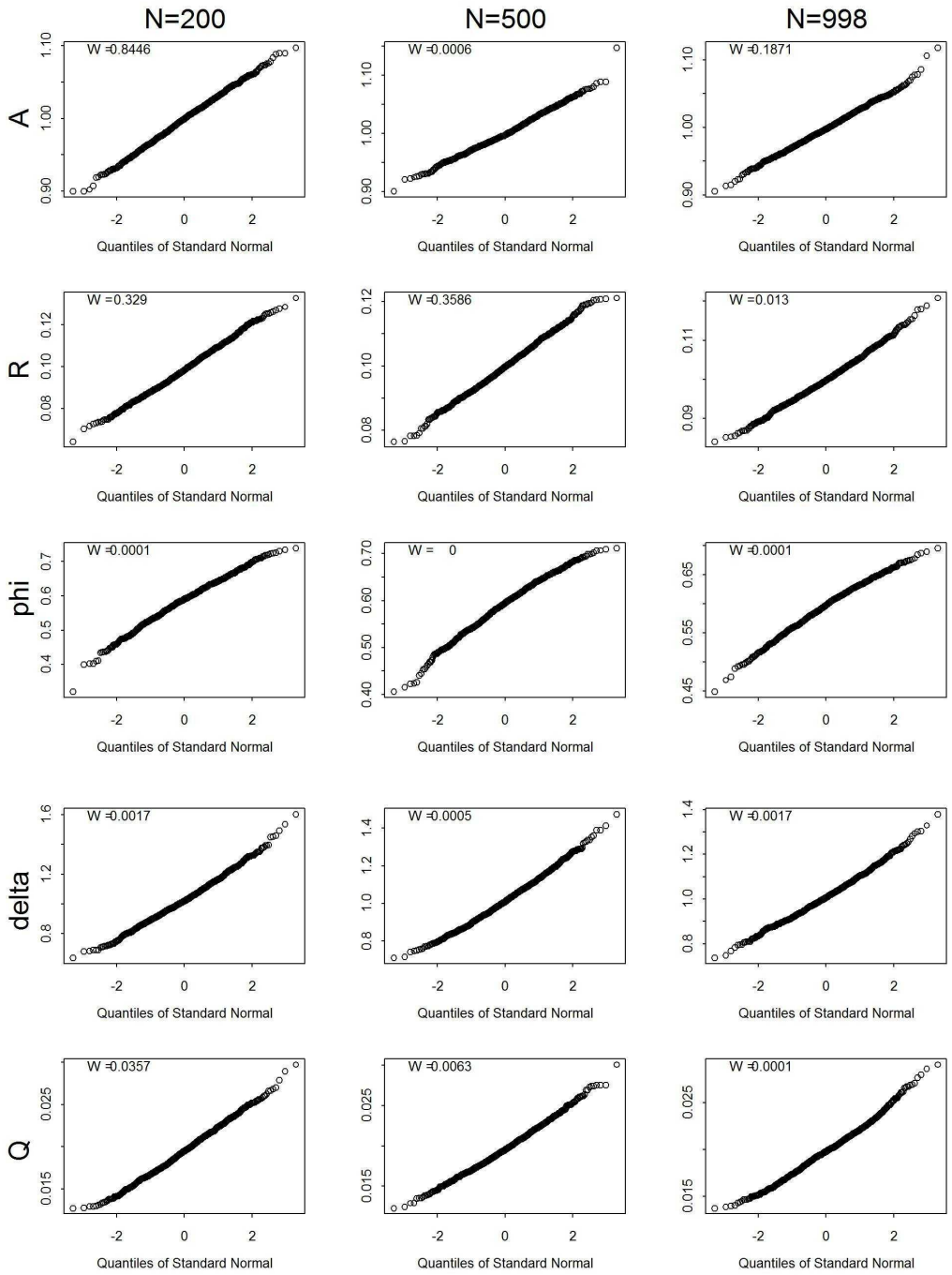

Figure 3.8: Normal probability plots for estimates obtained from the $\mathbb{ALL}$ approach with varying sample sizes. The p-values from the Shapiro-Wilk test are shown in the top left corner of each plot. Small p-value indicates evidence against the null hypothesis of normality.

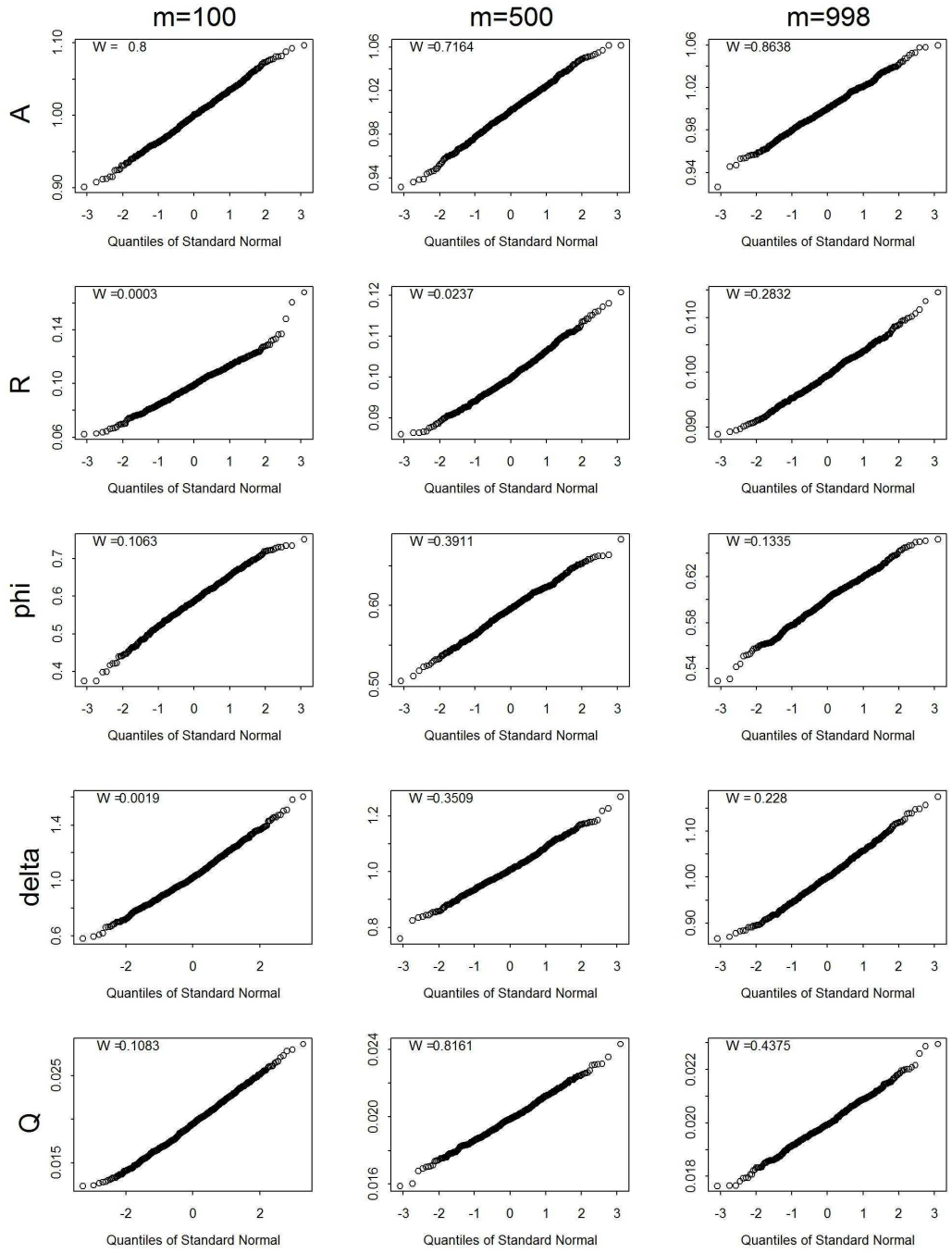

Figure 3.9: Normal probability plots for estimates obtained from the $\mathbb{CAL}$ approach with varying sample sizes. The p-values from the Shapiro-Wilk test are shown in the top left corner of each plot. Small p-value indicates evidence against the null hypothesis of normality.

3.4.2 Robustness of estimators with respect to observational noise structure

In the field of paleoclimatology, it has been argued that noise contained in a proxy series can behave like a red noise process, rather than a white noise process. To investigate the impact of red observational noise on the different estimators, another set of simulations is conducted. In this simulation, data is generated with red observational noise as stated below and the three estimation approaches are then applied to estimate the parameters.

Based on real-world proxy data, Mann et al. (2007) have suggested that a possible structure for the proxy noise is an AR(1) process with lag one autocorrelation around 0.3, i.e. a red noise series. Thus, the noise series e_t in the observation process can be written as,

$$e_t = \psi e_{t-1} + \kappa_t, \quad \kappa_t \sim \mathrm{N}(0, \mathcal{G}),$$

where, in the simulations, $\psi = 0.3$ and $\mathcal{G} = 0.1$. The value of $\mathcal{R}$ becomes

$$\begin{aligned} \mathcal{R} &= \mathrm{Var}(e_t) \\ &= \mathrm{Var}\left(\sum_{j=0}^{\infty} \psi^j \kappa_{t-j}\right) \\ &= \mathcal{G}/(1-\psi^2), \end{aligned}$$

which corresponds to $\mathcal{R} = 0.1/(1-0.3^2) \approx 0.11$ in the simulation. One thousand simulation runs are conducted. Same as before, sample sizes of 200, 500 and 998 are used. For the $\mathbb{CAL}$ approach, $m = 100$ regardless of N.

Figure 3.10 displays the estimated bias of the various estimators under these circumstances. Comparing the results with Figure 3.4, the red observational noise has a detrimental impact on the $\mathbb{PXY}$ approach and the previously well performing $\mathbb{ALL}$ approach. For example, the bias for $\hat{\mathcal{R}}$ has increased from 0.19% to 65% for the $\mathbb{PXY}$ approach and from 0.02% to 23% for the $\mathbb{ALL}$ approach when N=998. It is also worth noting that the estimator bias for the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches increases as sample size increases. The increase in bias for the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches is a result of the use of the Kalman filter estimate in the calculation of the parameter estimate. When deriving the Kalman filter, one needs to calculate the moments of

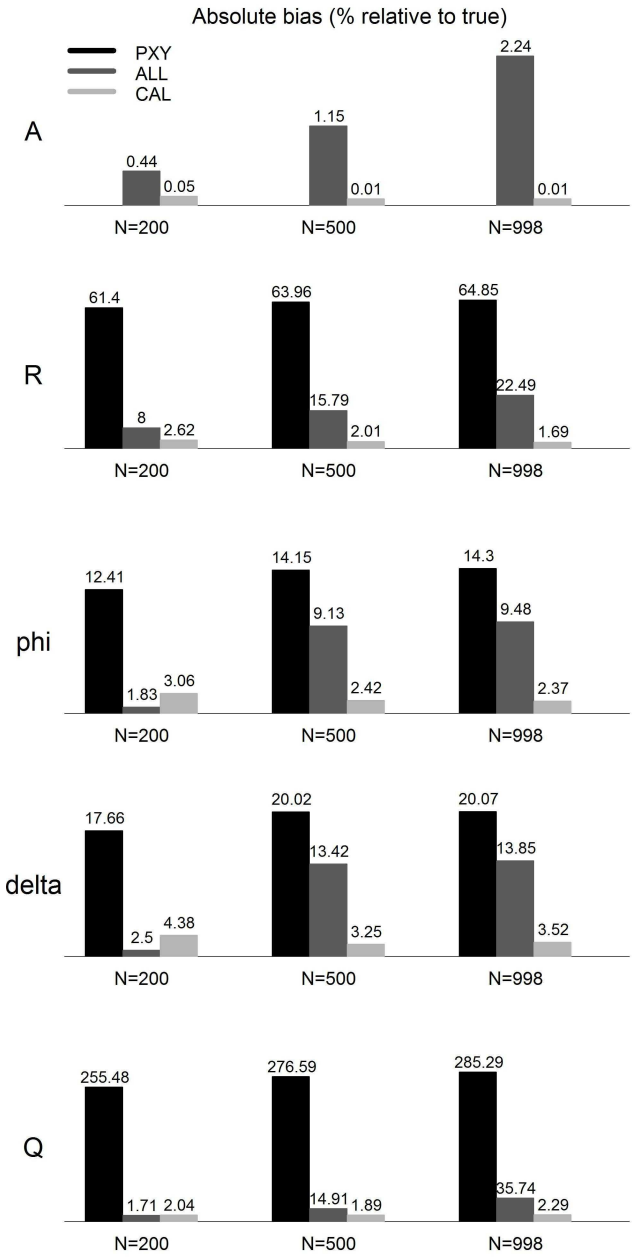

Figure 3.10: Absolute bias of the estimators obtained using the three different likelihood functions for three sample sizes. The simulated data is generated with red observational noise.

e_t conditional on $\mathbf{P}_{1:t-1}$ (see section 3.1.1). In the case where e_t is a white noise process, which is the assumption used when deriving the Kalman filter, the conditional moments are the same as the unconditional moments. However, if e_t is a red noise process, then e_t will depend on e_s for $s < t$. Since $\mathbf{P}_s = \mathcal{A}\mathbf{T}_s + e_s$, the conditional moments are no longer equivalent to the unconditional ones. Thus, the Kalman filter estimates used in the parameter estimation process are therefore incorrectly derived.

In contrast, the biasedness of the estimators for the $\mathbb{CAL}$ approach appears to be unaffected by the red observational noise. It is in fact possible to show that the estimators from the $\mathbb{CAL}$ approach are still asymptotically unbiased even with red observational noise. From the proof of Theorem 3.3.1, it is apparent that the error term e_t is not involved in deriving the asymptotic distribution of $\hat{\phi}$, $\hat{\delta}$ and $\hat{\mathcal{Q}}$. Thus, the structure of e_t has no effect on the asymptotic properties for these estimators. From the proof of the asymptotic distribution of $\hat{\mathcal{A}}$ in section 3.3.3, one has

$$\hat{\mathcal{A}} - \mathcal{A} = \frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sum_{t=n+1}^{N} \mathbf{T}_t^2}.$$

Hence

$$\mathrm{E}(\hat{\mathcal{A}}) = \mathcal{A} + \mathrm{E}\left(\frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sum_{t=n+1}^{N} \mathbf{T}_t^2}\right).$$

Note that

$$\mathrm{E}\left(\frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sum_{t=n+1}^{N} \mathbf{T}_t^2}\right) = \mathrm{E}\left[\mathrm{E}\left(\frac{\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{\sum_{t=n+1}^{N} \mathbf{T}_t^2}\middle| \mathbf{T}_{n+1}, \cdots, \mathbf{T}_N\right)\right] = 0,$$

since

$$\mathrm{E}(e_t) = \mathrm{E}\left(\sum_{j=0}^{\infty} \psi^j \kappa_{t-j}\right) = 0.$$

Thus

$$\mathrm{E}(\hat{\mathcal{A}}) = \mathcal{A}$$

which means $\hat{\mathcal{A}}$ is unbiased.

To prove that $\hat{R}$ is asymptotically unbiased, first note that $\hat{\mathcal{A}} \xrightarrow{p} \mathcal{A}$ even when the error

term e_t follows an AR(1) structure. To show this, note that

$$\hat{\mathcal{A}} - \mathcal{A} = \frac{m^{-1}\sum_{t=n+1}^{N} \mathbf{T}_t e_t}{m^{-1}\sum_{t=n+1}^{N} \mathbf{T}_t^2}.$$

Also note that $m^{-1}\sum_{t=n+1}^{N} \mathbf{T}_t^2$ does not depend on e_t and this term converges to $m^{-1}\sum_{t=n+1}^{N}(\mu_t^2 + \gamma_t)$ as $m \longrightarrow \infty$ (section 3.3.3). Hence, to prove $\hat{\mathcal{A}} \xrightarrow{p} \mathcal{A}$, it is suffice to show that $m^{-1}\sum_{t=n+1}^{N} \mathbf{T}_t e_t \xrightarrow{p} 0$. The error term under the AR(1) structure is $e_t = \psi e_{t-1} + \kappa_t$, where $\mathrm{Var}(\kappa_t) = \mathcal{G}$. Now consider

$$\begin{aligned}
& m^{-1} \sum_{t=n+1}^{N} \mathbf{T}_t e_t \\
=& m^{-1} \sum_{t=n+1}^{N} (\mu_t e_t + \widetilde{\mathbf{T}}_t e_t) \\
=& m^{-1} \sum_{t=n+1}^{N} \left[\mu_t \sum_{j=0}^{\infty} \psi^j \kappa_{t-j} + \left(\sum_{k=0}^{t-(n+2)} \phi^k w_{t-k} \right) \left(\sum_{i=0}^{\infty} \psi^i \kappa_{t-i} \right) \right] \\
=& \sum_{j=0}^{\infty} \left(\psi^j m^{-1} \sum_{t=n+1}^{N} \mu_t \kappa_{t-j} \right) + \sum_{i=0}^{\infty} \left\{ \psi^i m^{-1} \sum_{t=n+1}^{N} \left[\kappa_{t-i} \left(\sum_{k=0}^{t-(n+2)} \phi^k w_{t-k} \right) \right] \right\}.
\end{aligned}$$

The κ_t's are i.i.d. random variables with mean zero and κ_t and w_s are mutually independent and so, by Theorem 3.3.2 in section 3.3.3, as $m \longrightarrow \infty$,

$$\begin{aligned}
& m^{-1} \sum_{t=n+1}^{N} \mu_t \kappa_{t-j} \xrightarrow{p} 0, \\
& m^{-1} \sum_{t=n+1}^{N} \left[\kappa_{t-i} \left(\sum_{k=0}^{t-(n+2)} \phi^k w_{t-k} \right) \right] \xrightarrow{p} 0.
\end{aligned}$$

The second convergence result follows because

$$\begin{aligned}
&\mathrm{Cov}\left[\kappa_{t-i}\left(\sum_{k=0}^{t-(n+2)}\phi^k w_{t-k}\right),\ \kappa_{s-i}\left(\sum_{k=0}^{s-(n+2)}\phi^k w_{s-k}\right)\right] \\
=&\mathrm{E}(\kappa_{t-i})\mathrm{E}(\kappa_{s-i})\mathrm{E}\left[\left(\sum_{k=0}^{t-(n+2)}\phi^k w_{t-k}\right)\left(\sum_{k=0}^{s-(n+2)}\phi^k w_{s-k}\right)\right] \\
=&0,
\end{aligned}$$

$$\mathrm{E}\left[\kappa_{t-i}\left(\sum_{k=0}^{t-(n+2)}\phi^k w_{t-k}\right)\right]=0,$$

and

$$\mathrm{Var}\left[\kappa_{t-i}\left(\sum_{k=0}^{t-(n+2)}\phi^k w_{t-k}\right)\right]\leq\frac{\mathcal{GQ}}{1-\phi^2}.$$

Now, provided that $|\phi|$ and $|\psi|$ are less than 1,

$$m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_t e_t\xrightarrow{p}0.$$

This implies that $\hat{\mathcal{A}}-\mathcal{A}\xrightarrow{p}0$. Now from the proof of the asymptotic distribution of $\hat{R}$ in section 3.3.3, one has

$$\hat{\mathcal{R}}=m^{-1}\sum_{t=n+1}^{N}\left[(\mathcal{A}^2+\hat{\mathcal{A}}^2-2\mathcal{A}\hat{\mathcal{A}})\mathbf{T}_t^2+2(\mathcal{A}-\hat{\mathcal{A}})\mathbf{T}_t e_t+e_t^2\right].$$

Note that by Lemma 3 in section 3.3.3,

$$m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_t^2\xrightarrow{p}m^{-1}\sum_{t=n+1}^{N}(\mu_t^2+\gamma_t)<\infty$$

and by earlier arguments above,

$$m^{-1}\sum_{t=n+1}^{N}\mathbf{T}_t e_t\xrightarrow{p}0.$$

By Theorem 3.6.7 in section 3.6.3, $m^{-1}\sum_{t=n+1}^{N} e_t^2$ is asymptotically normal with mean $\mathrm{E}(e_t^2) = \mathcal{R}$. Therefore, since $\hat{\mathcal{A}} \xrightarrow{p} \mathcal{A}$, $\hat{\mathcal{R}}$ is asymptotically normal with mean $\mathcal{R}$ and thus $\hat{\mathcal{R}}$ is asymptotically unbiased.

Table 3.3 shows the change in estimator sample variance in the red noise simulation when compared to the white noise simulation with N=998. The sample variance of $\hat{\mathcal{A}}$ and $\hat{\mathcal{R}}$ from the $\mathbb{CAL}$ approach in the red noise simulation are respectively 89% and 34% larger compared to that obtained from the white noise simulations. In contrast, there is almost no change in the sample variance for $\hat{\phi}$, $\hat{\delta}$ and $\hat{\mathcal{Q}}$. This finding agrees with the fact that the asymptotic distribution for these estimators from the $\mathbb{CAL}$ approach is unaffected by the structure of the noise term e_t. For the estimators from the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches, there are substantial changes in the sample variances. However, both positive and negative changes can be found.

	$\hat{\mathcal{A}}$	$\hat{\mathcal{R}}$	$\hat{\phi}$	$\hat{\delta}$	$\hat{\mathcal{Q}}$
$\mathbb{PXY}$		104.1	-19.9	-11.4	566.1
$\mathbb{ALL}$	78.2	32.4	-31.3	-27.8	109.9
$\mathbb{CAL}$	89.4	33.7	-0.5	-0.9	-4.4

Table 3.3: Percentage change in estimates sample variance in the red noise simulation relative to the white noise simulation with $N = 998$.

The probability plots for the estimates from the three approaches are similar to that in the white noise simulation and thus not shown. This suggests that red observational noise has affected the location and variability of the estimators, but might not have affected their normality. Figure 3.11a shows the MSE from the three approaches relative to that obtained from the $\mathbb{ALL}$ approach for the red noise simulation. The $\mathbb{PXY}$ approach performs the worst among the three approaches regardless of the sample size. It also appears that the $\mathbb{CAL}$ approach was able to produce a slightly better estimate of the state process than the $\mathbb{ALL}$ approach. Figure 3.11b shows the MSE from the three approaches relative to the corresponding MSE from the white noise simulation. It is clear that the red observational noise has an detrimental impact on the $\mathbb{PXY}$ approach in term of the Kalman filter estimates of the state

process. The impact on the $\mathbb{ALL}$ and $\mathbb{CAL}$ approaches is much smaller and is very similar when sample size is small. However, when sample size gets large, the $\mathbb{CAL}$ approach tends to be more robust. Even though the parameter estimates from the $\mathbb{CAL}$ approach are asymptotically unbiased, the MSEs do increase by about 10% in the red noise simulation when compared to the white noise simulation. Such an increase is likely due to the increase in variability of the parameter estimates which is exhibited in Table 3.3.

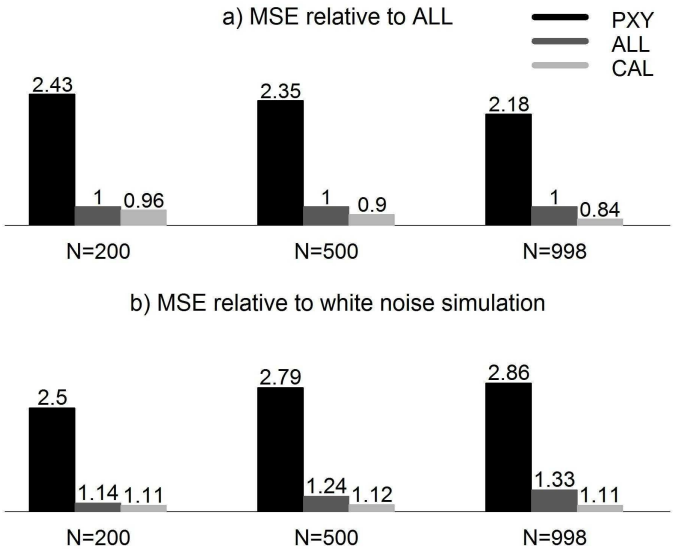

Figure 3.11: Mean square error (MSE) between the simulated state process and the Kalman filter estimated state process over 1000 simulation runs with red observational noise. The top plot shows the MSE from the different approaches relative to the $\mathbb{ALL}$ approach. The bottom plot displays the MSE obtained with red observational noise relative to the corresponding MSE obtained from the white noise simulation.

3.4.3 Robustness of estimators with respect to state equation specification

In this section, a simulation study is conducted to investigate the robustness of estimators with respect to state equation specification. From the PACF of the simulated NH temperature series in Figure 3.3, one might argue that the structure of the NH temperature might be that of an AR(2) process, instead of an AR(1) process. Simulations were therefore run to investigate the robustness of the estimators when the estimation approaches misspecify an AR(2) state process as an AR(1) process.

The simulated state process can be written as,

$$\mathbf{T}_t = \varphi_1 \mathbf{T}_{t-1} + \varphi_2 \mathbf{T}_{t-2} + \delta \mathbf{F}_t + w_t$$

where, in the simulation, $\varphi_1 = 0.51$, $\varphi_2 = 0.15$, $\delta = 1$ and $\text{Var}(w_t) = \mathcal{Q} = 0.02$. The value of φ is chosen according to the PACF in Figure 3.3, where the PACF equals to $\varphi_1/(1-\varphi_2)$ at lag 1 and equals to φ_2 at lag 2. Again, the simulation is run 1000 times and the three approaches are used to estimate the parameters. Again, the sample size for the $\mathbb{CAL}$ approach is $m = 100$ regardless of N.

Figure 3.12 displays the absolute bias for the estimators from the three approaches. The bias for $\hat{\phi}$ is not shown because the parameter ϕ is not involved in the AR(2) model. Among the three approaches, the $\mathbb{ALL}$ approach appears to produce better estimates regardless of the sample size. When $N = 998$, the estimator bias from the $\mathbb{ALL}$ approach is generally small, ranging from 0.5% to 2.7%. Similar to simulations conducted in the previous sections, the $\mathbb{PXY}$ approach produces the worst estimators, with the exception that the $\mathbb{CAL}$ approach has the largest bias for $\hat{\delta}$.

For the $\mathbb{PXY}$ approach, the bias for the parameter estimates is generally large when $N = 200$. For this approach, increasing the sample size increases the absolute bias of the estimates. This is not surprising given that the structure of the simulated data is different from what the estimation approach assumes. Thus, an estimation approach that produce biased estimates when sample size is small should not be able to benefit from an increase in sample size. However, the bias for $\hat{\delta}$ from the $\mathbb{ALL}$ approach decreases as sample size increase. Such an

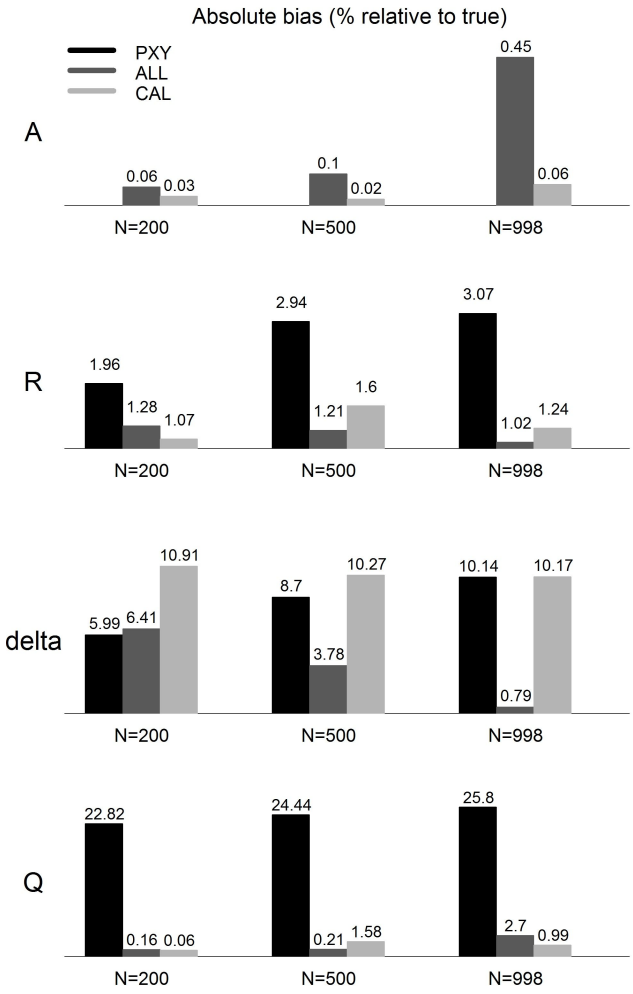

Figure 3.12: Absolute bias of the estimators obtained using the three different likelihood functions for three sample sizes. The simulated data is generated under an AR(2) model.

abnormal phenomenon can be explained by first noting that the bias for $\hat{\delta}$ is in fact negative for the $\mathbb{PXY}$ approach and gets more negative as N increases. On the other hand, the bias for $\hat{\delta}$ is positive for both the $\mathbb{ALL}$ and $\mathbb{CAL}$ approaches and is roughly constant over the different N for the $\mathbb{CAL}$ approach (since m=100 regardless of N). Summing up the biases for $\hat{\delta}$ from the $\mathbb{PXY}$ and $\mathbb{CAL}$ approaches at each N corresponds roughly to the bias for $\hat{\delta}$ from the $\mathbb{ALL}$ approach. This is reasonable given that the likelihood function used in the $\mathbb{ALL}$ approach is a combination of the likelihood functions used in the $\mathbb{PXY}$ and $\mathbb{CAL}$ approaches. So, it is intuitive that the bias of the parameter estimates from the $\mathbb{ALL}$ approach will be related to that

of the $\mathbb{PXY}$ and $\mathbb{CAL}$ approaches. The exact relationship between the biases from the three approaches will depend on the parameter of interest. It is hard to determine such relationship theoretically given that there are no explicit forms for the parameter estimators from the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches. Thus, one can only examine such relationships empirically.

The normal probability plots of the estimates are very similar to that in Figure 3.7, 3.8 and 3.9 and thus not shown. In terms of the estimator sample variances, depending on the parameter and estimation approach, positive or negative changes can be observed when compared to that obtained in the AR(1) simulations.

Figure 3.13 shows the MSE of the Kalman filter estimates obtained from the three estimation approaches. The MSEs between the different approaches are very similar. Compared to the MSE from the AR(1) simulations, there is almost no change in the MSE value for all approaches. This provides evidence that misspecifying the AR(2) state process under investigation as an AR(1) process has almost no impact on the accuracy of the Kalman filter estimates.

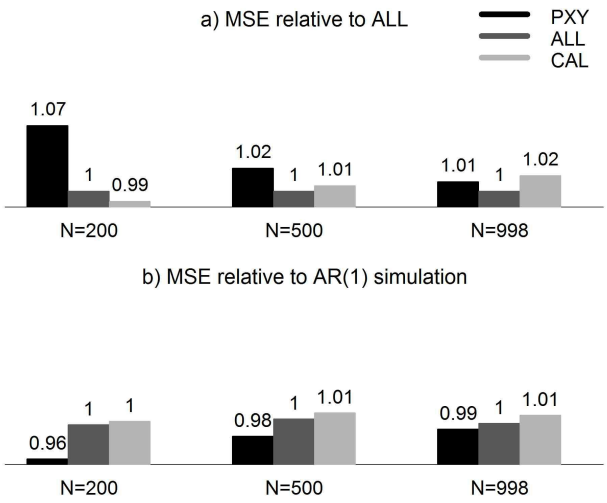

Figure 3.13: Mean square error between the simulated state process and the Kalman filter estimated state process over 1000 simulation runs with an AR(2) model. The top plot shows the MSE from the different approaches relative to the $\mathbb{ALL}$ approach. The bottom plot displays the MSE obtained under the AR(2) model relative to the corresponding MSE obtained from the AR(1) simulation.

3.5 Concluding remarks

In this chapter, the state-space model approach has been proposed as a method to reconstruct unknown historical temperature. Such an approach involves defining an appropriate state-space model for the reconstruction problem and then using the Kalman filter and smoother to provide historical temperature estimates. The use of the Kalman filter and smoother is complicated by the fact that parameters that specify the state-space model are unknown and estimation is required. Existing estimation methods do not take into account the observed temperature record during the calibration period and hence two new estimation approaches are proposed.

Both asymptotic and finite sample properties of the estimators obtained from the three competing estimation approaches have been examined. Under the univariate Gaussian state-space model assumption, the asymptotic distribution for estimator obtained from the $\mathbb{CAL}$ approach has been derived. On the other hand, the theoretical asymptotic distribution of the estimator obtained from the $\mathbb{ALL}$ approach is currently unknown due to the complexity of its likelihood function. However, simulation studies provide evidence that the estimator is likely to possess the same asymptotic property as that stated in Theorem 3.2.1 with $\mathcal{L}_{\mathbf{P}}(\Theta)$ replaced by $\mathcal{L}_{\mathbf{P},\mathbf{T}}(\Theta)$.

Assuming the data follows the same structure as that given by the specified state-space model, simulation results suggest that the parameter estimators from the $\mathbb{ALL}$ approach are more efficient than the other approaches. In terms of the Kalman filter estimates obtained using the estimators, the two new approaches produce comparable results even though the sample size for the $\mathbb{CAL}$ approach is fixed at $m = 100$ and is smaller than that for the $\mathbb{ALL}$ approach. Inter-comparison between the three approaches suggests that the two new approaches perform better than the existing approach in terms of the Kalman filter estimate.

The impact of departures in the assumed state-space model on the estimators was also considered through simulation studies. The impact of red observational noise, rather than white was considered and in this case, it was proved that the parameter estimators from the $\mathbb{CAL}$ approach remain asymptotically unbiased. This provides evidence that the $\mathbb{CAL}$ approach estimators are robust to such departures in model assumptions. In contrast, simulation re-

sults suggest that the $\mathbb{PXY}$ and $\mathbb{ALL}$ approaches tend to produce biased parameter estimators when observational noise is red. At the same time, the Kalman filter estimate from the $\mathbb{CAL}$ approach is much better than that obtained with the $\mathbb{PXY}$ approach and is slightly better than that from the $\mathbb{ALL}$ approach.

The consequence of misspecifying the state equation has also been examined. In particular, a simulation study was carried out where the simulated state process is AR(2) rather than AR(1). Evidence from the study indicates that the estimators from all three approaches may be biased if one misspecified the state equation as an AR(1) process when estimating the parameters. With respect to the Kalman filter estimate, all approaches are found to be robust to the mild state equation misspecification that was considered in the simulation experiments.

In summary, the use of the $\mathbb{PXY}$ approach should be avoided when the state process is partially observed. This approach consistently produces estimates that are worse than one or both of the new approaches. On the other hand, using the $\mathbb{ALL}$ approach to estimate the state-space model parameters is possibly the best choice for any sample size if one has strong belief that the assumed state-space model is correct. However, this is subject to the caveat that the true parameter values, which are unknown in real-world application, are used as the initial value of the EM algorithm in the simulation studies. In real-world applications, a reasonable choice of the initial value can be obtained using the $\mathbb{CAL}$ approach since the estimates from the $\mathbb{CAL}$ approach are asymptotically unbiased.

If one is only concerned with the Kalman filter estimate for the state process, either the $\mathbb{ALL}$ approach or the $\mathbb{CAL}$ approach can be used. Compared to the $\mathbb{ALL}$ estimation approach, the $\mathbb{CAL}$ approach has the advantages that the maximum likelihood estimates can be solved analytically and the asymptotic properties of the estimators are known. Departure from model assumption generally affects the biasedness of the estimators. Exception are the parameter estimators from the $\mathbb{CAL}$ approach in which they remain asymptotically unbiased when the observational noise is red rather than white. However, no approach seems to produce estimators that are robust to both of the departures from the assumed model that have been considered. Nevertheless, in both cases considered, the Kalman filter estimates obtained with parameters estimated from the $\mathbb{ALL}$ and $\mathbb{CAL}$ approaches are affected only slightly.

3.6 Appendix

3.6.1 Derivation of the EM estimation procedure with an unknown state process

The derivation for the updated parameter estimate $\Theta^{(j)}$ of the EM algorithm stated in (3.2.5) will be presented next. Shumway and Stoffer (1982, 2000) provided derivations for the case where the exogenous variables are excluded and the state process is assume to be completely unknown. Here, the procedure is extended to include exogenous variables.

The joint density function of the complete data $(\mathbf{P}_{1:N}, \mathbf{T}_{0:N})$ is

$$f_\Theta(\mathbf{P}_{1:N}, \mathbf{T}_{0:N}) = f_{\mu_0,\Sigma_0}(\mathbf{T}_0) \prod_{t=1}^{N} f_{\phi,\delta,\mathcal{Q}}(\mathbf{T}_t \mid \mathbf{T}_{t-1}) \prod_{t=1}^{N} f_{\mathcal{A},\mathcal{R}}(\mathbf{P}_t \mid \mathbf{T}_t).$$

Hence, under the normality assumption and ignoring constants, the log likelihood function is

$$\begin{aligned} -2\,\ell(\Theta) =& -2\ln f_\Theta(\mathbf{P}_{1:N}, \mathbf{T}_{0:N}) \\ =& \ln|\Sigma_0| + (\mathbf{T}_0 - \mu_0)^{\mathrm{T}}\Sigma_0^{-1}(\mathbf{T}_0 - \mu_0) \\ &+ N\ln|\mathcal{Q}| + \sum_{t=1}^{N}(\mathbf{T}_t - \phi\mathbf{T}_{t-1} - \delta\mathbf{F}_t)^{\mathrm{T}}\mathcal{Q}^{-1}(\mathbf{T}_t - \phi\mathbf{T}_{t-1} - \delta\mathbf{F}_t) \\ &+ N\ln|\mathcal{R}| + \sum_{t=1}^{N}(\mathbf{P}_t - \mathcal{A}\mathbf{T}_t)^{\mathrm{T}}\mathcal{R}^{-1}(\mathbf{P}_t - \mathcal{A}\mathbf{T}_t). \end{aligned} \tag{3.6.1}$$

For the case where the state process is assumed to be unobserved, to implement the EM algorithm, one needs to calculate

$$H\left(\Theta \mid \Theta^{(j-1)}\right) = \mathrm{E}_{\mathbf{T}}\left[\ln f_\Theta(\mathbf{P}_{1:N}, \mathbf{T}_{0:N}) \middle| \mathbf{P}_{1:N}, \Theta^{(j-1)}\right].$$

Using Theorem 3.6.8 in section 3.6.3, one can easily obtain the above expectation by first noting that (cf. (3.1.10))

$$\mathrm{E}\left(\mathbf{T}_t - \phi\mathbf{T}_{t-1} - \delta\mathbf{F}_t \middle| \mathbf{P}_{1:N}, \Theta^{(j-1)}\right) = \mathbf{T}_{t|N} - \phi\mathbf{T}_{t-1|N} - \delta\mathbf{F}_t$$

and

$$\begin{aligned}
&\mathrm{Var}\left(\mathbf{T}_t - \phi\mathbf{T}_{t-1} - \delta\mathbf{F}_t \middle| \mathbf{P}_{1:N}, \Theta^{(j-1)}\right) \\
&= \mathbf{S}_{t|N} + \phi\mathbf{S}_{t-1|N}\phi^{\mathrm{T}} - \mathrm{E}\left(\mathbf{T}_t\mathbf{T}_{t-1}^{\mathrm{T}} \middle| \mathbf{P}_{1:N}, \Theta^{(j-1)}\right)\phi^{\mathrm{T}} - \phi\mathrm{E}\left(\mathbf{T}_{t-1}\mathbf{T}_t^{\mathrm{T}} \middle| \mathbf{P}_{1:N}, \Theta^{(j-1)}\right) \\
&= \mathbf{S}_{t|N} + \phi\mathbf{S}_{t-1|N}\phi^{\mathrm{T}} - \mathbf{S}_{t,t-1|N}\phi^{\mathrm{T}} - \phi\mathbf{S}_{t,t-1|N}^{\mathrm{T}},
\end{aligned}$$

where the Kalman smoother and lag one covariance smoother are obtained under the current parameter value $\Theta^{(j-1)}$. By Theorem 3.6.8, ignoring a constant, one now has

$$\begin{aligned}
&-2\ H\left(\Theta|\Theta^{(j-1)}\right) \\
&= \ln|\Sigma_0| + N\ln|\mathcal{Q}| + N\ln|\mathcal{R}| + \mathrm{tr}\left\{\Sigma_0^{-1}[\mathbf{S}_{0|N} + (\mathbf{T}_{0|N} - \mu_0)(\mathbf{T}_{0|N} - \mu_0)^{\mathrm{T}}]\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{Q}^{-1}\sum_{t=1}^{N}\left(\mathbf{S}_{t|N} + \phi\mathbf{S}_{t-1|N}\phi^{\mathrm{T}} - \mathbf{S}_{t,t-1|N}\phi^{\mathrm{T}} - \phi\mathbf{S}_{t,t-1|N}^{\mathrm{T}}\right)\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{Q}^{-1}\sum_{t=1}^{N}(\mathbf{T}_{t|N} - \phi\mathbf{T}_{t-1|N} - \delta\mathbf{F}_t)(\mathbf{T}_{t|N} - \phi\mathbf{T}_{t-1|N} - \delta\mathbf{F}_t)^{\mathrm{T}}\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{R}^{-1}\sum_{t=1}^{N}\left[\mathcal{A}\mathbf{S}_{t|N}\mathcal{A}^{\mathrm{T}} + (\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|N})(\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|N})^{\mathrm{T}}\right]\right\}.
\end{aligned}$$

To find the $\phi^{(j)}$, $\delta^{(j)}$ and $\mu_0^{(j)}$ values that maximize $H(\Theta \mid \Theta^{(j-1)})$, one can take the derivatives with respect to ϕ, δ and μ_0, set them equal to zero and solve to obtain their optimal values. At the same time, by Theorem 3.6.9 in section 3.6.3, $H(\Theta|\Theta^{(j-1)})$ is maximized when

$$\mathcal{R} = N^{-1}\sum_{t=1}^{N}\left[\mathcal{A}\mathbf{S}_{t|N}\mathcal{A}^{\mathrm{T}} + (\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|N})(\mathbf{P}_t - \mathcal{A}\mathbf{T}_{t|N})^{\mathrm{T}}\right]$$

and when

$$\begin{aligned}
\mathcal{Q} =& N^{-1}\sum_{t=1}^{N}\left(\mathbf{S}_{t|N} + \phi^{(j)}\mathbf{S}_{t-1|N}\phi^{(j)\mathrm{T}} - \mathbf{S}_{t,t-1|N}\phi^{\mathrm{T}} - \phi\mathbf{S}_{t,t-1|N}^{\mathrm{T}}\right) \\
&+ N^{-1}\sum_{t=1}^{N}\left(\mathbf{T}_{t|N} - \phi^{(j)}\mathbf{T}_{t-1|N} - \delta^{(j)}\mathbf{F}_t\right)\left(\mathbf{T}_{t|N} - \phi^{(j)}\mathbf{T}_{t-1|N} - \delta^{(j)}\mathbf{F}_t\right)^{\mathrm{T}}.
\end{aligned}$$

After some simplifications, one will be able to obtain the expressions stated in (3.2.5). The parameter μ_0 and Σ_0 cannot be estimated simultaneously because in (3.6.1), the term involving

μ_0 and Σ_0 contains only one data point $\mathbf{T}_0$. Hence, one can only estimate one of the parameters and fix the other at some reasonable value. In particular, one can fix the covariance matrix and use the Kalman smoother at $t = 0$ to estimate μ_0.

3.6.2 Derivation of the EM estimation procedure with a partially known state process

The EM estimation procedure derived in the previous section will now be modified to account for a partially known state process where $\mathbf{T}_{n+1:N}$ is assumed to be known. For the EM algorithm, one now has

$$H\left(\Theta|\Theta^{(j-1)}\right) = \mathrm{E}_{\mathbf{T}}\left[\ln f_{\Theta}(\mathbf{P}_{1:N}, \mathbf{T}_{0:N})\middle|\mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}, \Theta^{(j-1)}\right],$$

where $\ln f_{\Theta}(\mathbf{P}_{1:N}, \mathbf{T}_{0:N})$ is defined as in (3.6.1). In order to derive the above expectation, one needs to first derive, for $t < n + 1$,

$$\mathrm{E}\left[\mathbf{T}_t\middle|\mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}, \Theta^{(j-1)}\right]$$
$$\mathrm{Var}\left[\mathbf{T}_t\middle|\mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}, \Theta^{(j-1)}\right]$$

and

$$\mathrm{Cov}\left[\mathbf{T}_t, \mathbf{T}_{t-1}\middle|\mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}, \Theta^{(j-1)}\right]$$

and subsequently apply Theorem 3.6.8 to obtain $H\left(\Theta|\Theta^{(j-1)}\right)$. From Figure 3.2, it is clear that the above quantities will be the same if one conditions only on $\mathbf{T}_{n+1}$ instead of $\mathbf{T}_{n+1:N}$. Hence, the first two quantities are just the modified Kalman smoother in Property 3.3. For the covariance, it turns out that it is equal to the lag covariance smoother stated in Property

3.4 because, for $t < n+1$,

$$\begin{aligned}
\mathrm{Cov}\left[\mathbf{T}_t, \mathbf{T}_{t-1} \middle| \mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}\right] &= \mathrm{Cov}\left[\mathbf{T}_t, \mathbf{T}_{t-1} \middle| \mathfrak{B}_N, \mathbf{T}_{n+1}\right] \\
&= \mathrm{E}\left[\mathrm{Cov}(\mathbf{T}_t, \mathbf{T}_{t-1} | \mathbf{T}_t, \mathbf{T}_{n+1}, \mathfrak{B}_N) \middle| \mathfrak{B}_N\right] \\
&\quad + \mathrm{Cov}\left[\mathrm{E}(\mathbf{T}_t | \mathbf{T}_t, \mathbf{T}_{n+1}, \mathfrak{B}_N), \mathrm{E}(\mathbf{T}_{t-1} | \mathbf{T}_t, \mathbf{T}_{n+1}, \mathfrak{B}_N) \middle| \mathfrak{B}_N\right] \\
&= \mathrm{Cov}\left[\mathbf{T}_t, \mathrm{E}(\mathbf{T}_{t-1} | \mathbf{T}_t, \mathfrak{B}_N) \middle| \mathfrak{B}_N\right] \\
&= \mathrm{Cov}\left(\mathbf{T}_t, \mathbf{T}_{t-1|t-1} + J_{t-1}(\mathbf{T}_t - \mathbf{T}_{t|t-1}) \middle| \mathfrak{B}_N\right) \\
&= \mathrm{Cov}(\mathbf{T}_t, J_{t-1}\mathbf{T}_t | \mathfrak{B}_N) \\
&= \mathbf{S}_{t,t-1|N}.
\end{aligned}$$

Thus, by Theorem 3.6.8, one now has

$$\begin{aligned}
&-2\ H\left(\Theta | \Theta^{(j-1)}\right) \\
&= \mathrm{E}_{\mathbf{T}}\left[-2 \ln f_\Theta(\mathbf{P}_{1:N}, \mathbf{T}_{0:N}) \middle| \mathbf{P}_{1:N}, \mathbf{T}_{n+1:N}, \Theta^{(j-1)}\right] \\
&= \ln|\Sigma_0| + N \ln|\mathcal{Q}| + N \ln|\mathcal{R}| \\
&\quad + \mathrm{tr}\left\{\Sigma_0^{-1}[\widetilde{\mathbf{S}}_{0|N} + (\widetilde{\mathbf{T}}_{0|N} - \mu_0)(\widetilde{\mathbf{T}}_{0|N} - \mu_0)^{\mathrm{T}}]\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{Q}^{-1} \sum_{t=1}^{N}\left[\widetilde{\mathbf{S}}_{t|N} + \phi\widetilde{\mathbf{S}}_{t-1|N}\phi^{\mathrm{T}} - \mathbf{S}_{t,t-1|N}\phi^{\mathrm{T}} - \phi\mathbf{S}_{t,t-1|N}^{\mathrm{T}}\right]\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{Q}^{-1} \sum_{t=1}^{N}\left[(\widetilde{\mathbf{T}}_{t|N} - \phi\widetilde{\mathbf{T}}_{t-1|N} - \delta\mathbf{F}_t)(\widetilde{\mathbf{T}}_{t|N} - \phi\widetilde{\mathbf{T}}_{t-1|N} - \delta\mathbf{F}_t)^{\mathrm{T}}\right]\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{Q}^{-1}\left[\phi\widetilde{\mathbf{S}}_{n|N}\phi^{\mathrm{T}} + (\mathbf{T}_{n+1} - \phi\widetilde{\mathbf{T}}_{n|N} - \delta\mathbf{F}_{n+1})(\mathbf{T}_{n+1} - \phi\widetilde{\mathbf{T}}_{n|N} - \delta\mathbf{F}_{n+1})^{\mathrm{T}}\right]\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{Q}^{-1} \sum_{t=n+2}^{N}(\mathbf{T}_t - \phi\mathbf{T}_{t-1} - \delta\mathbf{F}_t)(\mathbf{T}_t - \phi\mathbf{T}_{t-1} - \delta\mathbf{F}_t)^{\mathrm{T}}\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{R}^{-1} \sum_{t=1}^{n}\left[(\mathbf{P}_t - \mathcal{A}\widetilde{\mathbf{T}}_{t|N})(\mathbf{P}_t - \mathcal{A}\widetilde{\mathbf{T}}_{t|N})^{\mathrm{T}} + \mathcal{A}\widetilde{\mathbf{S}}_{t|N}\mathcal{A}^{\mathrm{T}}\right]\right\} \\
&\quad + \mathrm{tr}\left\{\mathcal{R}^{-1} \sum_{t=n+1}^{N}(\mathbf{P}_t - \mathcal{A}\mathbf{T}_t)(\mathbf{P}_t - \mathcal{A}\mathbf{T}_t)^{\mathrm{T}}\right\}.
\end{aligned}$$

Using Theorem 3.6.9 or by taking derivatives, the expressions stated in (3.3.3) can be obtained after some simplifications.

3.6.3 Miscellaneous results

Theorem 3.6.1 *Suppose a random $p \times 1$ vector $\mathbf{Y}$ has a multivariate normal distribution $\mathrm{N}_p(\mu, \Sigma)$ that is partitioned as*

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where $\mathbf{Y}_1$ and μ_1 are $r \times 1$, Σ_{11} is $r \times r$, $\mathbf{Y}_2$ and μ_2 are $(p-r) \times 1$, and Σ_{11} is $(p-r) \times (p-r)$. Then $\mathbf{Y}_1 \sim \mathrm{N}_r(\mu_1, \Sigma_{11})$ and $\mathbf{Y}_2 \sim \mathrm{N}_{p-r}(\mu_1, \Sigma_{22})$. Furthermore, the conditional distribution of $\mathbf{Y}_2$ given $\mathbf{Y}_1$ is described by

$$\mathbf{Y}_2 | \mathbf{Y}_1 \sim \mathrm{N}_{p-r}\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{Y}_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right). \tag{3.6.2}$$

□

Proof: See, for example, Anderson (2003, 33-34).

Theorem 3.6.2 *If every linear combination of the components of a vector $\mathbf{Y}$ is normally distributed, then $\mathbf{Y}$ is normally distributed.*

□

Proof: See, for example, Anderson (2003, 44).

Theorem 3.6.3 *Let $\mathbf{Y}$ be partitioned into $[\mathbf{Y}_1 \quad \mathbf{Y}_2]^{\mathrm{T}}$. If $\mathbf{Y}_1 \sim \mathrm{N}(\mu_1, \Sigma_{11})$ and $\mathbf{Y}_2 | \mathbf{Y}_1 \sim \mathrm{N}(A\mathbf{Y}_1 + b, \Psi)$ where A, b and Ψ do not depend on $\mathbf{Y}_1$, then $\mathbf{Y} \sim \mathrm{N}(\mu, \Sigma)$, where*

$$\mu = \begin{pmatrix} \mu_1 \\ A\mu_1 + b \end{pmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{11}A^{\mathrm{T}} \\ A\Sigma_{11} & \Psi + A\Sigma_{11}A^{\mathrm{T}} \end{bmatrix}.$$

□

Proof: By Theorem 3.6.1, if $\mathbf{Y}$ has a normal distribution, $\mathbf{Y}_2|\mathbf{Y}_1$ has a normal distribution defined by (3.6.2) and $\mathbf{Y}_1$ will be normally distributed. This relationship in the reverse direction is also true, because

$$f(\mathbf{Y}_1, \mathbf{Y}_2) = f(\mathbf{Y}_2|\mathbf{Y}_1)f(\mathbf{Y}_1).$$

So, in order to show the joint distribution of $(\mathbf{Y}_1, \mathbf{Y}_2)$ is the normal distribution defined in Theorem 3.6.3, it is sufficient to show that the condition in (3.6.2) holds. Now, the mean and variance of $\mathbf{Y}_2$ are

$$\begin{aligned}
\mu_2 &\stackrel{def}{=} \mathrm{E}(\mathbf{Y}_2) = \mathrm{E}[\mathrm{E}(\mathbf{Y}_2|\mathbf{Y}_1)] = A\mu_1 + b, \\
\Sigma_{22} &\stackrel{def}{=} \mathrm{Var}(\mathbf{Y}_2) = \mathrm{E}[\mathrm{Var}(\mathbf{Y}_2|\mathbf{Y}_1)] + \mathrm{Var}[\mathrm{E}(\mathbf{Y}_2|\mathbf{Y}_1)] = \Psi + A\Sigma_{11}A^{\mathrm{T}}
\end{aligned}$$

and

$$\begin{aligned}
\Sigma_{12} &\stackrel{def}{=} \mathrm{Cov}(\mathbf{Y}_1, \mathbf{Y}_2) \\
&= \mathrm{E}(\mathbf{Y}_1\mathbf{Y}_2^{\mathrm{T}}) - \mathrm{E}(\mathbf{Y}_1)E(\mathbf{Y}_2)^{\mathrm{T}} \\
&= \mathrm{E}[\mathrm{E}(\mathbf{Y}_1\mathbf{Y}_2^{\mathrm{T}}|\mathbf{Y}_1)] - \mathrm{E}(\mathbf{Y}_1)E(\mathbf{Y}_2)^{\mathrm{T}} \\
&= \mathrm{E}(\mathbf{Y}_1\mathbf{Y}_1^{\mathrm{T}}A^{\mathrm{T}} + \mathbf{Y}_1b^{\mathrm{T}}) - \mu_1(A\mu_1 + b)^{\mathrm{T}} = \Sigma_{11}A^{\mathrm{T}}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(\mu_1 - \mathbf{Y}_1) &= A\mu_1 + b - A\Sigma_{11}\Sigma_{11}^{-1}(\mu_1 - \mathbf{Y}_1) = A\mathbf{Y}_1 + b \\
\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} &= \Psi + A\Sigma_{11}A^{\mathrm{T}} - A\Sigma_{11}\Sigma_{11}^{-1}\Sigma_{11}A^{\mathrm{T}} = \Psi.
\end{aligned}$$

Thus (3.6.2) is satisfied.

■

Theorem 3.6.4 ***Jensen's inequality***

Let $f(\mathbf{X})$ *be a convex function. Then* $\mathrm{E}[f(\mathbf{X})] \geq f(E[\mathbf{X}])$.

□

Proof: See, for example, Chow and Teicher (1978, 103-104).

Theorem 3.6.5 *Let $\Theta^{(j-1)}$ be the current estimate of Θ for the likelihood function $l_{\mathbf{Y}}(\Theta)$ in (3.2.4) and choose $\Theta^{(j)}$ according to the EM algorithm, i.e. by maximizing $H(\Theta \mid \Theta^{(j-1)})$ in (3.2.4). Then*

$$l_{\mathbf{Y}}(\Theta^{(j)}) \geq l_{\mathbf{Y}}(\Theta^{(j-1)})$$

with equality if and only if

$$H(\Theta^{(j)} \mid \Theta^{(j-1)}) = H(\Theta^{(j-1)} \mid \Theta^{(j-1)}) \text{ and } Z(\Theta^{(j)} \mid \Theta^{(j-1)}) = Z(\Theta^{(j-1)} \mid \Theta^{(j-1)}).$$

□

Proof: Consider writing the change in the log likelihood function in successive iterations of the EM algorithm as

$$\begin{aligned} & l_{\mathbf{Y}}(\Theta^{(j)}) - l_{\mathbf{Y}}(\Theta^{(j-1)}) \\ &= \Big[H(\Theta^{(j)} \mid \Theta^{(j-1)}) - H(\Theta^{(j-1)} \mid \Theta^{(j-1)})\Big] + \Big[Z(\Theta^{(j-1)} \mid \Theta^{(j-1)}) - Z(\Theta^{(j)} \mid \Theta^{(j-1)})\Big] \end{aligned}$$

The first square bracket is non-negative since $\Theta^{(j)}$ is determined such that $H(\Theta^{(j)} \mid \Theta^{(j-1)}) \geq H(\Theta^{(j-1)} \mid \Theta^{(j-1)})$. The second square bracket can also be shown to be non-negative using Jensen's inequality.

$$\begin{aligned} & Z(\Theta^{(j-1)} \mid \Theta^{(j-1)}) - Z(\Theta^{(j)} \mid \Theta^{(j-1)}) \\ &= -\mathrm{E}_{\mathbf{X}}\left[\log \frac{f_{\Theta^{(j)}}(\mathbf{X}|\mathbf{Y})}{f_{\Theta^{(j-1)}}(\mathbf{X}|\mathbf{Y})}\middle| \mathbf{Y}, \Theta^{(j-1)}\right] \\ &\geq -\log \mathrm{E}_{\mathbf{X}}\left[\frac{f_{\Theta^{(j)}}(\mathbf{X}|\mathbf{Y})}{f_{\Theta^{(j-1)}}(\mathbf{X}|\mathbf{Y})}\middle| \mathbf{Y}, \Theta^{(j-1)}\right] \\ &= -\log \int \frac{f_{\Theta^{(j)}}(\mathbf{X}|\mathbf{Y})}{f_{\Theta^{(j-1)}}(\mathbf{X}|\mathbf{Y})} f(\mathbf{X}|\mathbf{Y}, \Theta^{(j-1)}) d\mathbf{X} \\ &= 0 \end{aligned}$$

■

Theorem 3.6.6 ***The Cramer-Wold device***

Let $\{\mathbf{X}_n\}$ *be a sequence of random k-vectors. Then* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *if and only if* $\lambda^{\mathrm{T}}\mathbf{X}_n \xrightarrow{d} \lambda^{\mathrm{T}}\mathbf{X}$ *for all* $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)^{\mathrm{T}} \in \mathbb{R}^k$.

□

Proof: See, for example, Brockwell and Davis (1991, 204-205).

Theorem 3.6.7 *If* $\mathbf{X}_t$ *is a stationary linear process which is a linear combination of white noise sequence* w_t *having variance* Q *and is of the form*

$$\mathbf{X}_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$$

with the coefficients satisfying

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty,$$

then provided that $\mathrm{E}(w_t^4)$ *is finite, as* $m \to \infty$, $m^{-1}\sum_{t=1}^{m} \mathbf{X}_t^2$ *is asymptotically normal with mean* $\mathrm{E}(\mathbf{X}_t^2)$.

□

Proof: See, for example, Shumway and Stoffer (2000, 73-75).

Theorem 3.6.8 *Let* $\mathbf{X} \sim \mathrm{N}(\mu, \Sigma)$, *where* μ *is a p-dimensional vector and* Σ *is a* $p \times p$ *positive definite matrix. Let* $\mathbf{Z}$ *be a* $p \times p$ *positive definite matrix and* ξ *be a p-dimensional constant vector. Then*

$$\begin{aligned}\mathrm{E}_{\mathbf{X}}\left[(\mathbf{X}-\xi)^{\mathrm{T}}\mathbf{Z}^{-1}(\mathbf{X}-\xi)\right] &= tr(\mathbf{Z}^{-1}\Sigma) + (\mu-\xi)^{\mathrm{T}}\mathbf{Z}^{-1}(\mu-\xi) \\ &= tr\left[\mathbf{Z}^{-1}(\Sigma + (\mu-\xi)(\mu-\xi)^{\mathrm{T}})\right].\end{aligned}$$

□

Proof: See, for example, Dethlefsen et al. (1997, 164).

Theorem 3.6.9 *Let* $\mathbf{Z}$ *and* Σ *be* $p \times p$ *positive definite matrices. The function*

$$f(\Sigma) = \ln|\Sigma| + tr(\Sigma^{-1}\mathbf{Z})$$

has a unique minimum at $\Sigma = \mathbf{Z}$.

Proof: See, for example, Dethlefsen et al. (1997, 165). □

4 Comparing historical temperature reconstruction methods

A number of studies have attempted to reconstruct hemispheric mean temperature for the past millennium from proxy climate indicators. However, there are considerable differences between the available reconstructions (Figure 1.2). In this chapter, analyses are carried out to investigate how much of this discrepancy is caused by the variations in reconstruction methods. The proposed state-space model approach discussed in the previous chapter will also be applied to reconstruct hemispheric mean temperature. The resulting reconstruction is then compared with reconstructed series from existing methods.

The remainder of this chapter is organized as follows. The existing statistical procedures used for reconstruction are discussed in section 4.1. Details on applying the state-space model to the reconstruction problem are given in section 4.2. Comparisons between different methods are made with the help of both climate model data and real-world paleoclimate proxy data. The results are presented in section 4.3 and 4.4. In terms of notation, $N = n + m$ and m are the length of the composite proxy record and instrumental record respectively, $\mathbf{T}_t$ is the mean hemispheric temperature at time t and $\mathbf{P}_t$ is the composite proxy record at time t.

4.1 A survey of existing reconstruction methods

In this section, existing reconstruction methods will be presented. Mann et al. (2005) have classified the existing approaches into two categories: ***composite plus scale (CPS)*** approaches and ***climate field reconstruction (CFR)*** approaches. In general, all reconstruction approaches involve statistical procedures that map the proxy series onto the reconstructed temperature. Such mapping involves first finding the relationship between the proxy data and the instrumental record over the calibration period. This relationship is then applied to the proxy data to reconstruct historical temperature.

4.1.1 Composite plus scale (CPS) approaches

In the CPS approach, a number of proxy series is first combined together to form a composite record, which is then used to reconstruct the temporal evolution of mean temperature over some spatial domain, typically the hemispheric mean. The composite record is typically formed by calculating a weighted average of all the available proxy series. The weights can either be uniform (e.g., Jones et al. 1998; Briffa et al. 2001; Esper et al. 2002) or can be determined using the correlation between the proxy series and the instrumental record during the calibration period (e.g., Hegerl et al. 2007). The correlation based weighting scheme has the advantage of minimizing the influence of potentially unreliable proxy series on the composite record. One recent study by Moberg et al. (2005) used a different averaging scheme, in which high-frequency and low-frequency composites were first formed individually using wavelet transformations. Each of these composites was formed using only proxy indicators that were thought to be able to capture variability at the corresponding frequency range. The two composites are then combined to form a single composite. This averaging scheme can be beneficial when the individual proxy series are known to capture variability at different timescales.

Once the composite is formed, it is then calibrated to produce the reconstructed temperature. One calibration approach, the ***forward regression approach***, utilizes the ordinary least squares regression model which assumes that, at time t, for $t = n+1,\ n+2,\ ...,\ N$,

$$\mathbf{T}_t = \alpha \mathbf{P}_t + \varepsilon_t \tag{4.1.1}$$

where ε_t represents error that results from incomplete spatial sampling in the instrumental record. The coefficient α scales $\mathbf{P}_t$ to $\mathbf{T}_t$ and is estimated by ordinary least squares regression which minimizes the residual sum of squares between $\alpha\mathbf{P}_t$ and $\mathbf{T}_t$ during the calibration period. By assuming that (4.1.1) also holds at times prior to the calibration period, the unknown historical hemispheric mean temperature can then be reconstructed by scaling the pre-calibration period composite record $\mathbf{P}_t$ by α, for $t = 1,\ 2,\ ...,\ n$. The estimate of α, denoted by $\hat{\alpha}$, is in general negatively biased because of the measurement error and non-

temperature variability inherit in the composite proxy series (see, e.g., Fuller 1987, 2-5; Allen and Stott 2003). This under-estimation will result in a loss of variance in the reconstructed series because $\mathrm{Var}(\hat{\alpha}\mathbf{P}) < \mathrm{Var}(\alpha\mathbf{P})$ for $0 < \hat{\alpha} < \alpha$.

One way to avoid this underestimation in α is to use the ***total least squares*** (TLS) method to estimate α (see Allen and Stott 2003 for an application of this technique in climate change detection analysis). In this case, the statistical relationship between $\mathbf{P}_t$ and $\mathbf{T}_t$ is defined by

$$\mathbf{T}_t = \alpha(\mathbf{P}_t - e_t) + \varepsilon_t \tag{4.1.2}$$

where the additional term e_t represents non-temperature variability in the composite proxy series. By explicitly incorporating e_t into the regression model, the underlying value of α, in theory, can be better estimated. Hegerl et al. (2007a) applied the total least squares method in their reconstruction. To derive the best guess estimate of α, one needs to know the ratio of $\mathrm{Var}(e_t)$ to $\mathrm{Var}(\varepsilon_t)$. Since this ratio is unknown in real-world applications, it is estimated using climate model simulations (see Hegerl et al. 2007a for detail). Hegerl et al. (2007a) find that their reconstruction is insensitive to the precise choice of the $\mathrm{Var}(e_t)$ to $\mathrm{Var}(\varepsilon_t)$ ratio.

Another CPS reconstruction method, which is termed the ***variance matching approach***, is favored by Jones et al. (1998) and others. In this approach, the pre-calibration period $\mathbf{P}_t$ is scaled by a parameter β, where β is determined so that, during the calibration period, $\mathrm{Var}(\mathbf{T})$=$\mathrm{Var}(\beta\mathbf{P})$.

A variant of the forward regression model (4.1.1) is the ***inverse regression*** model (see Brown 1993 for an introduction and Coehlo et al. 2004 for an example)

$$\mathbf{P}_t = \mathcal{A}\mathbf{T}_t + e_t \tag{4.1.3}$$

where e_t is defined similarly as in (4.1.2) and $\mathcal{A}$ is estimated by minimizing the residual sum of squares between $\mathcal{A}\mathbf{T}_t$ and $\mathbf{P}_t$ during the calibration period. The reconstructed hemispheric temperature is then estimated by $\mathbf{P}_t/\mathcal{A}$. The above equation is, in fact, the same equation as the observation equation in the Gaussian state-space model in (3.1.2). Different implementations of the inverse regression method have been used in the paleoclimate reconstruction

literature. In Mann et al. (1998), the regression is done between the principal components of the instrumental record and the individual proxy indictor (see latter part of this section for details), some of which were obtained by principle component analysis from a set of proxy series. In Juckes et al. (2006), the inverse regression is done between each individual proxy series and the Northern Hemisphere annual mean temperature. A weighted average of the individual proxy series, weights determined by the regression coefficients, is then used as the reconstructed series.

Note that the instrumental record $\mathbf{T}_t$ is subject to sampling uncertainty (see Jones et al., (1997) for an estimate of this uncertainty) and neglecting such uncertainty when estimating $\mathcal{A}$ in inverse regression will result in an estimate that is negatively biased. However, such bias would be smaller than the bias incurred by using forward regression because the measurement error inherent in the composite record is usually larger than the sampling uncertainty in the instrumental record. On the other hand, the total least squares approach used by Hegerl et al. (2007a), in theory, should provide a more accurate estimate of the regression coefficient when both $\mathbf{P}_t$ and $\mathbf{T}_t$ are noise contaminated. However, this only holds if the ratio of $\mathrm{Var}(e_t)$ to $\mathrm{Var}(\varepsilon_t)$, which is required to derive the estimate of α, is known. This ratio is unknown in real-world applications and is estimated using a limited number of climate model simulations. The bias from such estimation is hard to determine and thus its impact on the reconstruction is unclear. For this reason, it is not obvious whether inverse regression or total least squares regression will result in smaller bias.

4.1.2 Climate field reconstruction (CFR) approaches

In the case of the CFR approach, the proxy series are used to reconstruct both the underlying temporal and spatial patterns of historical temperature. The Mann et al. (1998) method (often referred to as the ***MBH method***) is an example of a technique that uses the climate field reconstruction approach. The MBH method brings together techniques used in principal component analysis and regression. The instrumental record is first decomposed into its spatial and temporal parts through principal component analysis and only a subset of the components are retained. Next, the relationship between the subset of temporal principal

components (PCs) and the i^{th} proxy series ($i = 1, 2, ..., p$) during the calibration period is obtained by inverse regression which estimates the coefficients $\beta_j^{(i)}$ ($j = 1, 2, ..., N_{eof}$) in

$$\begin{bmatrix} \mathbf{P}_{n+1}^{(i)} \\ \mathbf{P}_{n+2}^{(i)} \\ \vdots \\ \vdots \\ \mathbf{P}_{n+m}^{(i)} \end{bmatrix} = \mathbf{U} \begin{bmatrix} \beta_1^{(i)} \\ \beta_2^{(i)} \\ \vdots \\ \beta_{N_{eof}}^{(i)} \end{bmatrix} + \xi^{(i)},$$

where N_{eof} is the number of EOFs (temporal PCs) retained, $\mathbf{P}_t^{(i)}$ is the i^{th} proxy data at time t , $\mathbf{U}$ is the matrix of temporal PCs ($m \times N_{eof}$) of the instrumental record and $\xi^{(i)}$ is a noise series ($m \times 1$). This procedure is repeated for each proxy series and this yields a matrix of coefficients, denoted by $\mathbf{G}$ ($p \times N_{eof}$), which is defined as,

$$\mathbf{G} = \begin{bmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \cdots & \beta_{N_{eof}}^{(1)} \\ \beta_1^{(2)} & \beta_2^{(2)} & \cdots & \beta_{N_{eof}}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_1^{(p)} & \beta_2^{(p)} & \cdots & \beta_{N_{eof}}^{(p)} \end{bmatrix}.$$

The pre-calibration period temporal PCs at time t, denoted by $\mathbf{Z}_t$ ($N_{eof} \times 1$), are then reconstructed through least squares regression using

$$\begin{bmatrix} \mathbf{P}_t^{(1)} \\ \mathbf{P}_t^{(2)} \\ \vdots \\ \mathbf{P}_t^{(p)} \end{bmatrix} = \mathbf{G}\mathbf{Z}_t + \kappa_t,$$

where κ_t is a noise series ($p \times 1$). The sequence of reconstructed temporal PCs, $\hat{\mathbf{Z}}_t$, is then scaled to have the same variance as the instrumental temporal PCs over the calibration period and subsequently re-combined with the calibration period spatial PCs to provide an estimate of the unknown local temperature. A hemispheric mean reconstruction is then formed by the

appropriate spatial average of the reconstructed local temperatures. The regression procedures above are both inverse regression. However, unlike the CPS inverse regression, the individual proxy series is used in the regression process. Also, there are N_{eof} coefficients to estimate in each regression, compared to only one coefficient in the CPS case.

Another CFR method uses the ***regularized Expectation Maximization (RegEM)*** algorithm (Schneider, 2001) and is advocated by Rutherford et al. (2005) and others. Let $\mathbf{T}_t^{(j)}$ be the local temperature at time t at the j^{th} location ($j = 1, 2, ..., q$) and let $\mathbf{X}$ be a $(n+m) \times (q+p)$ matrix with missing data which is defined as

$$\mathbf{X} = \begin{bmatrix} \mathbf{T}_1^{(1)} & \mathbf{T}_1^{(2)} & \cdots & \mathbf{T}_1^{(q)} & \mathbf{P}_1^{(1)} & \mathbf{P}_1^{(2)} & \cdots & \mathbf{P}_1^{(p)} \\ \mathbf{T}_2^{(1)} & \mathbf{T}_2^{(2)} & \cdots & \mathbf{T}_2^{(q)} & \mathbf{P}_2^{(1)} & \mathbf{P}_2^{(2)} & \cdots & \mathbf{P}_2^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{T}_{n+m}^{(1)} & \mathbf{T}_{n+m}^{(2)} & \cdots & \mathbf{T}_{n+m}^{(q)} & \mathbf{P}_{n+m}^{(1)} & \mathbf{P}_{n+m}^{(2)} & \cdots & \mathbf{P}_{n+m}^{(p)} \end{bmatrix}$$

$$\stackrel{def}{=} \begin{bmatrix} \vec{\mathbf{T}}_1 & \vec{\mathbf{P}}_1 \\ \vec{\mathbf{T}}_2 & \vec{\mathbf{P}}_2 \\ \vdots & \vdots \\ \vec{\mathbf{T}}_{n+m} & \vec{\mathbf{P}}_{n+m} \end{bmatrix}$$

where

$$\vec{\mathbf{T}}_t = [\mathbf{T}_t^{(1)} \quad \mathbf{T}_t^{(2)} \quad \cdots \quad \mathbf{T}_t^{(q)}]$$
$$\vec{\mathbf{P}}_t = [\mathbf{P}_t^{(1)} \quad \mathbf{P}_t^{(2)} \quad \cdots \quad \mathbf{P}_t^{(p)}],$$

with

$$\mathrm{E}[\vec{\mathbf{T}}_t \quad \vec{\mathbf{P}}_t] = \mu = [\vec{\mu_T} \quad \vec{\mu_P}], \qquad \text{for } t = 1,\ 2,\ ...,\ n+m$$

$$\mathrm{Var}(\mathbf{X}) = \Sigma = \begin{bmatrix} \Sigma_{TT} & \Sigma_{TP} \\ \Sigma_{PT} & \Sigma_{PP} \end{bmatrix}$$

where $\vec{\mu_T}$ and $\vec{\mu_P}$ have length q and p respectively. The $(q+p) \times (q+p)$ matrix Σ is partitioned according to $\mathbf{T}$ and $\mathbf{P}$ so that Σ_{TT}, Σ_{PP} and $\Sigma_{TP} = \Sigma_{PT}^{\mathrm{T}}$ are $q \times q$, $p \times p$ and $q \times p$ matrices respectively. In the matrix $\mathbf{X}$, $\vec{\mathbf{T}}_t$ is unknown for $t = 1, 2, .., n$. It is assumed that each record

$\vec{\mathbf{T}}_t$ can be represented by a linear model of the form

$$\vec{\mathbf{T}}_t = \vec{\mu_T} + (\vec{\mathbf{P}}_t - \vec{\mu_P})\mathbf{B} + \vec{\epsilon_t}.$$

Here, $\vec{\epsilon_t}$ is the residual and $\mathbf{B}$ is a $p \times q$ matrix of regression coefficients. The reconstructed temperature $\hat{\vec{\mathbf{T}}}_s$ at time s $(s = 1, 2, ..., n)$, is thus defined as $\vec{\mu_T} + (\vec{\mathbf{P}}_s - \vec{\mu_P})\mathbf{B}$.

To calculate $\hat{\vec{\mathbf{T}}}_s$, estimates of $\mathbf{B}$ and μ are required. If $n + m \geq p + 1$, the Expectation Maximization (EM) algorithm provides a way to iteratively estimate these parameters (see, e.g., Schneider 2001). First, initialize the unknown $\hat{\vec{\mathbf{T}}}_t$ with some values and obtain an estimate of the mean and covariance of matrix $\mathbf{X}$ (see Schneider 2001 for the formulas). Then, estimate $\mathbf{B}$ using the maximum likelihood estimate $\hat{\mathbf{B}} = \hat{\Sigma}_{PP}^{-1}\hat{\Sigma}_{PT}$, where the $\hat{\Sigma}_{PP}$ and $\hat{\Sigma}_{PT}$ denote the partitioned covariance matrix estimate. If $n + m < p + 1$, the estimate of Σ_{PP} is singular and the coefficient $\mathbf{B}$ is not defined. Next, impute the missing $\hat{\vec{\mathbf{T}}}_t$ by $\hat{\vec{\mu_T}} + (\vec{\mathbf{P}}_t - \hat{\vec{\mu_P}})\hat{\mathbf{B}}$ and update the $\mathbf{X}$ matrix. Re-estimate μ, Σ and $\mathbf{B}$ using the updated matrix and then recalculate the missing temperatures. This process is continued until the change in imputed values become sufficiently small.

In a typical CFR application, $n + m$ is greater than $p + 1$. However, the estimate of Σ is rank deficient because $n + m < p + q + 1$, which can result in a poor estimate of the coefficient $\mathbf{B}$. Hence, the RegEM algorithm is used instead. The RegEM algorithm consists of the same steps as the EM algorithm, except the maximum likelihood estimate $\hat{\mathbf{B}}$ is replaced with a regularized estimate, which is obtained by a regularized regression procedure. Different regularization schemes have been used. In Rutherford et al. (2005) and Mann et al. (2005), the ridge regression scheme is used to estimate the coefficient matrix $\mathbf{B}$ (Schneider, 2001). In Mann et al. (2007), truncated total least squares (TTLS) (Fierro et al., 1997) is used to estimate $\mathbf{B}$. Given the complexity of these regularization procedures, readers are referred to the corresponding references for more details.

4.2 Applying the state-space model for reconstruction

In addition to the methods described above, one can also use the state-space model approach proposed in Chapter 3 to reconstruct historical temperature. The state-space representation of the hemispheric mean temperature $\mathbf{T}_t$ ($t = 1,\ 2,\ ...,\ n+m$) is the univariate Gaussian state-space model in (3.1.4), which is given by the following system of equations:

$$\begin{aligned} \mathbf{P}_t &= \mathcal{A}\mathbf{T}_t + e_t \\ \mathbf{T}_t &= \phi\mathbf{T}_{t-1} + \delta\mathbf{F}_t + w_t \end{aligned} \tag{4.2.1}$$

where $\delta = [\delta_0 \quad \delta_{\mathbf{GS}} \quad \delta_{\mathbf{VOL}} \quad \delta_{\mathbf{SOL}}]$ are constants, $\mathbf{F}_t = \mathbf{X}_t - \phi\mathbf{X}_{t-1}$ and $\mathbf{X}_t = [1 \quad \mathbf{GS}_t \quad \mathbf{VOL}_t \quad \mathbf{SOL}_t]^{\mathrm{T}}$ are estimated responses to greenhouse gas and sulphate aerosol forcing combined (GS), volcanic forcing (VOL) and solar forcing (SOL). Unlike the simulation studies in Chapter 3, the response to each individual forcing is used rather than the combined response. Same as in Chapter 3, $\mathbf{X}_t$ is obtained from EBM simulations that are forced with estimates of historical forcing. The EBM simulation used here is the same as that used in Hegerl et al. (2007a), from which the 30N-90N average response to greenhouse gas, sulfate aerosol, volcanic and solar forcing are available from 1000-1997. Alternatively, one can also define a state equation that does not contain the response to forcings. This can be achieved by setting $\mathbf{X}_t = 1$ and $\delta = \delta_0$. The observation equation, which is the same equation as that in inverse regression, simply assumes that the relationship between the composite proxy series and hemispheric temperature that holds during the calibration period in (4.1.3) also holds in the pre-calibration period.

The problem of estimating the pre-calibration period $\mathbf{T}_t$ in a state-space model can be approached by using the Kalman filter and smoother algorithm as presented in Chapter 3. Alternatively, since the state process is partially observed in the reconstruction problem, one can use the modified Kalman smoother (Property 3.3) to reconstruct the unknown historical temperature. At the time of running the analysis in this chapter, the modified Kalman smoother has not yet been derived and so historical temperature is reconstructed using the existing Kalman smoother algorithm (Property 3.2). As pointed out in Chapter 3, the difference between the reconstructed series from the two smoothers should be small given that ϕ is

likely not close to 1 in the reconstruction problem (Figure 3.3).

The state-space model approach can also be extended to produce forecasts. Provided that $\mathbf{F}_t$ is known for the future time period, one can use the state equation to generate a forecast recursively. Recall that the state equation at time t is given by $\mathbf{T}_t = \phi\mathbf{T}_{t-1} + \delta\mathbf{F}_t + w_t$. By forecasting the error term w_t as zero, the forecast of $\mathbf{T}_{n+m+s}$ ($s = 1,\ 2,\ ....$) can be given by

$$\grave{\mathbf{T}}_{n+m+s} = \phi\grave{\mathbf{T}}_{n+m+s-1} + \delta\mathbf{F}_{n+m+s} \tag{4.2.2}$$

with $\grave{\mathbf{T}}_{n+m} = \mathbf{T}_{n+m}$ being the known instrumental record at time $n+m$. Forecasts can also be made in the case when the response to external forcing is not included in (4.2.1), i.e., when $\mathbf{X}_t = 1$ and $\delta = \delta_0$. However, such forecasts will likely not be very useful because they do not take the impact of forcings into account and thus quickly revert to a forecast of the mean δ_0. Confidence bound on the forecast can be derived based on the assumption that w_t is normally distributed. The variance of the forecast is given by $\mathrm{Var}(\grave{\mathbf{T}}_{n+m+s}) = \phi^2\mathrm{Var}(\grave{\mathbf{T}}_{n+m+s-1}) + \mathcal{Q}$ with $\mathrm{Var}(\grave{\mathbf{T}}_{n+m}) = 0$.

A difficulty in using the state-space time series model is that the parameters, $\{\mathcal{A}, \mathcal{R}, \phi, \delta, \mathcal{Q}, \mu_0, \Sigma_0\}$, are unknown and need to be estimated. As pointed out in Chapter 3, the method of maximum likelihood can be used to estimate the parameters. Note that during maximum likelihood estimation, the exogenous variable $\mathbf{F}_t$ is assumed to be known. If ϕ is unknown, the exogenous variables therefore cannot contain the parameter ϕ. Thus, for the state-space representation of hemispheric mean temperature in (4.2.1), one needs to fix the parameter ϕ that is involved in $\mathbf{F}_t = \mathbf{X}_t - \phi\mathbf{X}_{t-1}$ before estimating the other parameters. More details on the choice of ϕ are given in the next section. Note that there is no need to fix the parameter ϕ in front of $\mathbf{T}_{t-1}$ in (4.2.1) and this parameter can be estimated together with the other parameters through maximum likelihood estimation. From Chapter 3, it has been shown that the estimates from the $\mathbb{ALL}$ estimation approach are the most efficient when the state process is partially observed. Hence the $\mathbb{ALL}$ approach will be used in here to estimate the parameters in the state-space model. The EM algorithm will be used to numerically solve for the optimal parameter value. Note that the parameter Σ_0 is not estimatable using the EM algorithm and

hence it is fix at 0.05 for the following analysis. Results are in fact almost identical when different Σ_0 value is used.

An advantage of using the state-space model to reconstruct historical temperature is that it provides the flexibility to incorporate forcing response information into the estimation of the unknown temperature. This is achieved in a two step process that first determines the impact of each forcing on the unknown temperature through the estimation of the forcing coefficients. The estimation process uses the information available in both the proxy series and the calibration data. From (4.2.1), it is obvious that if the forcing coefficient is significantly different from zero, one can claim that the corresponding forcing has a significant impact on the hemispheric mean temperature. Such an assessment can be made using the confidence bound of each coefficient obtained during the parameter estimation step. As discussed in Chapter 3, confidence bounds for the estimates obtained from the $\mathbb{ALL}$ estimation approach can be approximated based on the normality assumption and with the Hessian matrix. Once the coefficients are estimated, the forcing information is then incorporated into the final estimate of the unknown temperature using the Kalman filter and smoother algorithm. In other words, the use of the state-space model allows one to simultaneously reconstruct the unknown hemispheric mean temperature and conduct a detection assessment on the importance of the response to GS, VOL and SOL forcing on hemispheric temperature. Furthermore, the use of the state-space model allows one to provide projections of future climate which are based on parameters that are estimated using the proxy records and past observations.

4.3 Comparison using climate model simulations

In general, it is difficult to compare the performance of different reconstruction methods using the past instrumental temperature record because the instrumental period is simply too short to calibrate such techniques and reliably assess their performance. Climate model simulations, however, can provide a test bed for assessing the reliability of these methods as first introduced by Zorita et al. (2003; see also von Storch et al. 2004; Mann et al. 2005; Zorita and von Storch 2005 and others). In this section, the methodology proposed by von Storch et al. (2004) will be used. Figure 4.1 visualizes the testing procedures. The idea is to generate

pseudo-proxy records by sampling a selection of simulated grid-box temperatures from the climate model and degrading them with additive noise. The reconstruction method is then applied to these pseudo-proxy records and the resulting reconstruction is validated against the known simulated hemispheric mean temperature record. To reflect the differences in the quality and properties of real-world proxy data, different colours of noise (e.g., red rather than white) and varying amplitudes of the noise variance have been used in different studies.

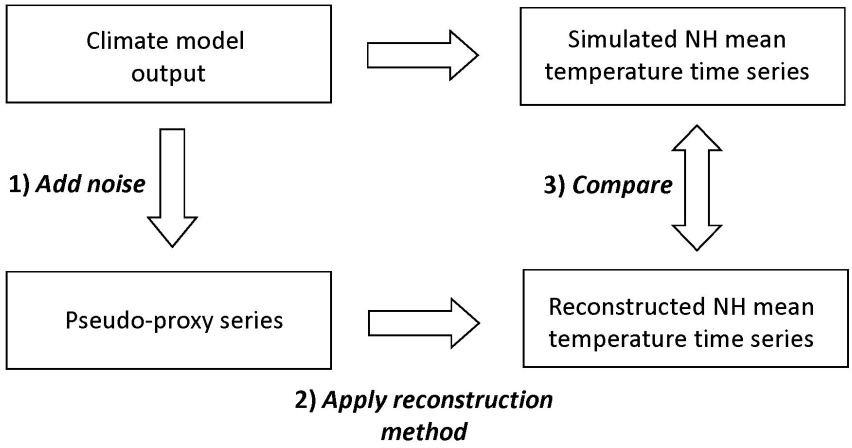

Figure 4.1: Visualization of the testing procedures proposed by von Storch et al. (2004).

A suite of experiments has been conducted to explore the sensitivity of each method to 1) the climate model simulation used, 2) the length of the calibration period, and 3) the amount of noise introduced into the pseudo-proxy series. The two simulations used are the GKSS ECHO-G simulation (von Storch et al. 2004) and a simulation from the NCAR CSM 1.4 model (Ammann et al. 2007). Both simulations were forced with reconstructions of solar, volcanic and greenhouse gas forcing. The CSM 1.4 run was also forced with a reconstruction of aerosol forcing over the millennium. For both simulations, only output between years 1000 and 1990 is used. To represent the varying calibration intervals that were used in actual reconstructions, two calibration periods were used in this analysis (1880-1960 and 1860-1970). These calibration periods are similar to that used in Hegerl et al. (2007a) and Moberg et al. (2005). In the following experiments, pseudo-proxy records were formed by degrading the grid box temperatures with additive white noise. The amount of noise introduced into the pseudo-

proxy series is expressed in terms of the ***signal-to-noise ratio (SNR)***, which is defined as $\sqrt{\mathrm{Var}(\mathcal{X})/\mathrm{Var}(\mathcal{N})}$, where $\mathcal{X}$ is the grid box temperature series and $\mathcal{N}$ is the additive noise series. For all methods, experiments with SNR = 0.5 and 1 were performed.

First, a pseudo northern hemisphere instrumental record is obtained from the simulation. CFR methods use a set of continuous grid-box temperatures for analysis; therefore the spatial coverage of the entire pseudo-instrumental record is fixed at that of the instrumental network in 1920, as represented in the HadCRUT3 data set (Brohan et al., 2006). The choice of the year 1920 is arbitrary, but it corresponds roughly to the mid points of the calibration periods that were used in the analysis. For the other methods, the pseudo NH mean instrumental record is calculated from the appropriate areal average of the same grid-box temperatures used above. This ensures that all methods are provided with the same information for calibration.

Next, two pseudo-proxy networks of 15 and 100 randomly selected model grid boxes are defined. These networks were sampled from the 321 NH grid boxes that are co-located with actual tree ring data found in the International Tree Ring Data Base (http://www.ncdc.noaa.gov/paleo/treering.html). These networks of grid-box temperature were converted to pseudo-proxy records by degrading the grid-box temperatures with added white noise to mimic the measurement error that is inherent in the proxy records. The resulting pseudo-proxy series were then standardized relative to the calibration period. Since the RegEM method is computationally intensive, this method will only be applied to the larger network of the two. The size of this network is similar to networks that are used in the real-world application of this technique. On the other hand, the CPS methods and state-space model approach are less computationally intensive, and these methods were tested using both networks.

Uniformly weighted spatial averages are used to form the composite record in the analysis; composites that are formed by wavelet transformation (Moberg et al., 2005) or correlation based weighted averages (Hegerl et al., 2007a) are not be considered here. Since the SNR and colour of the noise in each pseudo-proxy series is the same, the different weighting scheme should not substantially impact the resulting composite record.

For the total least squares method, $\mathrm{Var}(\varepsilon_t)$ is estimated by calculating the variability of the difference between the pseudo-instrumental record and the climate model simulated

NH temperature over the whole reconstruction period. Since the pre-noise contaminated composite record is known in the analysis, $\mathrm{Var}(e_t)$ is estimated by calculating the variability of the difference between the pre-noise contaminated and noise contaminated composite pseudo-proxy record, instead of following the procedure described in Hegerl et al. (2007a). This is a rather idealized situation in the sense that information which is not available in real-world application is used to estimate the two variances.

For the MBH method, the 10 largest EOFs are retained instead of using the selection rule that was described in Mann et al. (1998). Analysis is repeated by retaining the 5 or 15 largest EOFs and the results are almost identical to that obtained with the 10 largest EOFs. Thus, these results will not be shown. Von Storch et al. (2004, 2006) included a detrending step in their test of the MBH method, where the linear trend in the calibration period data is removed prior to calibration. Such detrending procedure is not involved in the original implementation of the MBH method in Mann et al. (1998). However, von Storch et al. (2006) argued that climate proxy might contain nonclimatic trends and hence the use of non-detrended data can be dangerous. On the other hand, detrending the calibration data removes the 20^{th} century warming trend, which contains important information for the reconstruction method to correctly identify the relationship between the proxy and the instrumental record. Further discussions on the use of detrended data can be found in von Storch et al. (2006) and Wahl et al. (2006). In the following experiments, no detrending is done prior to calibration.

For the RegEM method, the hybrid non-stepwise approach is used to reconstruct the grid box temperatures (Rutherford et al. 2005; Mann et al. 2005). As in Rutherford et al. (2005) and Mann et al. (2005), the ridge regression procedure is used to regularize the EM algorithm. Prior to reconstruction, a weight is applied to each standardized proxy series to ensure that the error variance of the signal in the series are homogenous among all records. The weight for the i^{th} proxy series is defined as $\sqrt{\mathrm{Var}(\mathbf{P}^{(i)})/\mathrm{Var}(\mathbf{S}^{(i)})}$, where $\mathbf{P}^{(i)}$ is the i^{th} proxy series and $\mathbf{S}^{(i)}$ is the signal in the i^{th} proxy series. However, $\mathbf{S}^{(i)}$ is unknown in real-world applications. An approximation of this weight can be provided by the sample correlation coefficient between the proxy series and the associated grid box temperature over the calibration period (M. Mann and S. Rutherford, pers. comm., 2006). Such a weighting

approach is not implemented in Rutherford et al. (2005) and Mann et al. (2005). However, through experiments with the RegEM method (not shown) and personal communications (2006) with M. Mann and S. Rutherford, it has been confirmed that reconstruction of the CSM hemispheric mean temperature is sensitive to whether weighted proxy series are used. However, this only applies to reconstructions that use ridge regression to regularize the EM algorithm. Mann et al. (2007) found that results obtained using TTLS regression for regularization are insensitive to whether weights are used. They also pointed out that regularization using TTLS regression is less computationally intensive and tends to provide more robust results than regularization with ridge regression. Hence regularization using TTLS regression would be preferable. However, we were not aware of these advantages at the time of running our experiments, and hence results obtained with the ridge regression procedure will be reported. Mann et al. (2005) found that RegEM reconstructions are relatively insensitive to the use of a shorter calibration period and hence, in this analysis, RegEM experiments were only carried out for the 1860-1970 calibration period.

For the state-space model approach, two separate types of experiments are run in which the variable $\mathbf{X}_t$ in (4.2.1) is defined differently. The first type of experiment accounts for the impact of external forcing when reconstructing the unknown temperature. This is achieved by setting $\mathbf{X}_t = [1 \quad \mathbf{GS}_t \quad \mathbf{VOL}_t \quad \mathbf{SOL}_t]^{\mathrm{T}}$. The second type of experiment only uses information from the proxy data for reconstruction and is done by using $\mathbf{X}_t = 1$ and $\delta = \delta_0$. For experiments that investigate the impact of external forcing, the variable $\mathbf{X}_t$ is obtained from an energy balanced model (EBM) driven with reconstructed solar, volcanic and anthropogenic forcings (Hegerl et al. 2003, 2007a) . The EBM simulation used here is the same as that used in Chapter 3 and Hegerl et al. (2007a), from which the 30N-90N average response to greenhouse gas, sulfate aerosol, volcanic and solar forcing are available.

In all the experiments, all parameters in (4.2.1) except one are estimated through the EM algorithm. Exception is the parameter ϕ in the exogenous variable $\mathbf{F}_t$, which needs to be estimated outside of the EM algorithm. The value of ϕ will be data dependent because ϕ represents the lag-one autocorrelation of internal variability. For the following analysis, it is estimated by calculating the lag-one autocorrelation of the residuals that result from

fitting (3.1.3) using the CSM or the ECHO-G model simulated northern hemisphere mean temperature as $\mathbf{T}_t$ and the EBM simulated response to forcings mentioned above as $\mathbf{X}_t$. For the CSM and ECHO-G simulations, ϕ is estimated to be 0.581 and 0.741 respectively. The exogenous variable $\mathbf{F}_t$ used in the analysis is then obtained using these ϕ values. The precise choice of ϕ turns out to have very little impact on the resulting reconstruction (not shown).

Annual mean data are used for all reconstruction methods with some exceptions. Following the typical CPS procedure, to estimate the parameters α and β in the forward regression and the variance matching methods, decadally smoothed data are used. The estimated parameters are then used to scale the annual mean pseudo-proxy series to provide the reconstructed annual hemispheric mean temperature. For comparison, temperatures are also reconstructed using parameters that are estimated with annual mean data and these reconstructions will be denoted as the non-smoothed forward regression and non-smoothed variance matching reconstructions. On the other hand, as in Mann et al. (1998), the pseudo-instrumental record used in the principle component analysis of the MBH method is expressed as monthly means and annual mean PCs are subsequently obtained from the monthly mean PCs for analysis.

Figure 4.2a shows examples of reconstructed NH temperature evolution simulated by the CSM. The reconstructions are based on 15 pseudo-proxy series with SNR=0.5. It should be noted that the same pseudo-proxy series and pseudo-instrumental record was used in each method to ensure that differences in performance are solely due to the methods themselves. Among the methods that are considered, most did provide a faithful estimate for the simulated NH temperature. The forward regression (smoothed and non-smoothed) and the non-smoothed variance matching methods are exceptions. Comparing the series reconstructed with inverse regression and total least squares regression, it is clear that the two series are visually indistinguishable, suggesting that the neglect of sampling uncertainty in the instrumental record when estimating $\mathcal{A}$ does not have a substantial impact on the results. On the other hand, neglecting the measurement error in the composite record when estimating α can cause a substantial bias in the reconstruction. This can be observed by comparing the reconstructed series obtained with forward regression to that obtained with total least squares regression. Experiments using 100 pseudo-proxy series with SNR=0.5 are displayed in Figure 4.2b. It is

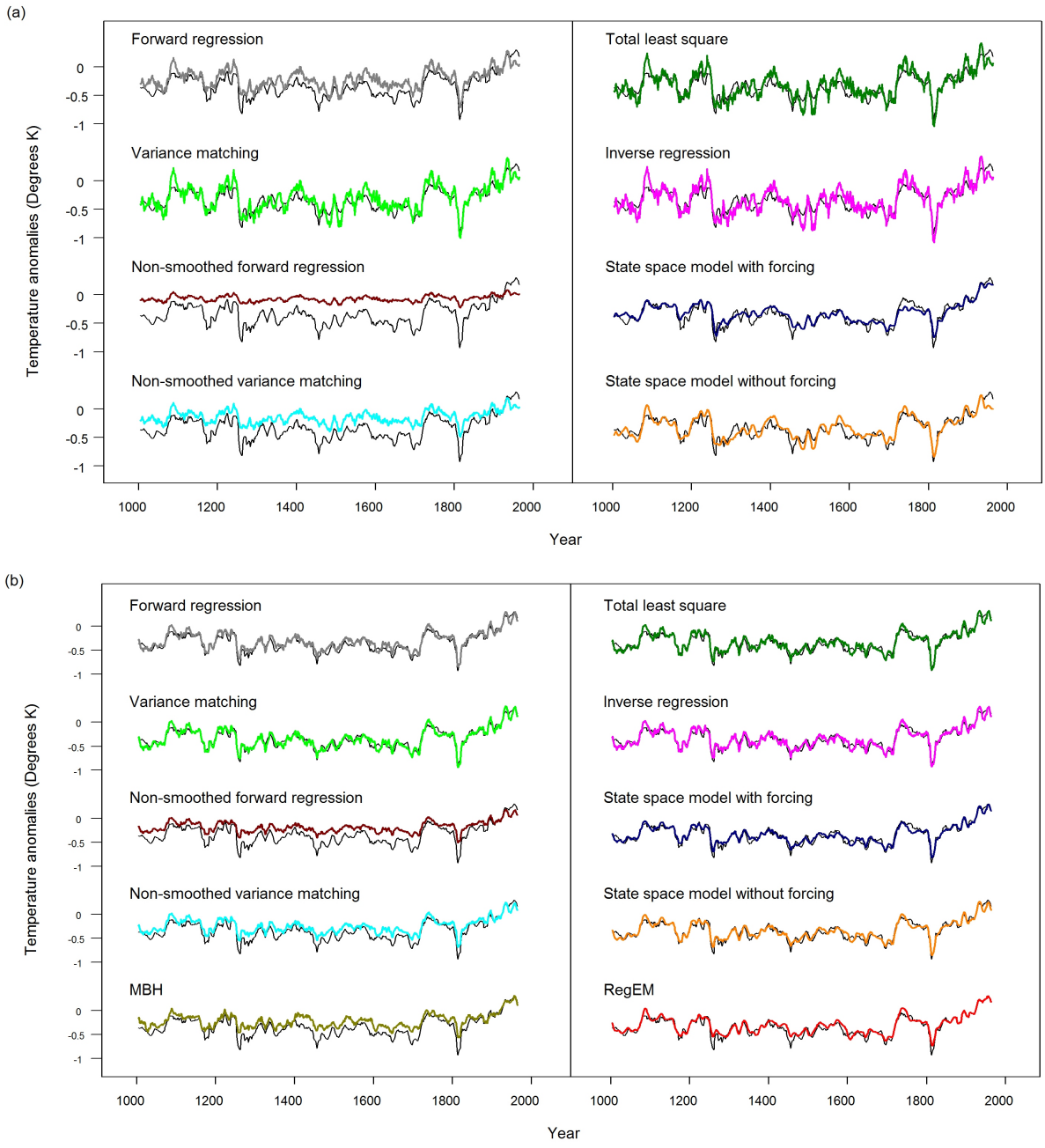

Figure 4.2: Examples of reconstructed CSM northern hemisphere mean temperature series. Experiments were run using SNR=0.5 and the 1860-1970 calibration period with a) 15 and b) 100 pseudo-proxy series. 11-yr moving averages are shown. The CSM NH mean is shown in black. Reconstructions are shown as colour line. All series are express as anomalies relative to the calibration period. Units (degrees Kelvin).

clear that the increase in the number of pseudo-proxy series does seem to alleviate the problem in the smoothed forward regression method. The improvement gained when going from 15 to 100 pseudo-proxy series is expected because the noise variance in the composite record of 100pseudo-proxy series is only 15% of that with 15 pseudo-proxy series. However, such improvement may be less significant in real-world applications given that errors can be spatially correlated. Results obtained with the ECHO-G simulation are very similar and thus not shown.

The test result for a reconstruction method can be affected by both the specific realizations of noise that are added when creating the pseudo-proxy record, and the locations of the pseudo-proxy record. One would expect the reconstruction to differ as a result of sampling variability from at least these two sources. Therefore, to get a better picture of the performance of each reconstruction method, it is necessary to repeat the reconstruction methods on a number of realizations of the pseudo-proxy record. Hence, 100 temperature series are reconstructed for each method using 100 different realizations of pseudo-proxy records. For each realization, the locations of the pseudo-proxies also change randomly within the 321 grid boxes that are specified before, as well as the noise. For the network with 100 proxy series, only 40 reconstructions are produced for each method due to computational limitations. However, 100 reconstructions were produced for each method for the smaller 15 locations proxy network. To provide a quantitative assessment for each reconstruction method, we computed the relative root mean squared error (RRMSE) of the reconstruction error, expressed relative to the variability of the model simulated NH temperature during the pre-calibration period. The RRMSE is simply defined as

$$\text{RRMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\mathbf{T}_t - \hat{\mathbf{T}}_t)^2}{\sum_{t=1}^{n}(\mathbf{T}_t - \overline{\mathbf{T}})^2}}$$

where $\mathbf{T}_t$, $\hat{\mathbf{T}}_t$ and $\overline{\mathbf{T}}$ are the model simulated NH temperature, the reconstructed NH temperature and the temporal mean of the model simulated NH temperature during the pre-calibration period respectively. In general, a smaller RRMSE means a better reconstruction. The RRMSE can lie between zero and infinity and a RRMSE value of less than 1 indicates that the reconstructed series is better than a reconstruction that has a constant value equal

to the climatology of the pre-calibration period.

Figure 4.3 shows the median of the RRMSEs obtained from the 100 (or 40) realizations, as a function of the degree of smoothing of the climate model simulated annual mean series $\mathbf{T}_t$ and the reconstructed annual mean series $\hat{\mathbf{T}}_t$ at which the RRMSE is calculated. An estimated 5-95% uncertainty range of the RRMSE is also displayed in the figure, which is obtained by using the 5^{th} and 95^{th} percentiles of the sample of RRMSEs. Comparing the results obtained between the two simulations, the results are robust for most of the methods. The performance of most of the methods considered is very similar at decadal and lower resolution. The non-smoothed forward regression, non-smoothed variance matching and MBH methods are exceptions. The performance of all methods was found to be insensitive to the two choices of calibration period, and thus results obtained using the shorter calibration periods are not shown. In fact, for both the CSM and ECHO-G simulation, there is almost no change in the median value of RRMSE when the shorter calibration period is used.

At the annual resolution, the RegEM and state-space model approaches produce the smallest RRMSEs. This is a result of the more sophisticated procedures that are involved in these methods, which in effect filter out the measurement errors in the pseudo-proxy series. In contrast, the reconstructed series from a typical CPS method is merely a scaled version of the composite series, with the result that the measurement error contained in the composite series is directly transferred to the reconstructed series. This problem is more severe when only 15 proxy series are available for reconstruction (Figure 4.4). Hence, the simple CPS methods should be avoided if the goal is to reconstruct high frequency climate variation. At the same time, the MBH method is observed to be the worst performer when SNR=0.5. However, when SNR is increased to 1, the MBH method, at annual resolution, produces comparable RRMSEs to that of most CPS methods considered.

The estimated RRMSE uncertainty range provides information on the sensitivity to sampling variability for each method. From Figure 4.3, it is obvious that such sensitivity is larger when only 15 pseudo-proxy series are used, reflecting the greater sampling variability in the composite record when only 15 proxy series are used. Inter-comparison between the different methods suggests that the sensitivity to sampling variability at annual resolution is somewhat

smaller for the RegEM and state-space model approaches. At decadal or lower resolution, the RRMSE uncertainty range is very similar across most methods.

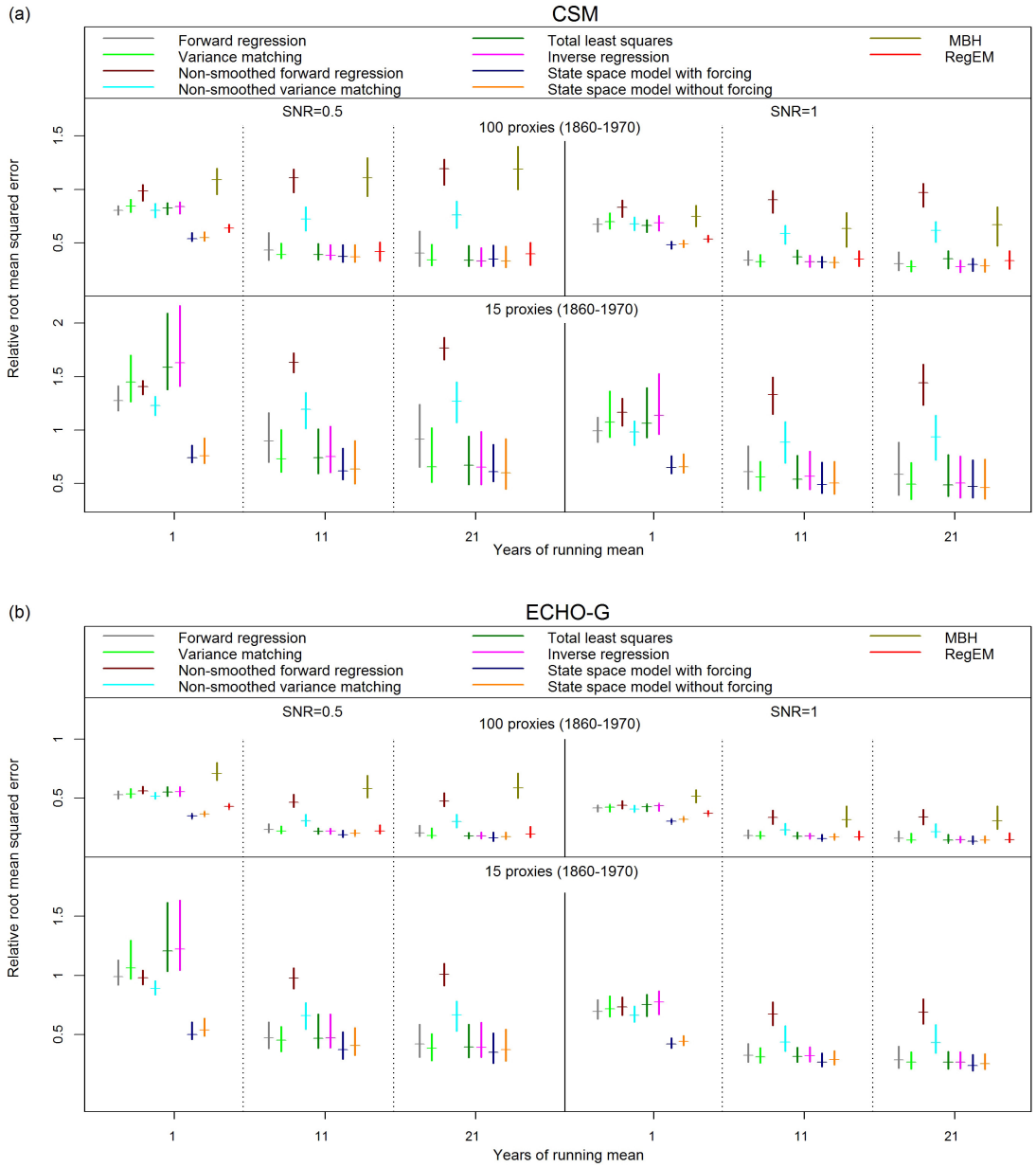

Figure 4.3: Relative root mean squared error (RRMSE) of the reconstruction error, expressed relative to the variability of the simulated hemispheric temperature. Units (degrees Kelvin). The median RRMSE is indicated with horizontal bars and the estimated 5-95% range of the RRMSEs is shown with vertical lines. Results using the 1860-1970 calibration period with different signal-to-noise ratios and varying number of pseudo-proxies are shown for the two climate model simulations: a) CSM and b) ECHO-G.

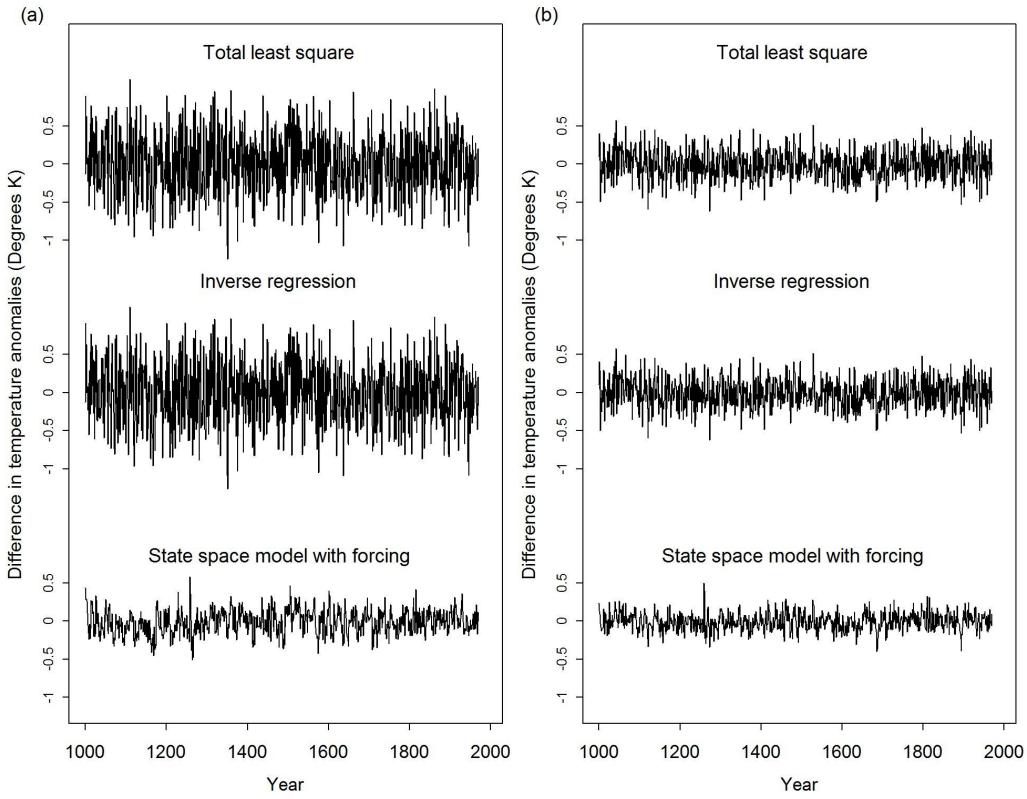

Figure 4.4: Example of the difference in temperature anomalies between the ECHO-G simulation and the reconstructed ECHO-G series using SNR=0.5 and the 1860-1970 calibration period, plotted at annual resolution with a) 15 pseudo-proxy series and b) 100 pseudo-proxy series.

By comparing the results of the two state-space model approaches (with external forcing response and without), it is clear that the RRMSE is insensitive to the inclusion or exclusion of the EBM estimated forcing response information. Nevertheless, the reconstructed series from the two approaches are slightly different when only 15 pseudo-proxy series are used (Figure 4.2a). This suggests that the Kalman filter and smoother algorithm relies more heavily on the proxy series than the estimated response to forcings to estimate the unknown temperature. Even though the reconstructions from the two state-space model approaches are very similar, the approach that takes forcing changes into account may be more useful in some instances since it may be possible to use it to provide a detection assessment and perhaps also a projection of future climate.

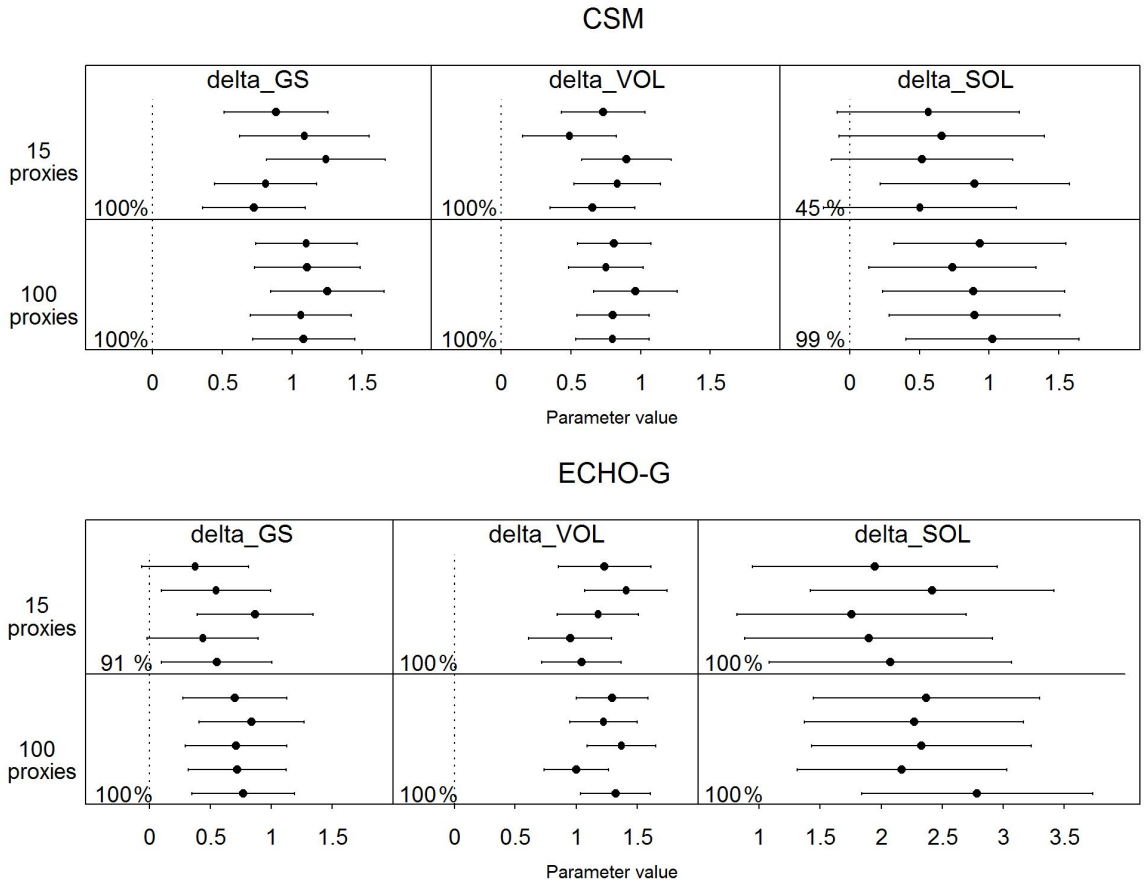

Figure 4.5: 95% confidence bounds for the coefficients used to scale the EBM simulated responses to external forcing in the state-space model when the pseudo-proxy SNR is 0.5 and using the 1860-1970 calibration period. Only results from 5 out of the 100 (or 40) experiments are displayed. The number in the bottom left corner of each box indicates the percentage of confidence bounds (out of 100 or 40) that excludes zero.

As mentioned above, one can use the state-space model to simultaneously reconstruct the NH temperature and conduct detection analysis. Both the CSM and ECHO-G simulations are forced with a combination of external forcing factors and therefore one should be able to detect the effect of external forcing in the experiments provided that these models respond similarly to forcing as the EBM. Figure 4.5 shows the confidence bounds on the forcing response coefficients for the experiments using SNR=0.5. For $\delta_{\mathbf{GS}}$ and $\delta_{\mathbf{VOL}}$, the confidence bounds do not include zero for almost all experiments that were run, indicating that the EBM simulated GS and VOL

signals are detectable in the reconstruction of the CSM and ECHO-G simulations. However, the response to solar forcing as simulated by the EBM is not detectable in about half of the CSM reconstructions obtained using 15 pseudo-proxy series. This fraction is reduced to near zero when the SNR is increased to 1. In contrast, the EBM simulated SOL signal is detectable in all the CSM experiments that use 100 pseudo-proxy series and in all ECHO-G experiments. The inability to detect the SOL forcing in some experiments may be due to the fact that the climate response to solar forcing is relatively weaker than that to the other forcings and hence may be harder to detect when the noise contamination in the pseudo-proxy series increases. At the same time, unlike the ECHO-G and EBM simulations, the solar forcing estimates used in the CSM simulation excluded the 11-yr solar cycle and this may also contribute to the varying detection results for the response to solar forcing. The inability to consistently detect the response to SOL forcing in our experiments is consistent with detection work on real-world paleo-reconstructions (Hegerl et al. 2003, 2007a).

Figure 4.6 displays hindcasts of annual NH mean temperature for 1971 to 1990 with 100 pseudo-proxy series, SNR=0.5 and the 1860-1970 calibration period. The hindcasts are produced using (4.2.2). For comparison, the sum of the responses to external forcings for the average of 30N-90N as simulated by the EBM is also displayed in the figure. Two hindcasts were produced for each climate model using the same set of pseudo-proxy series, one using the full 1000-1970 period to estimate the parameters for the state-space model and another using only data from 1800 to 1970. It can be observed that the skill of the hindcast varies between the two analysis periods. In fact, the estimates of the forcing response coefficients, which are influential to the hindcast values, are substantially different for the two analysis periods (Table 4.1). However, these differences have only a minor influence on the reconstructed series (not shown) because the Kalman smoother algorithm relies more heavily on the proxy data than the EBM to reconstruct the unknown temperature.

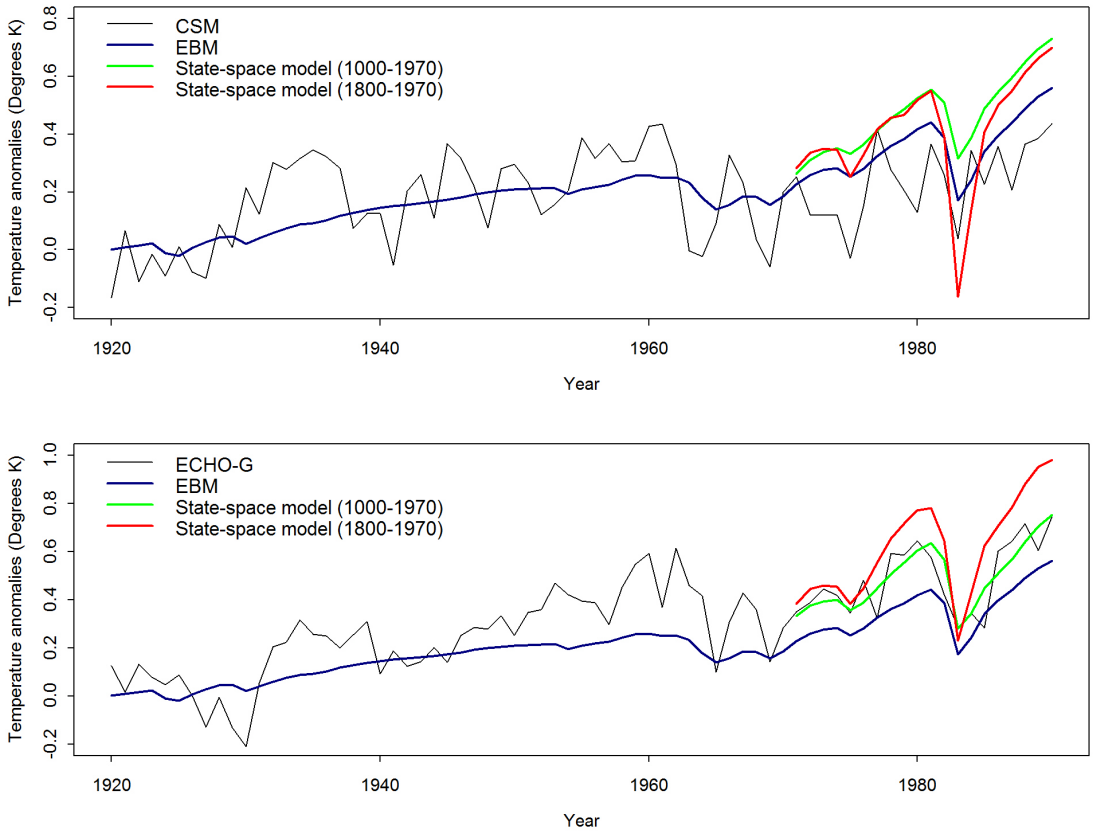

Figure 4.6: Hindcasts of annual mean NH temperature based on the estimated state equation from 100 proxy series for two analysis periods: 1000-1970 and 1800-1970. The sum of response to external forcings for the 30N-90N average as simulated by the energy balance model is also displayed for comparison purposes. All series are expressed as anomalies relative to the 1860-1970 period.

	CSM		ECHO-G	
Parameter	1000-1970	1800-1970	1000-1970	1800-1970
δ_{GS}	**1.315**	**2.105**	**0.775**	**1.978**
δ_{VOL}	**0.994**	**2.334**	**1.270**	**1.875**
δ_{SOL}	**0.891**	-0.298	**2.779**	**6.456**

Table 4.1: Estimated parameter values of the state-space model obtained with 100 pseudo-proxy series using two analysis periods. Boldfaced values are significant.

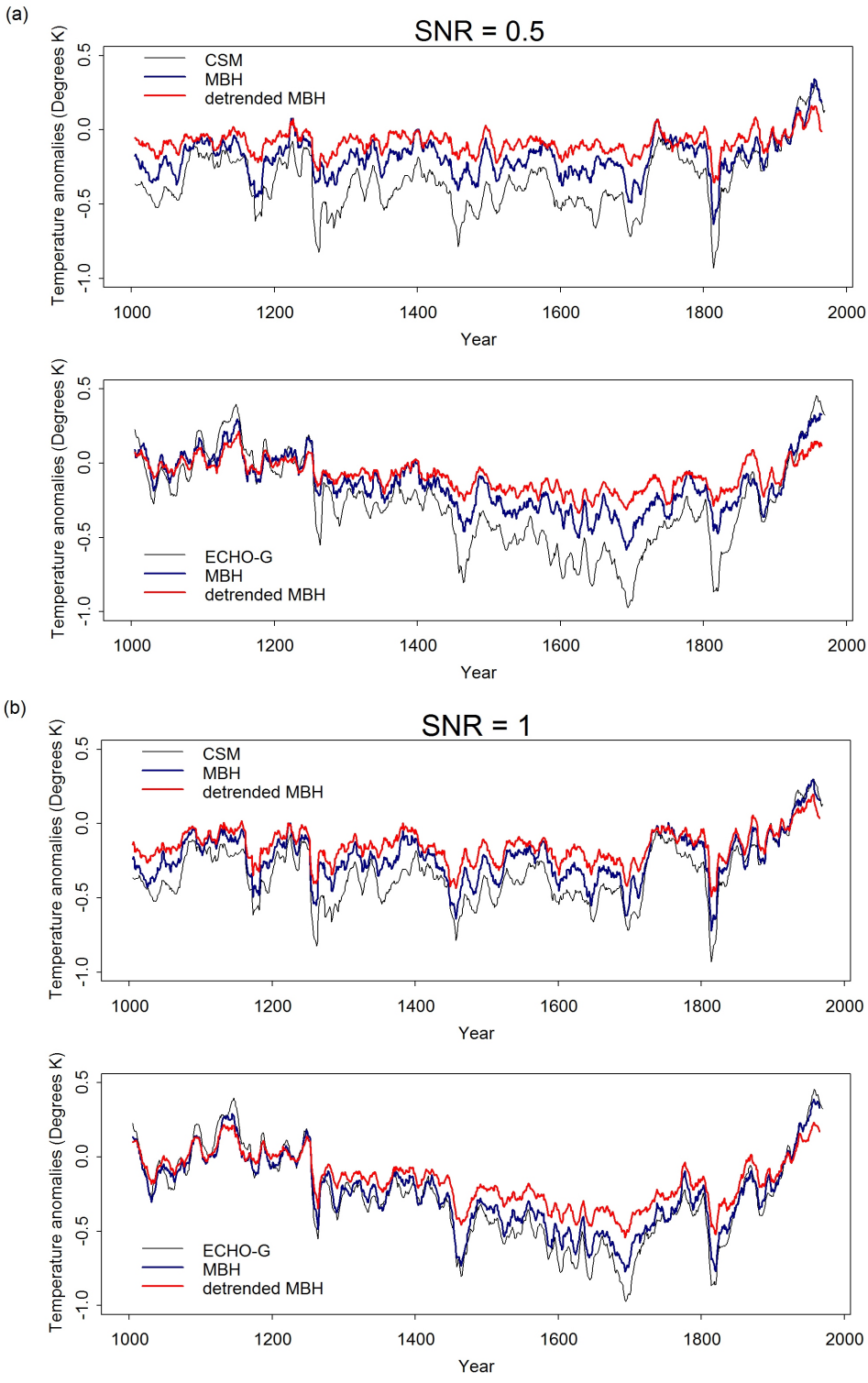

Figure 4.7: Comparison of the reconstructed hemispheric mean temperature series obtained using the MBH method with non-detrended or detrended data at two signal-to-noise ratios a) SNR=0.5 and b) SNR=1 with 100 pseudo-proxy series and the 1860-1970 calibration period. All series are expressed as 11-year running means and as anomalies relative to the calibration period.

Figure 4.7 displays examples of reconstructed NH temperature from the MBH method using the two different simulations and different SNR with the 1860-1970 calibration period. For comparisons, experiments were repeated with detrended calibration data and those results are also shown in Figure 4.7. When the variance of the noise added in the pseudo-proxy series is the same as the grid-box temperature variance, that is, when SNR=1, the non-detrended MBH method was able to provide a reasonable reconstruction of the ECHO-G simulated hemispheric mean temperature. However, results become unsatisfactory when the SNR is decreased to 0.5. For the reconstruction of CSM hemispheric mean temperature, the result is poor with both SNRs. Although the calibration period used here is longer than that used in Mann et al. (1998), the result does not change substantially when the shorter 1880-1960 calibration period is used (not shown). While the result suggests the MBH method underestimates long term variability, the under-estimation is smaller when non-detrended data is used (see Wahl et al., 2006; von Storch et al., 2006 for further discussions).

The CPS methods mentioned in the previous section were also tested using detrended calibration data (not shown). The performance of the CPS methods varies across different realizations of pseudo-proxy records. In some cases, detrending does not affect the reconstructed series irrespective of which CPS method is considered. On the other hand, there were also realizations of pseudo proxies for which there was under-estimation of variability when detrended data are used. Therefore, in the context of pseudo-proxies constructed with white noise, detrending results in less robust reconstructions of hemispheric mean temperature variability.

4.3.1 Analysis with red pseudo-proxy noise

Up to this point, experiments have not taken into account the possibility that the proxies consist of a temperature signal plus correlated errors. Therefore, the impact of red pseudo-proxy noise on the reconstruction methods is now examined. Following Mann et al. (2007), red pseudo-proxy noise is generated from an AR(1) process with lag-one autocorrelation equal to 0.32. As before, analyses are conducted using two calibration periods and with SNR=0.5 and 1.

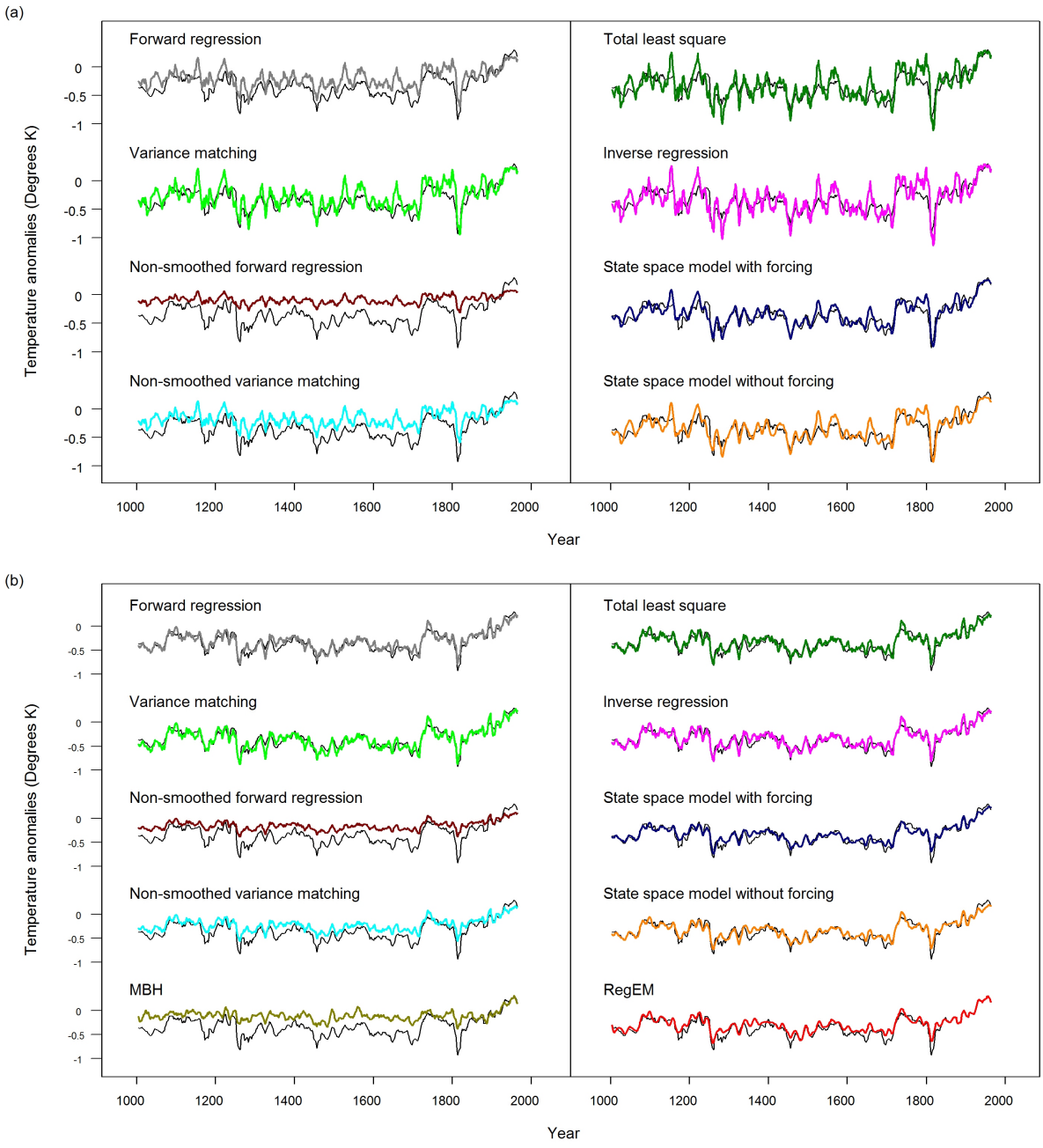

Figure 4.8: Examples of reconstructed CSM northern hemisphere mean temperature series. Experiments were run using SNR=0.5 and the 1860-1970 calibration period with a) 15 and b) 100 red pseudo-proxy series. 11-yr moving averages are shown. The CSM NH mean is shown in black. Reconstructions are shown as coloured lines. All series are express as anomalies relative to the calibration period. Units (degrees Kelvin).

Examples of reconstructed NH temperature evolution simulated by the CSM based on red pseudo-proxy series with SNR=0.5 are displayed in Figure 4.8. It is clear that the redness of the pseudo-proxy noise has slightly increased the variability of the reconstructed series when 15 pseudo-proxy series are used. On the other hand, series reconstructed with 100 pseudo-proxy series do not seem to be affected. Results obtained with the other calibration period and with the ECHO-G simulation are very similar and thus not shown.

The estimated RRMSEs obtained with red pseudo-proxy series (not shown) are very similar to those shown in Figure 4.3. In particular, the estimated median RRMSEs are almost unchanged and the relative ranking of the RRMSEs between the different reconstruction methods remains the same as in the case of white pseudo-proxy series. Hence, similar conclusions regarding the performance of the different methods can be drawn as before. However, the RRMSE uncertainty ranges are slightly larger than before when 15 pseudo-proxy series are used.

4.4 Comparison using real-world paleoclimate proxy data

The CPS methods and state-space model approach is now applied to the paleoclimate proxy data used in Hegerl et al. (2007a). This data set consists of 14 proxy series. All records are available as decadally smoothed series. As in Hegerl et al. (2007a), correlation based weighted averages were calculated to form the composite record. Using 1880-1960 as the calibration period, the 30N-90N mean temperature is reconstructed for the period 1510-1960. As noted previously, the use of the state-space model approach requires fixing the parameter ϕ in the exogenous variable $\mathbf{F}_t$. Here, the value of this parameter is estimated by the lag-one autocorrelation of the decadally smoothed 30N-90N mean temperature from a control simulation of the CCCma CGCM2 (Flato and Boer, 2001). The resulting ϕ value, 0.982, was used to obtain the exogenous variable $\mathbf{F}_t$.

Figure 4.9 compares the reconstructed series obtained from the variance matching method, state-space model approach and Hegerl et al. (2007a), who used the total least squares method. The reconstruction estimates obtained from these approaches are nearly identical. Reconstructed series from the other CPS approaches also agree closely with the series in Figure 4.9

(not shown). This finding is consistent with results obtained using climate model simulations in the previous section where it was seen that these methods have similar RRMSEs at the decadal resolution.

Estimates of the parameters of the state-space model (with forcing) are given in Table 4.2. The parameter ϕ is estimated to be close to 1, which is larger than that obtained using climate model simulations with annual mean data, which ranges from 0.3 to 0.7. This is not surprising given that decadally smoothed data is strongly dependent between successive time points. The estimate parameter value for $\delta_{\mathbf{GS}}$ and $\delta_{\mathbf{VOL}}$ is significantly different from zero, suggesting that the response to GS and VOL forcing are detectable. On the other hand, the response to SOL forcing is not detected. These detection results agree with the findings reported in Hegerl et al. (2007a).

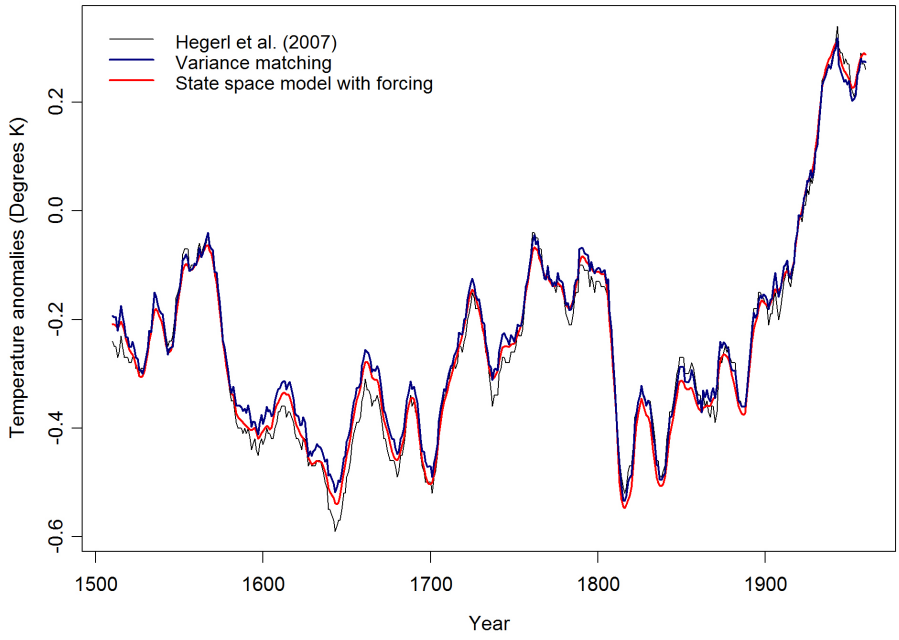

Figure 4.9: Reconstructed series for the 30N-90N mean using the variance matching method and state-space model approach for real-world paleoclimate proxy data. Temperature series are expressed as 11-yr moving averages. All series are expressed as anomalies relative to the 1880-1960 period and in units of degrees Kelvin.

Parameter	1270-1960	1510-1960	1800-1960
ϕ	**0.981** (0.967, 0.994)	**0.979** (0.961, 0.997)	**0.944** (0.900, 0.988)
δ_{GS}	**1.022** (0.043, 2.001)	**1.128** (0.112, 2.143)	**4.271** (1.054, 7.489)
δ_{VOL}	**0.953** (0.685, 1.220)	**0.814** (0.522, 1.106)	**0.998** (0.461, 1.535)
δ_{SOL}	0.784 (-0.592, 2.161)	0.368 (-1.218, 1.955)	-0.203 (-3.211, 2.805)

Table 4.2: Estimated parameter values of the state-space model obtained with paleoclimate proxy data for three reconstruction periods. The 95% confidence interval is listed in brackets. Boldfaced values are significant.

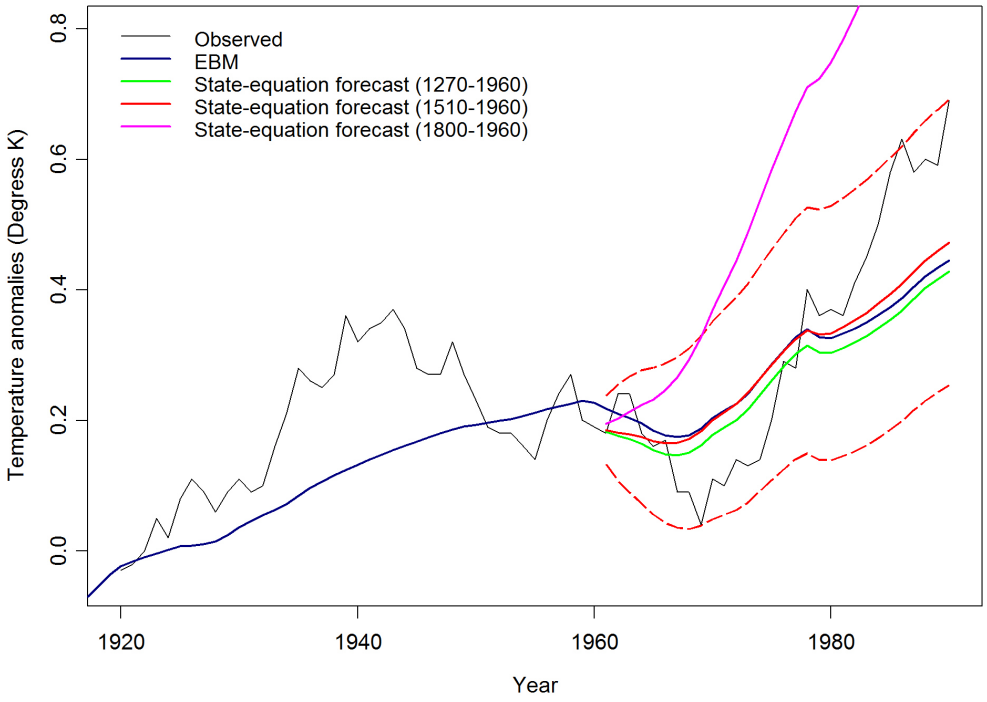

Figure 4.10: State equation hindcast of decadally smoothed 30N-90N mean temperature with real-world paleoclimate proxy data. All hindcasts are based on the estimated state equation for three analysis periods: 1270-1960, 1510-1960 and 1800-1960. Nine proxy series are available for the 1270-1960 analysis period and 14 proxy series are available for the other two analysis periods. 95% confidence bounds of the hindcasts for the 1510-1960 analysis period are shown as dashed lines. All series are expressed as anomalies relative to the 1880-1960 period and in units of degrees Kelvin.

The 30N-90N decadally smoothed mean temperature hindcast for 1961-1990 is displayed in Figure 4.10. Hindcasts are generated for three analysis periods (1270-1960, 1510-1960 and 1800-1960). Confidence bounds on the hindcast for the 1510-1960 analysis period are also displayed in the figure. The hindcasts obtained from the 1270-1960 and 1510-1960 analysis periods are very similar and is very close to the sum of the response to external forcings as simulated by the EBM. The confidence bounds on the hindcasts generated for the 1510-1960 analysis were able to include almost all the observed temperature anomalies. Similar results can be obtained for the 1270-1960 analysis period. On the other hand, the hindcasts obtained from the 1800-1960 analysis period warm too quickly. This is because the estimated value of $\delta_{\mathbf{GS}}$ is four times larger than that in the other two analysis periods (Table 4.2).

4.5 Concluding remarks

In this chapter, the skill of several different reconstruction methods were compared using climate model simulations. At the annual resolution, the state-space model and RegEM approaches provide the best reconstructions. On the other hand, when compared at decadal or lower resolution, most methods can provide satisfactory and similar results. Exceptions are the MBH, non-smoothed forward regression and non-smoothed variance matching methods. When analyzed with decadally smoothed real-world paleoclimate proxy data, all the CPS methods provide almost identical results. The similarity in performance provides evidence that the difference between many real-world reconstructions is more likely to be due to the choice of the proxy series, or the use of difference target seasons or latitudes than to the choice of statistical reconstruction method (see also Juckes et al., 2006).

The past two chapters have put forward another approach to historical temperature reconstruction that is based on a state-space time series model and the Kalman filter and smoother algorithm. This approach allows the possibility of incorporating additional non-proxy information into the reconstruction analysis, such as the estimated response to external forcing. However, experiments show that the state-space model approach does not produce substantially different reconstructions when such information is included. Nevertheless, both state-space model approaches provided better reconstructions than existing CPS methods at

annual resolutions. At the same time, including forcing response terms in the state-space model allows one to carry out a simultaneous reconstruction and detection analysis. It can also be used to provide forecasts of future climates. Consistent with the results of Hegerl et al. (2007a), the effects of anthropogenic forcing (greenhouse gas and aerosol) and volcanic forcing is detected in real-world paleoclimate proxy data.

Extending the state-space model approach to reconstruct both the underlying temporal and spatial patterns of historical temperature using multiple proxy series is feasible by defining $\mathbf{T}_t$ as a vector that represents the spatial patterns of temperature and $\mathbf{P}_t$ as a vector containing the multiple proxy records. In this case, the Kalman filter and smoother algorithm can continue to be used to estimate the unknown temperature. However, due to high dimensionality of the problem, modifications on the maximum likelihood estimation methods proposed in Chapter 3 might be required to give estimates of the unknown parameters.

It would also be useful to account for uncertainty more completely in the climate reconstructions that are produced with the state-space model approach. Confidence bounds for the Kalman filter and smoother estimates of the unknown temperature can be derived using the fact that these estimates are normally distributed (see section 3.1.1). The variance of the Kalman filter and smoother estimates at time t are $\mathbf{S}_{t|t}$ and $\mathbf{S}_{t|N}$ respectively (see Property 3.1 and 3.2). However, these variances only account for uncertainty resulting from internal variability of the climate system and non-temperature variability in the composite climate proxy. They do not take into account the uncertainty resulting from variability in the state-space model parameter estimates. Nevertheless, Monte Carlo simulation can possibly provide a way to take parameter uncertainty into account when constructing confidence bounds for the temperature estimates. This will require first generating the state-space model parameters from the asymptotic distribution of the maximum likelihood estimates obtained using the proxy record and observed temperature. Then, the Kalman filter and smoother estimates are calculated using the generated parameter value. By repeating the procedure a large number of times, the variability of the Kalman filter and smoother estimates in the resulting sample could possibly be used to reflect the variability due to uncertainty in parameter estimates.

5 References

Allen, M.R. and S.F.B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, **15**, 419-434.

Allen, M.R. and P.A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, Part I: Theory. *Climate Dynamics*, **21**, 477-491.

Ammann, C.M., F. Joos, D. Schimel, B.L. Otto-Bliesner and R. Tomas, 2007: Solar influence on climate during the past millennium: results from transient simulations with the NCAR Climate System Model. Proceedings of the National Academy of Sciences of the United States of America, **104**, 3713-3718.

Berliner, L.M., R.A. Levine and D.J. Shea, 2000: Bayesian climate change assessment. *Journal of Climate*, **13**, 3805-3820.

Bhat, B.R., 1974: On the method of maximum likelihood for dependent observations. *Journal of Royal Statistical Society B*, **36**, 48-53.

Billingsley, P., 1961: *Statistical inference for Markov processes.* University of Chicago Press, 75pp.

Boer, G.B., 2004: Long time-scale potential predictability in an ensemble of climate climate models. *Climate Dynamics*, **23**, 29-44.

Braganza, K., D.J. Karoly, A.C. Hirst, P. Stott, R.J. Stouffer and S.F.B. Tett, 2004: Simple indices of global climate variability and change: Part II: Attribution of climate change during the 20th century. *Climate Dynamics*, **22**, doi:10.1007/S00382-004- 0413-1.

Brier, G.W., 1950: Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, **78**, 1-3.

Briffa, K.R., T. J. Osborn, F. H. Schweingruber, I. C. Harris, P. D. Jones, S. G. Shiyatov and E. A. Vaganov, 2001: Low-frequency temperature variations from a northern tree-ring density network. *Journal of Geophysical Research*, **106**, 2929-2941.

Brockwell P.J. and R.A. Davis, 1991: *Time series: theory and methods.* Springer-Verlag, New York, 577pp.

Brohan, P., J.J. Kennedy, I. Haris, S.F.B. Tett and P.D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *Journal of Geophysical Research*, **111**, D12106.

Brown, B.M., 1971: Martingale central limit theorems. *Annals of Mathematical Statistics*, **42**, 59-66.

Brown, P.J., 1993: *Measurement, regression, and calibration.* Oxford University Press, 201pp.

Bürger G. and U. Cubasch, 2005: Are multiproxy climate reconstructions robust? *Geophysical Research Letters*, **32**, L23711.

Bürger G., I. Fast and U. Cubasch, 2006: Climate reconstruction by regression - 32 variations on a theme. *Tellus*, **58A**, 227-235.

Caines, P.E. 1988: *Linear Stochastic Systems*. Wiley, 847pp.

Chow, Y.S. and H. Teicher, 1978: *Probability theory: independence, interchangeability, martingales*. Springer-Verlag, New York, 455pp.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, J. J. Doblas-Reyes and D. Stephenson, 2004: Forecast calibration and combination: A simple Bayesian approach for ENSO. *Journal of Climate*, **17**, 1504-1516.

Collins, M. and M.R. Allen, 2002: Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *Journal of Climate*, **15**, 3104-3109.

Cramér, H., 1946: *Mathematical methods of statistics*. Princeton University Press.

D'Arrigo, R., E.R. Cook, R.J. Wilson, R. Allan and M.E. Mann, 2005: On the variability of ENSO over the past six centuries. *Geophysical Research Letters*, **32**, L03711.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **B39**, 1-38.

Derome, J., G. Brunet, A. Plante, N. Gagnon, G.J. Boer, F.W. Zwiers, S.J. Lambert, J. Sheng and H. Ritchie, 2001: Seasonal predictions based on two dynamical models. *Atmosphere-Ocean*, **39**, 485-501.

Dethlefsen, C., B. Klein and H. Thomsen, 1997: *State space models and Kalman filtering in the global positioning system*, Master thesis, Aalborg University, Denmark, 221pp. Available from http://www.math.aau.dk/∼dethlef/pub/master.pdf.

Durbin, J., and S.J. Koopman, 2001: *Time series analysis by state space methods*. Oxford University Press, 253pp.

Esper, J., E.R. Cook and F.H. Schweinngruber, 2002: Low-frequency in long treering chronologies for reconstruction past temperature variability. *Science*, **295**, 2250-2253.

Esper, J., D.C. Frank, R.J. Wilson and K.R. Briffa, 2005: Effect of scaling and regression on reconstructed temperature amplitude for the past millennium. *Geophysical Research Letters*, **32**, L07711.

Fierro, R. D., G. H. Golub, P. C. Hansen and D. P. O'Leary, 1997: Regularization by truncated total least squares. *SIAM Journal of Scientific Computing*, **18**, 1223-1241.

Flato, G.M. and G.J. Boer, 2001: Warming asymmetry in climate change simulations. *Geophysical Research Letter* **28**, 195-198.

Fuller, W.A., 1987: *Measurement error models*. Wiley, 440 pp.

Gillett, N.P., A.J. Weaver, F.W. Zwiers and M.D. Flannigan, 2004a: Detecting the effect of human induced climate change on Canadian forest fires. *Geophysical Research Letters*, **31**, L18211, doi:10.1029/2004GL020876.

Gillett, N.P., M.F. Wehner, S.F.B. Tett and A.J. Weaver, 2004b: Testing the linearity of the response to combined greenhouse gas and sulfate aerosol forcing. *Geophysical Research Letters*, **31**, L14201.

Gillett, N.P., F.W. Zwiers, A.J. Weaver, G.C. Hegerl, M.R. Allen and P.A. Stott, 2002: Detecting anthropogenic influence with a multi-model ensemble. *Geophysical Research Letters*, **29**, 1970.

Grotzner, A., M. Latif, A. Timmermann and R. Voss, 1999: Interannual to decadal predictability in a coupled ocean-atmosphere general circulation model. *Journal of Climate*, **12**, 2607-2624.

Hall, P. and C.C. Heyde, 1980: *Martingale limit theory and its application.* Academic Press, New York, 308pp.

Harvey, A., 1989: *Forecasting, structural time series models and the Kalman filter.* Cambridge University Press, 554pp.

Hegerl, G. C., T.J. Crowley, S.K. Baum, K.-Y. Kim, and W.T. Hyde, 2003: Detection of volcanic, solar, and greenhouse gas signals in paleo-reconstructions of Northern Hemispheric temperature. *Geophysical Research Letters*, **30**, 1242, doi:10.1029/ 2002GL016635.

Hegerl, G.C. , T.J. Crowley, M.R. Allen , W.T. Hyde, H.N. Pollack, J.E. Smerdon and E. Zorita, 2007a: Detection of human influence on a new, validated 1500 year temperature reconstruction. *Journal of Climate*, **20**, 650-666.

Hegerl, G.C., F.W. Zwiers, P. Braconnot, N.P. Gillett, Y. Luo, J.A. Marengo Orsini, N. Nicholls, J.E. Penner and P.A. Stott, 2007b: Understanding and Attributing Climate Change. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IDAG (International Ad-hoc Detection Group), 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *Journal of Climate*, **18**, 1291-1314.

IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

Jansen, E., J. Overpeck, K.R. Briffa, J.-C. Duplessy, F. Joos, V. Masson-Delmotte, D. Olago, B. Otto-Bliesner, W.R. Peltier, S. Rahmstorf, R. Ramesh, D. Raynaud, D. Rind, O. Solomina, R. Villalba and D. Zhang, 2007: Palaeoclimate. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jensen, J.L. and N.V. Petersen, 1999: Asymptotic normality of the maximum likelihood estimator in state space models. *The Annals of Statistics*, **27**, 514-535.

Jones, P.D., T.S. Osborn and K.R. Briffa, 1997: Estimating sampling errors in large-Scale temperature averages. *Journal of Climate*, **10**, 2548-2568.

Jones, P.D., K.R. Briffa, T.P. Barnett and S.F.B. Tett, 1998: High-resolution palaeoclimatic records for the last millennium: Integration, interpretation and comparison with General Circulation Model control run temperatures. *Holocene*, **8**, 455-471.

Jones, P.D., T.J. Osborn, K.R. Briffa, C.K. Folland, E.B. Horton, L.V. Alexander, D.E. Parker and N.A. Rayner, 2001: Adjusting for sampling density in grid box land and ocean surface temperature time series. *Journal of Geophysical Research*, **106**, 3371-3380.

Juckes M.N., M.R. Allen, K.R. Briffa, J. Esper, G.C. Hegerl, A. Moberg, T.J. Osborn, S.L. Weber and E. Zorita, 2006: Millennial temperature reconstruction intercomparison and evaluation. *Climate of the Past Discussions*, **2**, 1001-1049.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**, 34-45.

Karoly, D.J. and Q. Wu, 2005: Detection of regional surface temperature trends. *Journal of Climate*, **18**, 4337-4343.

Kharin, V.V. and F. Zwiers. 2002: Climate Predictions with Multimodel Ensembles. *Journal of Climate*, **15**, 793-799.

Lee, T.C.K., 2003: *Identifying anthropogenic causes of the observed twentieth century surface temperature change: frequentist and Bayesian approach*, Master thesis, University of Canada, Canada, 73 pp.

Lee, T.C.K., F. Zwiers, X. Zhang, G. Hegerl and M. Tsao, 2005: A Bayesian approach to climate change detection an attribution. *Journal of climate*, **18**, 2429-2440.

Lee, T.C.K., F. Zwiers, X. Zhang and M. Tsao, 2006: Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing. *Journal of climate*, **19**, 5305-5318.

Lee, T.C.K., F. Zwiers and M. Tsao, 2008: Evaluation of proxy-based millennial reconstruction methods. *Climate Dynamics*, **in press**.

Lehmann, E.L. and G. Casella, 2001: *Theory of point estimation*. Springer-Verlag, New York, 589 pp.

Mann, M.E., R.S. Bradley and M.K. Hughes, 1998: Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, **392**, 779-787.

Mann, M.E., R.S. Bradley and M.K. Hughes, 1999: Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophysical Research Letters*, **26**, 759-762.

Mann, M.E., and P.D. Jones, 2003: Global surface temperatures over the past two millennia. *Geophysical Research Letters*, **30**, 1820.

Mann, M.E., S. Rutherford, E. Wahl and C.M. Ammann, 2005: Testing the fidelity of methods used in proxy-based reconstructions of past climate. *Journal of Climate*, **18**, 4097-4107.

Mann, M.E., S. Rutherford, E. Wahl and C.M. Ammann, 2007: Robustness of Proxy-Based Climate Field Reconstruction Methods. *Journal of Geophysical Research*, **112**, D12109.

Mason, S.J., L. Goddard, N.E. Graham, E. Yulaeva, I. Sun and P.A. Arkin, 1999: The IRI seasonal climate prediction system and the 1997/98 El-Nino event. *Bulletin of the American Meteorological Society*, **80**, 1853-1873.

Min, S.-K., A. Hense, H. Paeth and W.-T. Kwon, 2004: A Bayesian decision method for climate change signal analysis. *Meteorologische Zeitschrift*, **13**, 421-436.

Mitchell, J.F.B., D.J. Karoly, G.C. Hegerl, F.W. Zwiers, M.R. Allen and J. Marengo, 2001: Detection of climate change and attribution of causes. Houghton J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell and C.A. Johnson (eds.), *Climate Change 2001: The Scientific Basis*. Cambridge University Press, Cambridge, United Kingdom and New York, 695-738.

Moberg, A., D.M. Sonechkin, K. Holmgren, N.M. Datsenko and W. Karlen, 2005: Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data. *Nature*, **433**, 613-617.

Nakicenovic, N., J. Alcamo, G. Davis, B. de Vries, J. Fenhann, S. Gaffin, K. Gregory, A. Grbler, T. Y. Jung, T. Kram, E. L. La Rovere, L. Michaelis, S. Mori, T. Morita,W. Pepper, H. Pitcher, L. Price, K. Raihi, A. Roehrl, H.-H. Rogner, A. Sankovski, M. Schlesinger, P. Shukla, S. Smith, R. Swart, S. van Rooijen, N. Victor and Z. Dadi, 2000: *IPCC Special Report on Emissions Scenarios*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 599 pp.

Osborn, T.J., S.C.B. Raper and K.R. Briffa, 2006: Simulated climate change during the last 1,000 years: comparing the ECHO-G general circulation model with the MAGICC simple climate model. *Climate Dynamics*, **27**, 185-197.

O'Lenic, E., 1994: *Operational long-lead forecast for the climate outlook.* Technical Procedures Bulletin 418. NOAA/NWS/CPC. 30 pp. [Available from NOAA/CPC, 5200 Auth Rd., Camp Springs, MD 20746.]

Pohlmann, H., M. Botzet, M. Latif, A. Roesch, M. Wild and P. Tschuck, 2004: Estimating the decadal predictability of a coupled AOGCM. *Journal of Climate*, **17**, 4463-4472.

Ramaswamy, V., O. Boucher, J. Haigh, D. Hauglustaine, J. Haywood, G. Myhre, T. Nakajima, G.Y. Shi and S. Solomon, 2001: Radiative forcing of climate change. Houghton J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell and C.A. Johnson (eds.), *Climate Change 2001: The Scientific Basis*. Cambridge University Press, Cambridge, United Kingdom and New York, 349-416.

Rao, C.R., 1973: *Linear statistical inference and its applications.* 2nd ed., Wiley, New York.

Rutherford, S., M.E. Mann, T. J. Osborn, R. S. Bradley, K. R. Briffa, M. K. Hughes and P. D. Jones, 2005: Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to methodology, predictor network, target season, and target domain. *Journal of Climate*, **18**, 2308-2329.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, **14**, 853-871.

Schweppe, F., 1965: Evaluation of likelihood functions for Gaussian signals. *IEEE Transaction on Information Theory*, **IT-4**, 294-305.

Schnur, R. and K. Hasselmann, 2005: Optimal filtering for Bayesian detection and attribution of climate change. *Climate Dynamics*, **24**, 45-55.

Serfling, R.J., 1980: *Approximation theorems of mathematical statistics.* Wiley, 371 pp.

Shumway, R.H. and D.S. Stoffer, 1982: An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, **3**, 253-264.

Shumway, R.H. and D.S. Stoffer, 2000: *Time series analysis and its applications.* Springer, 549 pp.

Somerville, R., H. Le Treut, U. Cubasch, Y. Ding, C. Mauritzen, A. Mokssit, T. Peterson and M. Prather, 2007: Historical Overview of Climate Change. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Stott, P.A., 2003: Attribution of regional-scale temperature changes to anthropogenic and natural causes. *Geophysical Research Letters*, **30**, 1728, doi:10.1029/2003GL017324.

von Storch, H., E. Zorita, J.M. Jones, Y. Dimitriev, F. Gonzalez-Rouco and S.F.B. Tett, 2004: Reconstructing past climate from noisy data. *Science*, **306**, 679-682.

von Storch, H., E. Zorita, J.M. Jones, F. Gonzalez-Rouco and S.F.B. Tett, 2006: Response to Comment on "Reconstructing past climate from noisy data". *Science*, **312**, 529c.

Wahl, E.R., D.M. Ritson and C.M. Ammann, 2006: Comment on "Reconstructing past climate from noisy data". *Science*, **312**, 529b.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Wu, C.F., 1983: On the convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95-103.

Zhang, X., F.W. Zwiers and P.A. Stott, 2005; Multi-model multi-signal climate change detection at regional scale. *Journal of Climate*, **19**, 4294-4307.

Zhang, X., F.W. Zwiers, G. Hegerl, F.H. Lambert, N.P. Gillett, S. Solomon, P.A. Stott and T. Nozawa, 2007: Detection of human influence on twentieth-century precipitation trends. *Nature*, **448**, 461-465.

Zorita, E., J. F. Gonzalez-Rouco and S. Legutke, 2003: Testing the Mann et al. (1998) approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model. *Journal of Climate*, **16**, 1378-1390.

Zorita, E. and H. von Storch, 2005: Methodical aspects of reconstructing non-local historical temperatures. *Memorie della Societa Astronomica Italia*, **76**, 794-801.

Zwiers, F.W., 2002: The 20-year forecast. *Nature*, **416**, 690-691.

Zwiers, F.W. and X. Zhang, 2003: Towards regional scale climate change detection. *Journal of Climate*, **16**, 793-797.