Assignment 4
Multivariate Processes

Christian Glissov, s146996

May 7, 2018

DTU Compute - Institute for Mathematics and Computer science
Time Series Analysis 02417 May 7, 2018
Supervisor: Lasse Engbo Christiansen
Code made in collaboration with Frederik Boe Hüttel, s151689 and Thomas Ørkild, s154433

Technical University of Denmark

Contents

Question 4.1: Presenting the data

The following dataset has been given:

- Year: Year of observation.
- sh: Temperature anomaly of southern hemisphere in Celsius.
- nh: Temperature anomaly of northern hemisphere in Celsius.

The temperatures are estimated based on a number of measurement stations and do include measurement errors. Looking at figure (1) one can see that the two hemispheres seem to be somehow correlated. Taking the correlation of the two hemispheres shows that there is a heavy correlation between them $cor(X_{sh}, X_{nh}) = 0.909$. It can also be seen that the variance is not stationary, since both hemispheres seem to have a positive trend. The variance also seems to be heteroscedastic, with short sub-periods of increased variance or spikes. The heteroscedasticity might be due to measurement errors or extreme climate events, but one can only assume, due to lack of information about the dataset.

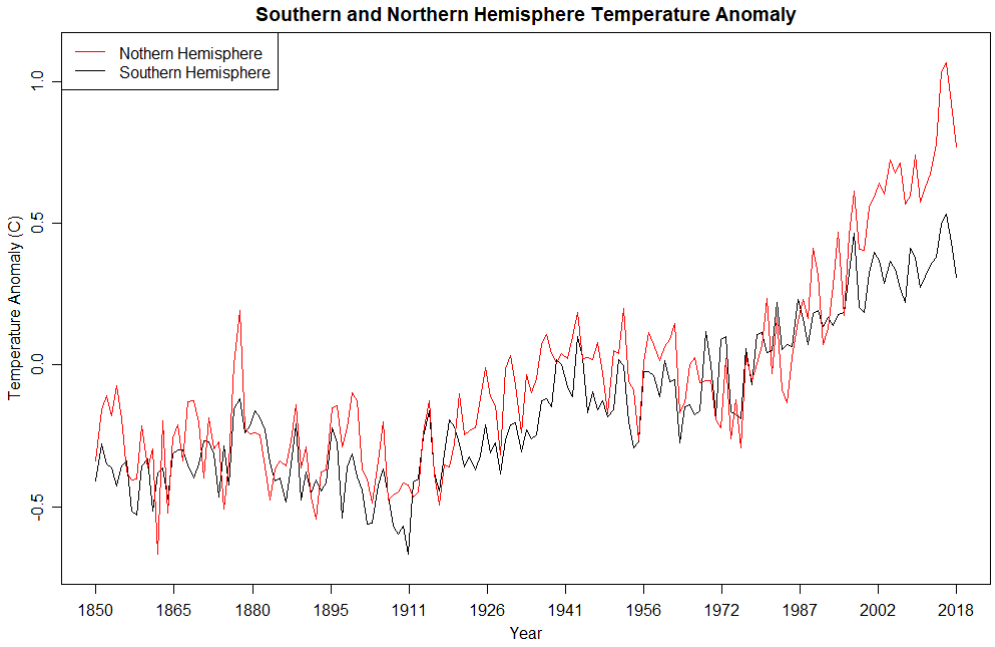


Figure 1: The two hemispheres and their temperature anomaly.

Looking at the box-plot in figure (2) it is seen that the northern hemisphere seems to have more outliers and to be slightly warmer.

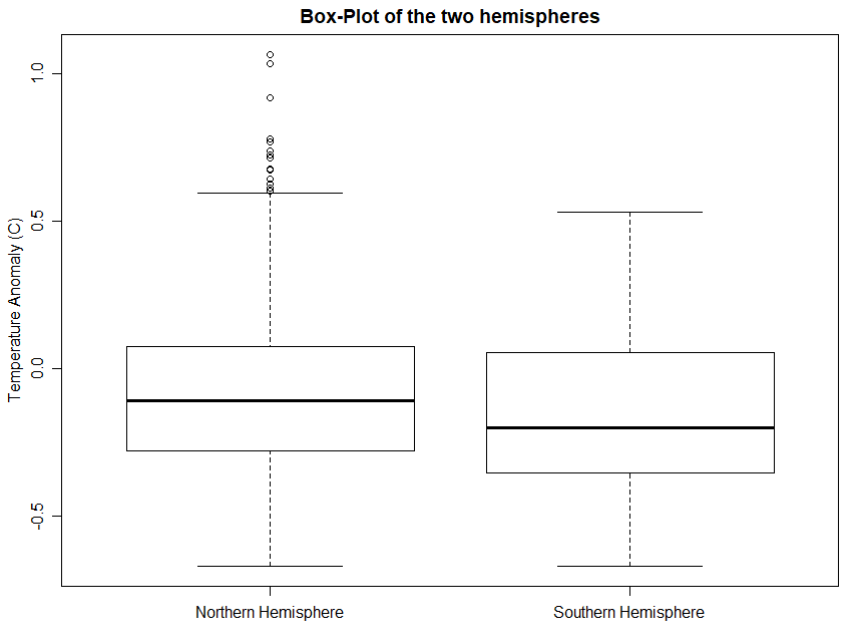


Figure 2: Box-plot of the two hemispheres.

Due to the outliers for the northern hemisphere, it should be expected that the variability is slightly higher, which is also reflected in the prediction intervals later.

Question 4.2: Formulating state space model

Using (10.1) and (10.2) in [1] to formulate a bi-variate state space model, first the system equation is defined. X_t must follow the Markov property. Therefore the latent stochastic state vector X_{t-1} is given by the two hemispheres X_{sh} and X_{nh}:

$$\mathbf{X}_{t-1} = \begin{bmatrix} X_{sh,t-1} \\ X_{nh,t-1} \end{bmatrix}$$

Now assuming the temperature anomalies for the two hemispheres follow independent random walks, the matrix of A_t is easily found by matching the dimensions 2×2:

$$\mathbf{A}_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

There is no input to the system, so the input vector is simply 0 in this case. For the observation equation the output of both hemispheres are wanted. Therefore C_t is defined as:

$$C_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The full bi-variate state space model is then:

$$\mathbf{X}_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_{sh,t-1} \\ X_{nh,t-1} \end{bmatrix} + \mathbf{e}_{1,t}$$
$$\mathbf{Y}_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_{sh,t} \\ X_{nh,t} \end{bmatrix} + \mathbf{e}_{2,t}$$

Where $\mathbf{e}_{2,t} \sim N(0, \Sigma_{2,t})$ and $\mathbf{e}_{1,t} \sim N(0, \Sigma_{1,t})$ are random noise vectors with the assumptions given in (10.3), (10.4), (10.5) and (10.6) in [1]. Where

$$\Sigma_1 = \begin{bmatrix} \sigma^2_{\mathbf{e}_1,11,t} & 0 \\ 0 & \sigma^2_{\mathbf{e}_1,22,t} \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} \sigma^2_{\mathbf{e}_2,11,t} & 0 \\ 0 & \sigma^2_{\mathbf{e}_2,22,t} \end{bmatrix}$$

These will be initialized in the next question.

Question 4.3: Kalman filtering

It's now assumed that the initial temperature anomaly is -0.4C for the southern and -0.3C northern hemisphere:

$$\mathbf{X}_0 = \begin{bmatrix} -0.4 \\ -0.3 \end{bmatrix}$$

The initial variance is 0.01 and so is the variance of the system and observation noise. To apply the Kalman filtering, a package in *R* is used, *FKF*. The inputs are then the same A and C matrix as in the previous question, the initial variance for the start guesses is:

$$P_0 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}$$

And the variance of the system and observation noise is also given as:

$$\Sigma_1 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}$$

After giving the inputs to the *FKF* function, the reconstruction of the Kalman filter for each hemisphere can be seen in figure (4) and (3) with their confidence intervals. The reconstruction is found by using eq. (10.73) and (10.74) with the

Kalman gain given in (10.60) in [1], this is continuously used for each Kalman filter, so it will be omitted in the next questions.

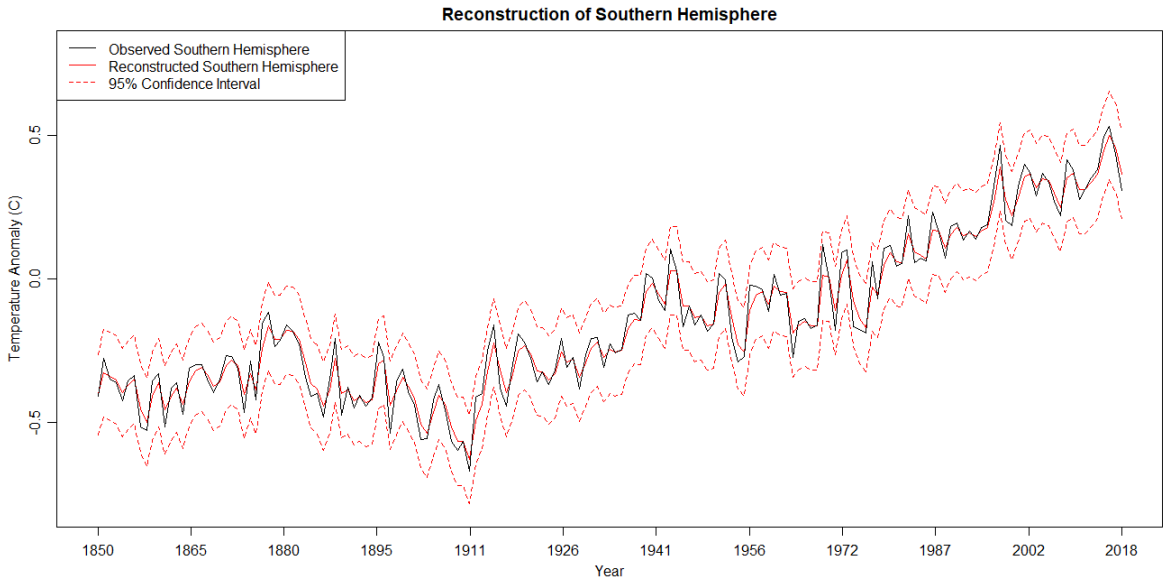


Figure 3: Reconstruction of the southern hemisphere and its temperature anomaly.

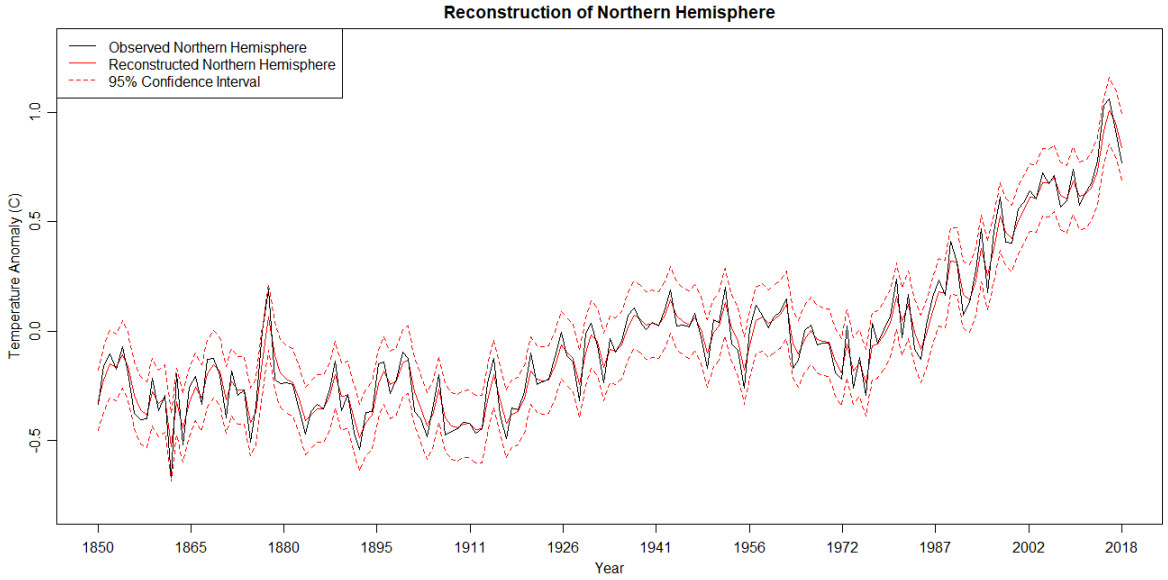


Figure 4: Reconstruction of the northern hemisphere and its temperature anomaly.

The Kalman filter seems to do a good job of reconstructing the data, with no observations outside the confidence interval. The **log-likelihood** is 204.9797 and will be used to compare models later. The likelihood is found by using (10.149) in [1]. Now looking at the predictions. Predictions are found by using (10.76) and updating the co-variance matrices by using (10.77) and (10.78). Once again this is continuously used for each Kalman filter, so it will be omitted in the next questions.

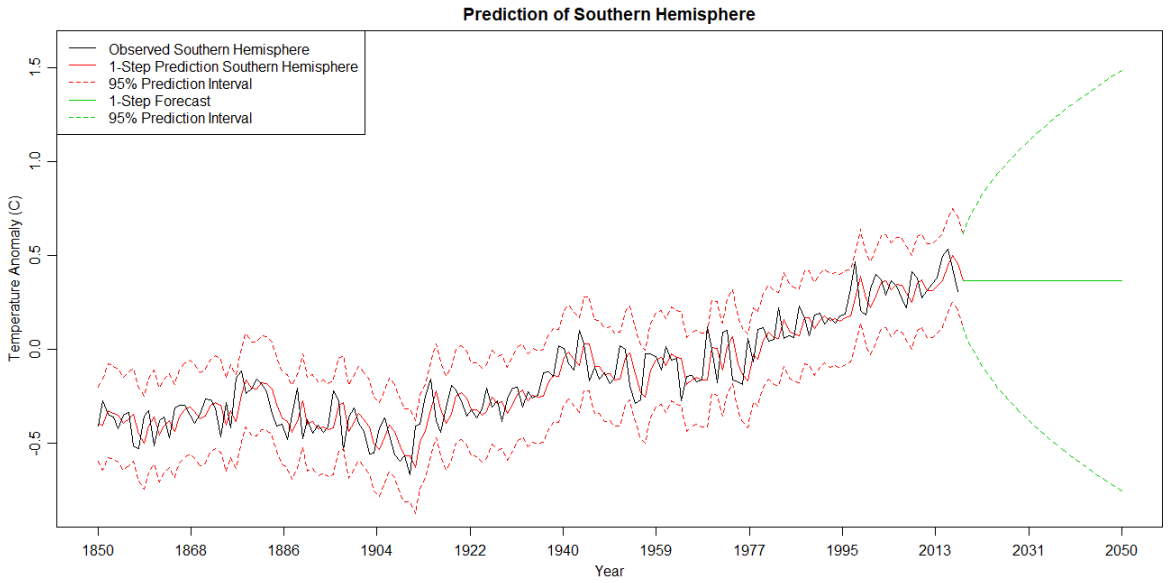


Figure 5: Prediction of the southern hemisphere and its temperature anomaly.

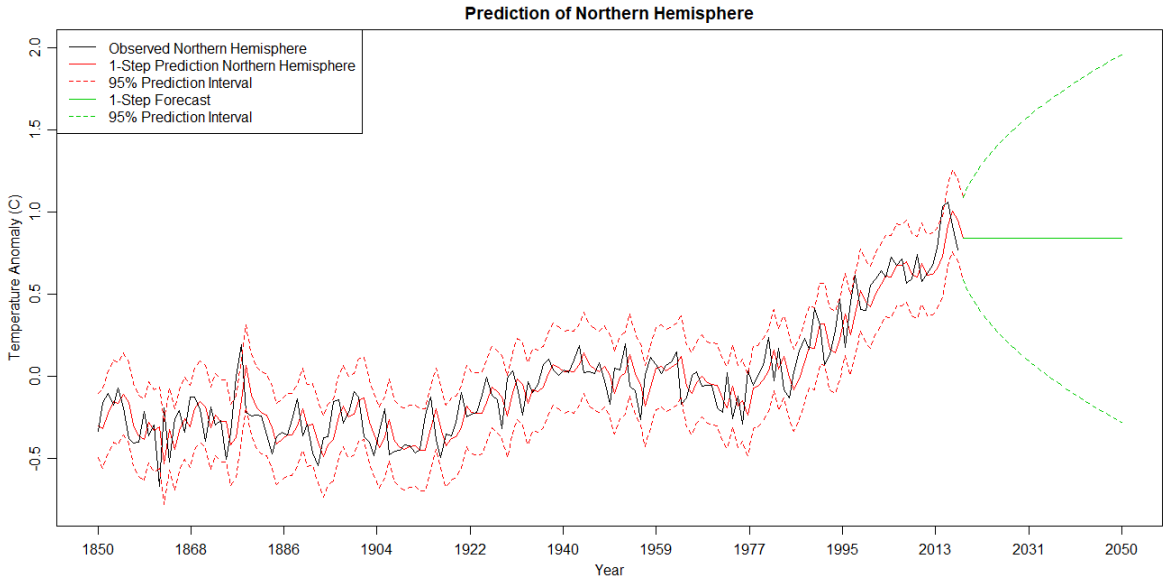


Figure 6: Prediction of the northern hemisphere and its temperature anomaly.

It can be seen that a constant prediction is achieved. This makes sense, since no trend is taken into account, which means the prediction will be the last in-sample observation predicted. The confidence interval looks fine, it is wide, but also captures a lot of the previous observations, which is good. The further we predict, the more uncertain the predictions will be as seen in the confidence interval. All the intervals was found by using equation 1:

$$conf(N_t) = N_t \pm 1.96 \cdot \sqrt{var(N_t)} \tag{1}$$

Where N_t is the values of the reconstruction or the 1-step predictions. It is assumed that the t-distribution has enough degrees of freedom to approximate a normal distribution, hence the 1.96. The values of the predictions can be seen in table 1.

Table 1: Table of predictions

Southern Hemisphere			Northern Hemisphere		
Year	Prediction	95% Confidence Interval	Year	Prediction	95% Confidence Interval
2020	0.3653181	[0.0482,0.6824]	2020	0.8393056	[0.5222,1.1564]
2030	0.3653181	[-0.3309,1.0616]	2030	0.8393056	[0.1431, 1.5355]
2040	0.3653181	[-0.5668,1.2975]	2040	0.8393056	[-0.0928,1.7715]
2050	0.3653181	[-0.7541,1.4847]	2050	0.8393056	[-0.2801,1.9587]

The predictions are very simplified, with it being based on the last in-sample 1-step prediction. As can be seen from the previous data, there seems to be a trend, which is not defined in the state space model. Therefore one shouldn't trust the predictions. Also, the values chosen are not optimal, but rather naive guesses. This will be optimised in the next question.

Question 4.4: Optimising parameters

In this question the initial values will be optimised. To do this the optimise function, *optim*, in *R* will be used. Optimal starting values will be found by minimising the negative log-likelihood. All values will be found by using numerical optimisation in the form of the limited memory *BFGS* algorithm, where the relative convergence tolerance of *optim* is the default of $1.490116\text{e}-08$. To avoid negative variance all the parameters which are being numerically optimised is transformed by using a log-transform, except the initial temperature, which is allowed to be within the negative domain. After optimising the parameters, the **log-likelihood** increases to 238.1264. This is a good improvement. The parameters optimised is the initial variance P_0, the variance of the system and observation noise Σ_1, Σ_2, and finally the initial guess values. This means 8 parameters in total to optimise. This gives the following initial matrices, for the start value:

$$\mathbf{X}_0 = \begin{bmatrix} -0.3638 \\ -0.2319 \end{bmatrix}$$

The initial variance matrix is:

$$P_0 = \begin{bmatrix} 1.360\text{e}-09 & 0 \\ 0 & 1.777\text{e}-08 \end{bmatrix}$$

And the variance of the system and observation noise is given as:

$$\Sigma_1 = \begin{bmatrix} 0.0015 & 0 \\ 0 & 0.0034 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} 0.0067 & 0 \\ 0 & 0.0113 \end{bmatrix}$$

Looking at the reconstructing in figure (7) and (8):

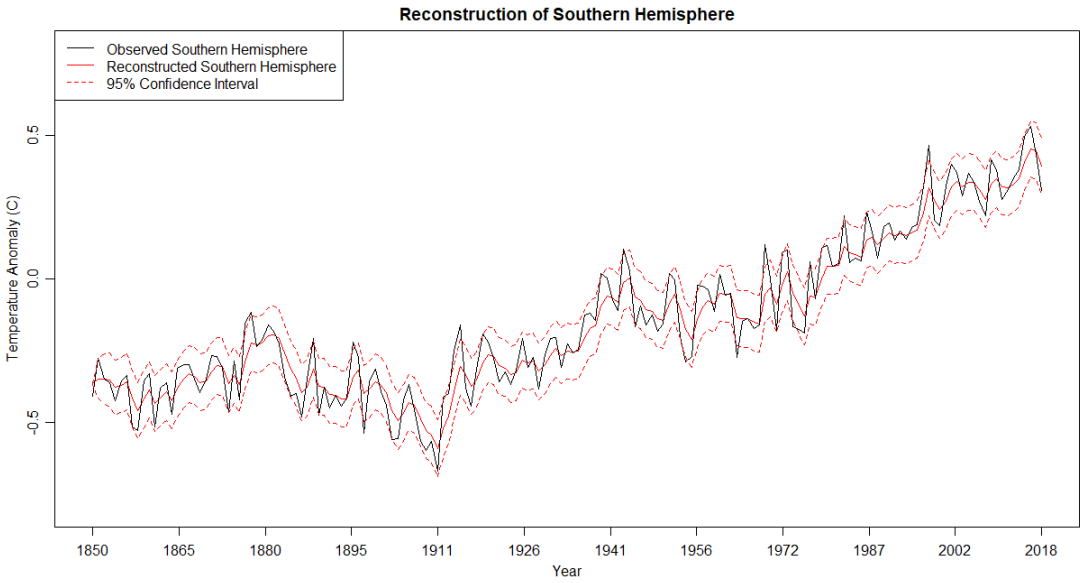


Figure 7: Reconstruction of the southern hemisphere and its temperature anomaly.

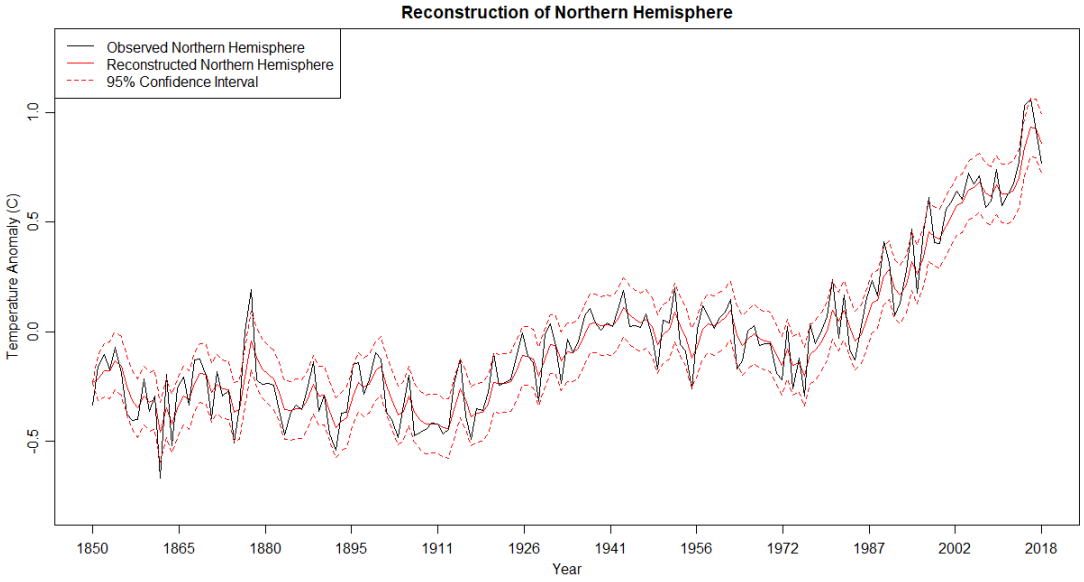


Figure 8: Reconstruction of the northern hemisphere and its temperature anomaly.

The Kalman filter reconstructing the data looks fine, a few observations outside the confidence interval, but it also got a lot more narrow, which means less uncertainty. Now looking at the predictions:

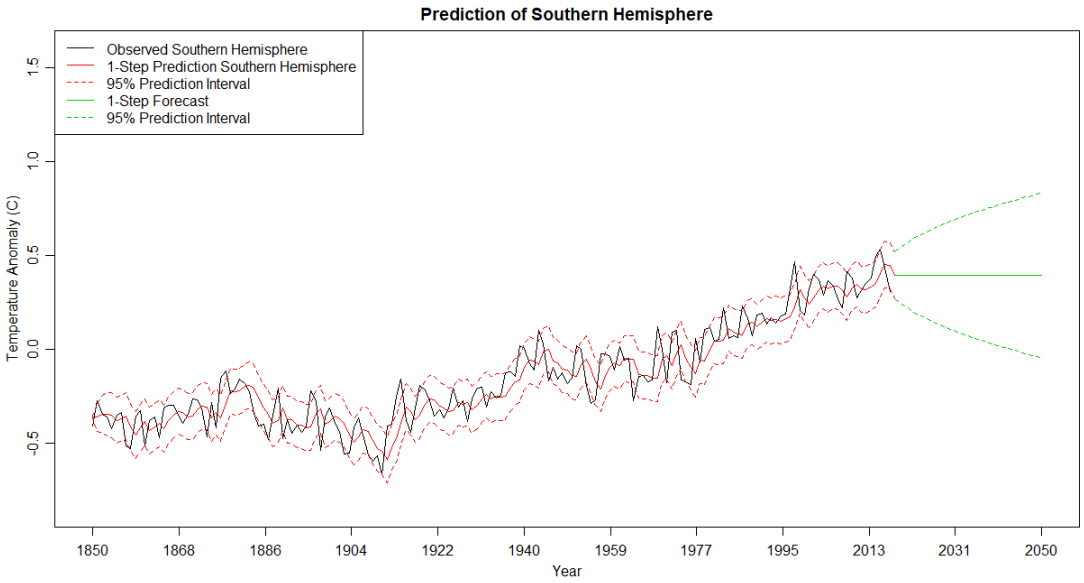


Figure 9: Prediction of the southern hemisphere and its temperature anomaly.

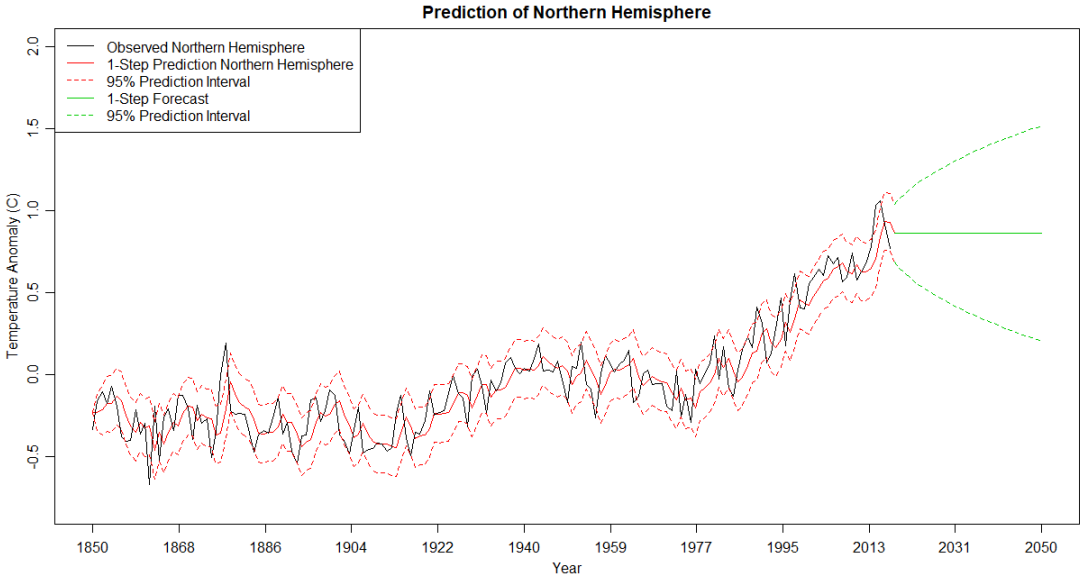


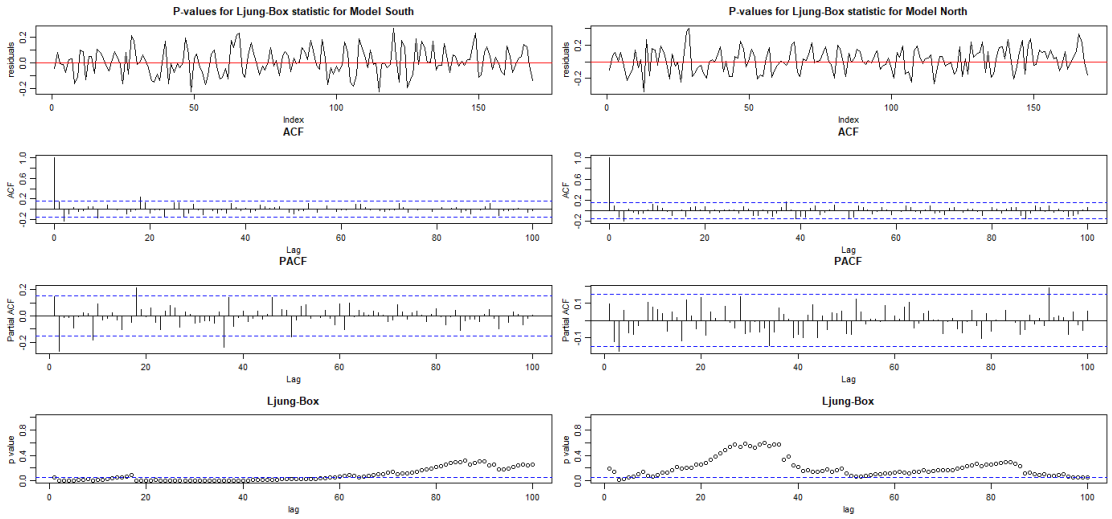
Figure 10: Prediction of the northern hemisphere and its temperature anomaly.

The prediction interval is more narrow, the optimisation made the predictions less uncertain by optimising the parameters. Predicting further into the future increases the uncertainty. Below a table of the predictions can be seen:

Table 2: Predictions for Kalman filter with optimized parameters

Southern Hemisphere			Northern Hemisphere		
Year	Prediction	95% Confidence Interval	Year	Prediction	95% Confidence Interval
2020	0.3936376	[0.2486,0.5387]	2020	0.8613905	[0.6517,1.0711]
2030	0.3936376	[0.1139,0.6734]	2030	0.8613905	[0.4451,1.2777]
2040	0.3936376	[0.02552,0.7617]	2040	0.8613905	[0.3112,1.4116]
2050	0.3936376	[-0.0454,0.8327]	2050	0.8613905	[0.2041,1.5187]

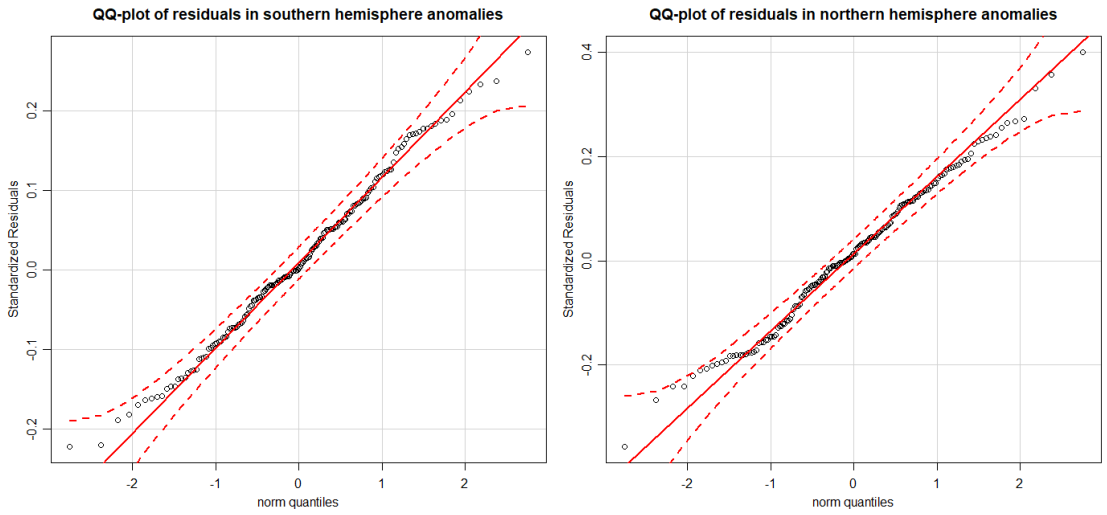
The predictions are constant, since no trend is assumed, eq. (10.76) will therefore just be based on the last in-sample 1-step prediction. In the next question the same system will be set up, but allowing correlation between the latent variables. Finally the residuals of the state space model is checked to see if they follow the IID property and are random noise.



(a) Southern hemisphere (b) Northern hemisphere

Figure 11: Two plots showing the residuals, ACF and PACF and Ljung-Box of the residuals.

The noise looks random and indeed gives the following confidence intervals when making a sign-test for south and north, ([0.3755840 0.5308878]), which contain $P(S) = 0.5$ in the interval. There seems to be a few significant lags in the PACF, which means not all information is contained in the model. Finally the Ljung-Box test could be better for the southern residuals.



(a) Southern hemisphere (b) Northern hemisphere

Figure 12: Two plots showing the QQ-plot of the standardized residuals.

The residuals are well contained within the confidence interval and the residuals are assumed normally distributed, which is good.

Question 4.5: Optimize parameters - with correlation

It was seen in question 4.1 that the anomalies of each hemisphere were rather similar. Therefore the model is adjusted to allow the system noise to be correlated. This is done by changing the variance matrix of the system noise:

$$\Sigma_1 = \begin{bmatrix} \sigma_{\mathbf{e}_1,11,t} & \sigma_{\mathbf{e}_1,21,t} \\ \sigma_{\mathbf{e}_1,12,t} & \sigma_{\mathbf{e}_1,22,t} \end{bmatrix}$$

Where $\sigma_{12,t} = \sigma_{21,t}$. Therefore 9 parameters will now be optimised. The reconstruction of the model including correlation is then seen in (13 and 14) and the initial matrices can be seen below. The initial temperature anomaly values:

$$\mathbf{X}_0 = \begin{bmatrix} -0.4021 \\ -0.2760 \end{bmatrix}$$

The initial variance matrix is:

$$P_0 = \begin{bmatrix} 4.222\text{e}-07 & 0 \\ 0 & 1.4538\text{e}-06 \end{bmatrix}$$

And the variance of the system and observation noise is given as:

$$\Sigma_1 = \begin{bmatrix} 0.0051 & 0.0068 \\ 0.0068 & 0.0103 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} 0.0038 & 0 \\ 0 & 0.0061 \end{bmatrix}$$

The reconstruction is:

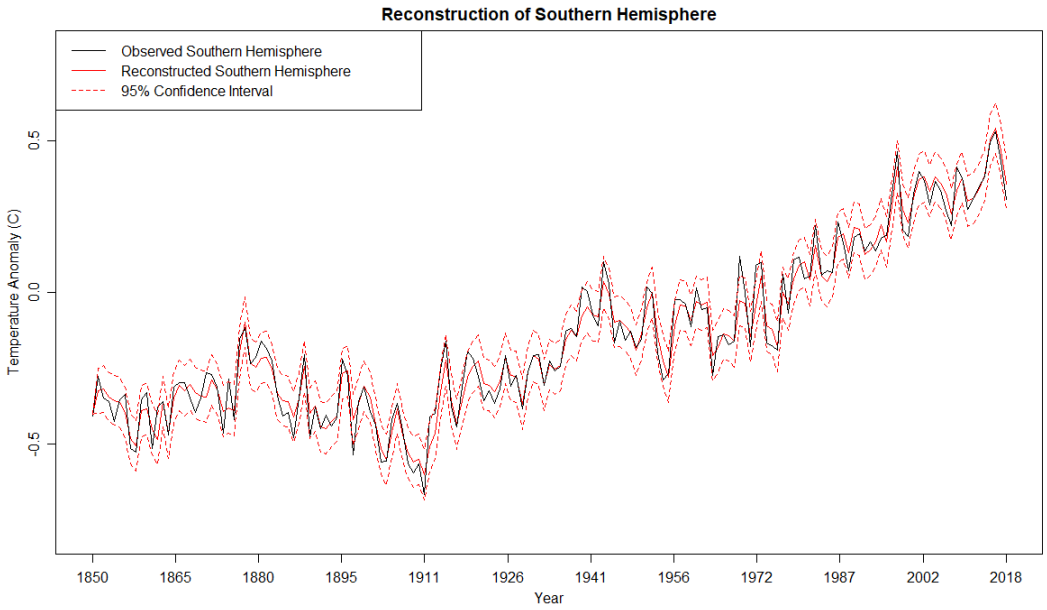


Figure 13: Reconstruction of the southern hemisphere and its temperature anomaly.

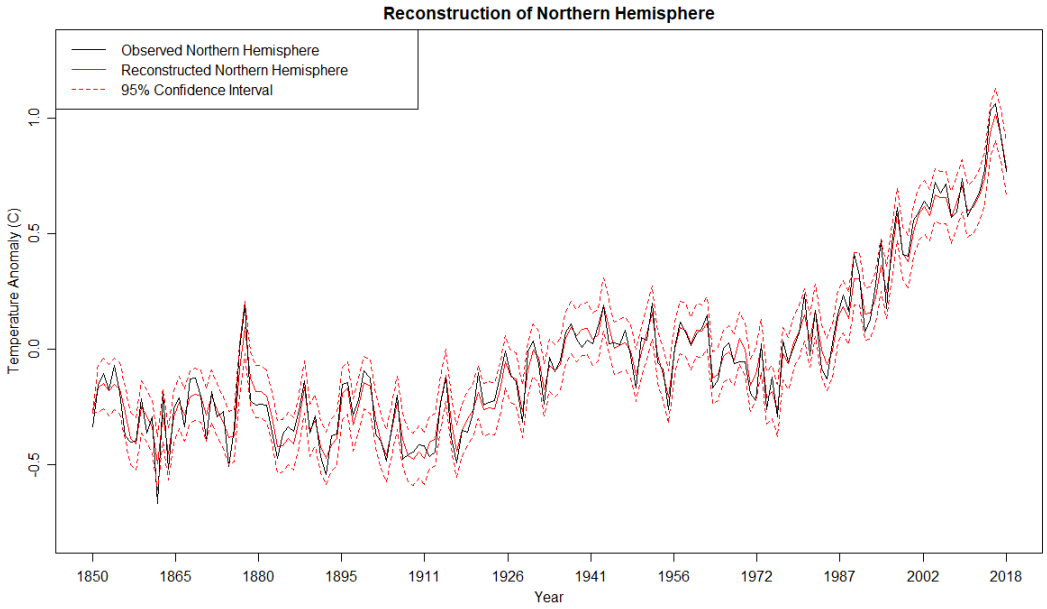


Figure 14: Reconstruction of the northern hemisphere and its temperature anomaly.

The Kalman filter reconstructing the data looks good, less observations are outside the confidence interval, but it is still narrow, which means less uncertainty. Now looking at the predictions:

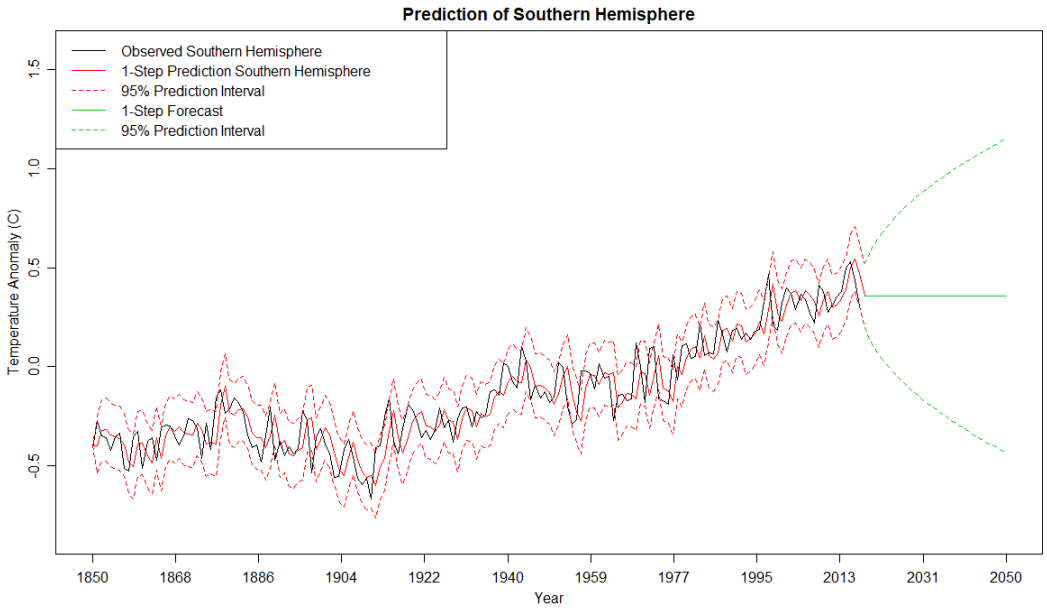


Figure 15: Prediction of the southern hemisphere and its temperature anomaly.

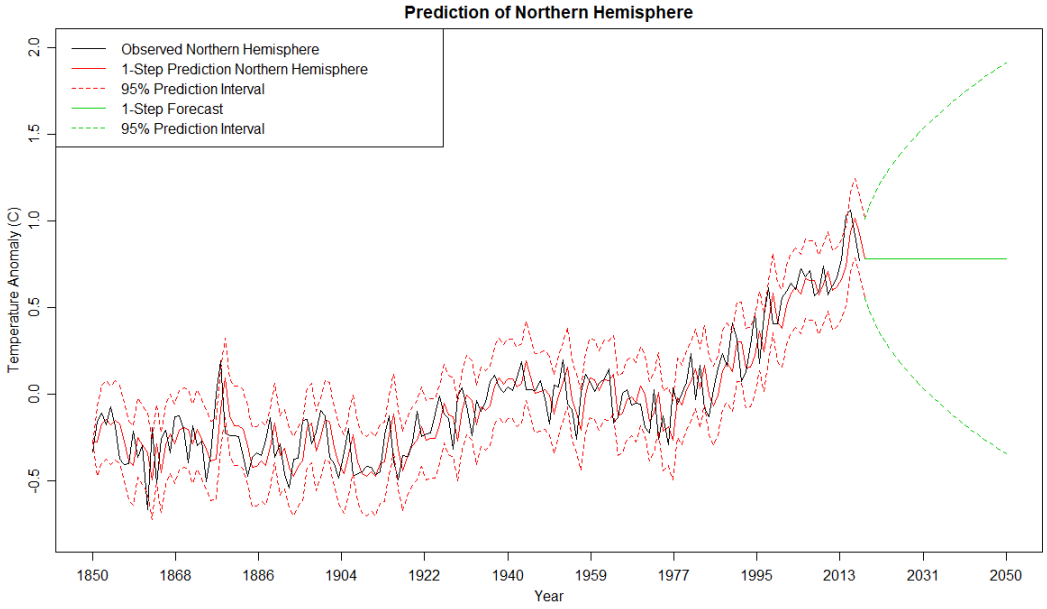


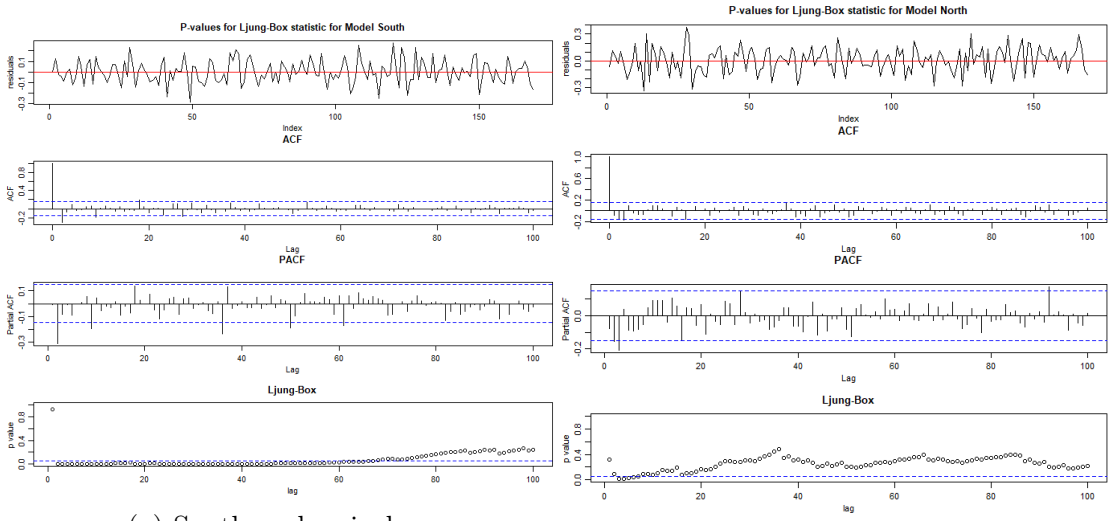
Figure 16: Prediction of the northern hemisphere and its temperature anomaly.

The prediction interval is a lot wider. This means more uncertainty, but the correlation is also allowed now and therefore more of the information is explained. This implicit a wider interval as expected. The **log-likelihood** with correlation included is 268.5282. An improvement compared to the other models. The predictions can be seen in table 3

Table 3: Table of predictions.

Southern Hemisphere			Northern Hemisphere		
Year	Prediction	95% Prediction Interval	Year	Prediction	95% Prediction Interval
2020	0.3576152	[0.1433,0.5720]	2020	0.784497	[0.4811,1.0879]
2030	0.3576152	[-0.1328,0.8480]	2030	0.784497	[0.0861,1.4829]
2040	0.3576152	[-0.3019,1.0172]	2040	0.784497	[-0.1554,1.7244]
2050	0.3576152	[-0.4358,1.1511]	2050	0.784497	[-0.3465,1.9155]

Once again the predictions are constant, as expected from eq. (10.76) in [1]. Looking at the residuals of the observational equation:

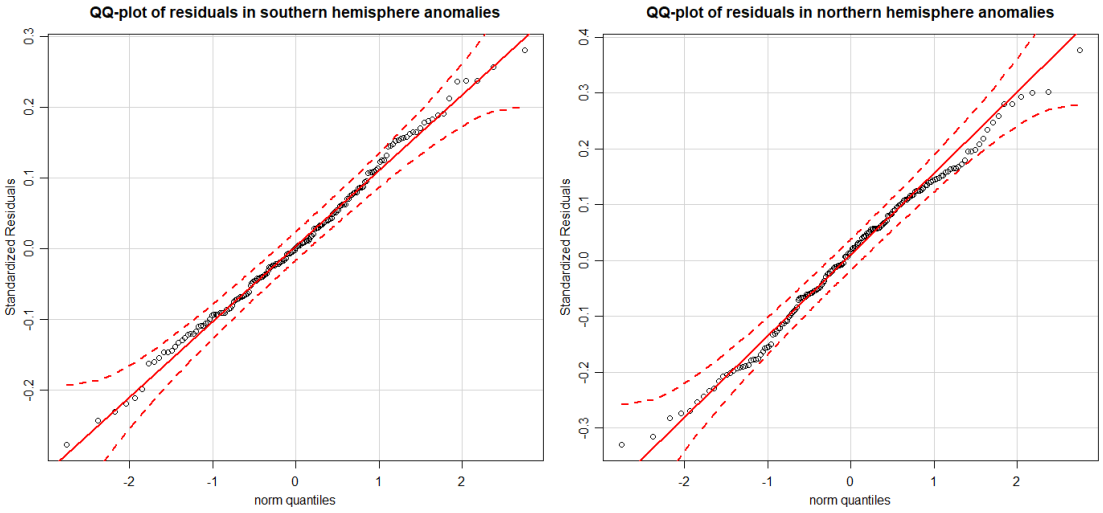


(a) Southern hemisphere

(b) Northern hemisphere

Figure 17: Two plots showing the residuals, ACF and PACF and Ljung-Box of the residuals.

The noise looks random and indeed gives the following confidence intervals when making a sign-test for south, ([0.4220095, 0.5779905]), and for the north ([0.3987119 0.5545238]), which both contain $P(S) = 0.5$ in the interval. There seems to be a few significant lags in the PACF and ACF, which means not all information is contained in the model. Finally the Ljung-Box test could again be better for the southern residuals.



(a) Southern hemisphere (b) Northern hemisphere

Figure 18: Two plots showing the QQ-plot of the standardized residuals.

The residuals are well contained within the confidence interval and the residuals are assumed normally distributed, which is good. The next question will look at the trend of the data.

Question 4.6: Formulating state space model with common trend

It was clear from question 4.1 that the data is not stationary. Therefore a common trend will be included in the state space model in this question. The trend is assumed to be a random walk itself and will be added to the latent state space vector as a stochastic variable $X_{T,t}$:

$$\mathbf{X}_t = \begin{bmatrix} X_{sh,t} \\ X_{nh,t} \\ X_{T,t} \end{bmatrix}$$

Once again $\mathbf{A}_t$ and $\mathbf{C}_t$ is defined with suitable dimensions. $\mathbf{A}_t$ will now be a 3×3 matrix and $\mathbf{C}_t$ will be 2×3. For $\mathbf{A}_t$ the states for the anomalies should have an added trend, therefore $\mathbf{A}_t$ is:

$$\mathbf{A}_t = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

For the observation equation, the trend is not of interest, but only the output of the anomalies. This means $\mathbf{C}_t$ is given as:

$$\mathbf{C}_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Finally the variance of the random noise vectors are:

$$\Sigma_1 = \begin{bmatrix} \sigma^2_{\mathbf{e}_1,11,t} & \sigma^2_{\mathbf{e}_1,12,t} & 0 \\ \sigma^2_{\mathbf{e}_1,21,t} & \sigma^2_{\mathbf{e}_1,22,t} & 0 \\ 0 & 0 & \sigma^2_{\mathbf{e}_1,33,t} \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} \sigma^2_{\mathbf{e}_2,11,t} & 0 & 0 \\ 0 & \sigma^2_{\mathbf{e}_2,22,t} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Σ_1 also include the correlation between the southern and northern anomalies. The full state space model will now be illustrated:

$$\mathbf{X}_t = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_{sh,t-1} \\ X_{nh,t-1} \\ X_{T,t-1} \end{bmatrix} + \begin{bmatrix} \mathrm{e}_{1,sh,t} \\ \mathrm{e}_{1,nh,t} \\ \mathrm{e}_{1,T,t} \end{bmatrix}$$

$$\mathbf{Y}_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_{sh,t} \\ X_{nh,t} \\ X_{T,t} \end{bmatrix} + \begin{bmatrix} \mathrm{e}_{2,sh,t} \\ \mathrm{e}_{2,nh,t} \\ \mathrm{e}_{2,T,t} \end{bmatrix}$$

Where $\mathbf{e}_{2,t} \sim N(0, \Sigma_{2,t})$ and $\mathbf{e}_{1,t} \sim N(0, \Sigma_{1,t})$ and following the same assumptions as the errors in 4.2. This is the state space model including a trend and correlation, which will be optimized in question 4.7.

Question 4.7: Optimize parameters - with correlation and common trend

The same procedure will be used as in the previous questions. This time the trend is included and the number of parameters optimised will increase. The initial variance will have an extra parameter to optimise for the trend, the starting guess will be set to 0.01. The initial guess for the trend will be 0.0001, since it seems to increase rather slowly. Finally the variance matrix of the system noise includes the trend parameter. One should be careful with the starting guess of this parameter, since the *BFGS* algorithm might converge incorrectly. Looking at the figure of the data in question 4.1, (1), one can see that the trend increases very little over the entire span of the time series. Since the standard deviation of the series seems to be range from 0.001 to 0.003. Let's assume the variance will be around 0.002^2, therefore $4e-06$ will be our starting guess. The reconstruction and optimised initial values can be seen below. The initial temperature anomaly values:

$$\mathbf{X}_0 = \begin{bmatrix} -0.4041 \\ -0.2757 \\ 0.0027 \end{bmatrix}$$

The initial variance matrix is:

$$P_0 = \begin{bmatrix} 1.1686\text{e}-06 & 0 & 0 \\ 0 & 3.2716\text{e}-06 & 0 \\ 0 & 0 & 2.1763\text{e}-07 \end{bmatrix}$$

And the variance of the system and observation noise is given as:

$$\Sigma_1 = \begin{bmatrix} 0.0049 & 0.0067 & 0 \\ 0.0067 & 0.0103 & 0 \\ 0 & 0 & 9.001\text{e}-13 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} 0.0038 & 0 \\ 0 & 0.0061 \end{bmatrix}$$

The reconstruction is:

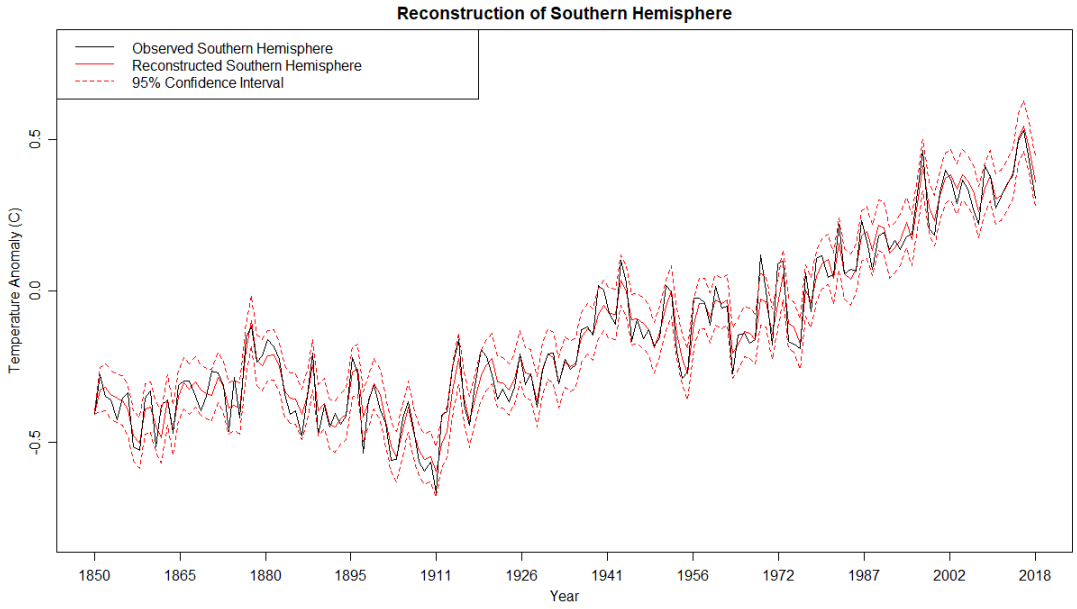


Figure 19: Reconstruction of the southern hemisphere and its temperature anomaly.

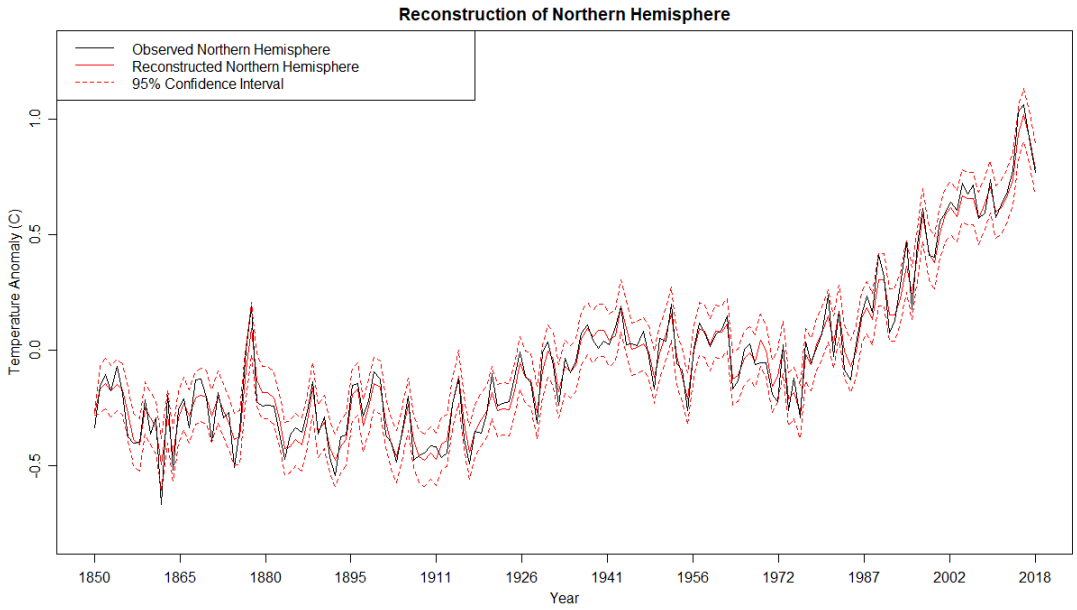


Figure 20: Reconstruction of the northern hemisphere and its temperature anomaly.

The reconstruction looks good. All the data seems to be within the confidence interval which is very narrow. The predictions are now shown:

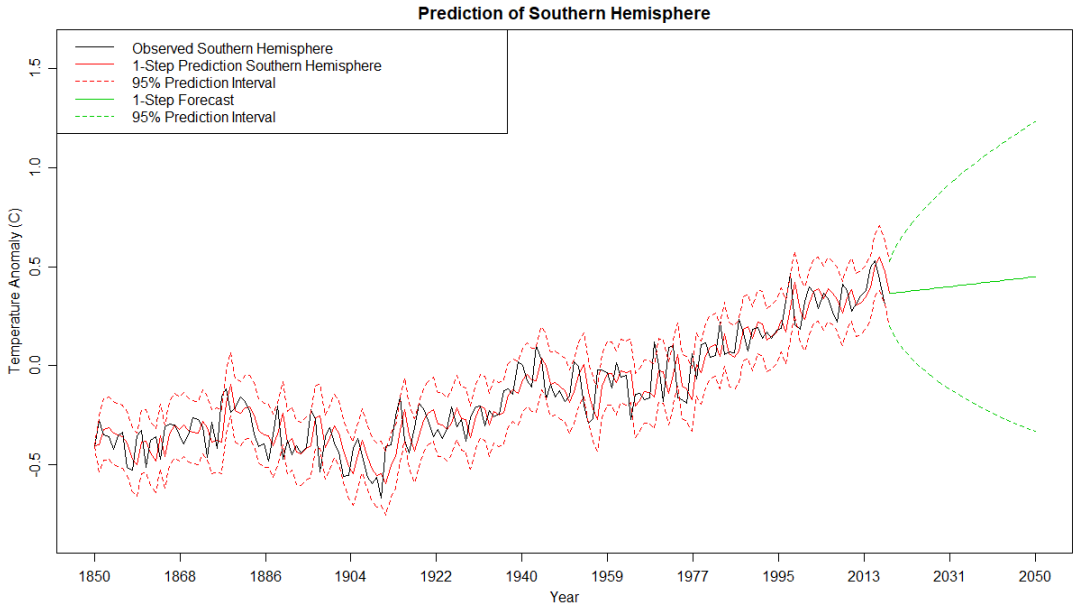


Figure 21: Prediction of the southern hemisphere and its temperature anomaly.

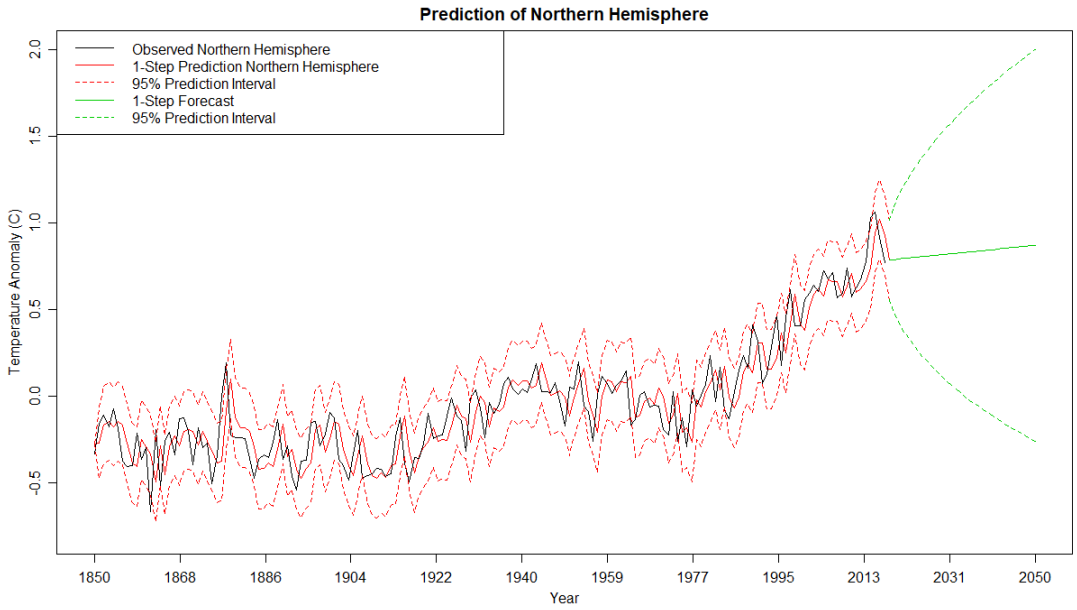


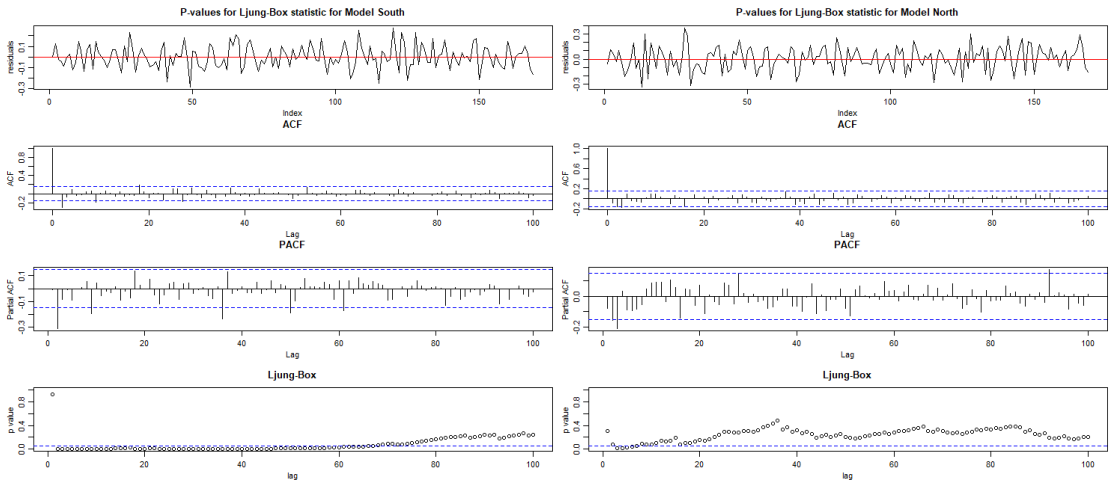
Figure 22: Prediction of the northern hemisphere and its temperature anomaly.

The **log-likelihood** is 268.7179, which is the best log-likelihood so far. The predictions also slowly increases now due to the trend, which seems to be a better prediction than the previous models, as the current trend is positive. The prediction intervals are large and encapsulate most of the time series. This seems to be an overall improvement compared to the other models. The values of the predictions can be seen in table 4:

Table 4: Table of prediction for the state space model including the trend and correlation.

Southern Hemisphere			Northern Hemisphere		
Year	Prediction	95% Prediction Interval	Year	Prediction	95% Prediction Interval
2020	0.3663202	[0.1545,0.5782]	2020	0.7895906	[0.4859,1.0933]
2030	0.3936827	[-0.0902,0.8776]	2030	0.816953	[0.1179,1.5160]
2040	0.4210452	[-0.2297,1.0718]	2040	0.8443155	[-0.0965,1.7851]
2050	0.4484076	[-0.3344,1.2312]	2050	0.8716779	[-0.2604,2.0037]

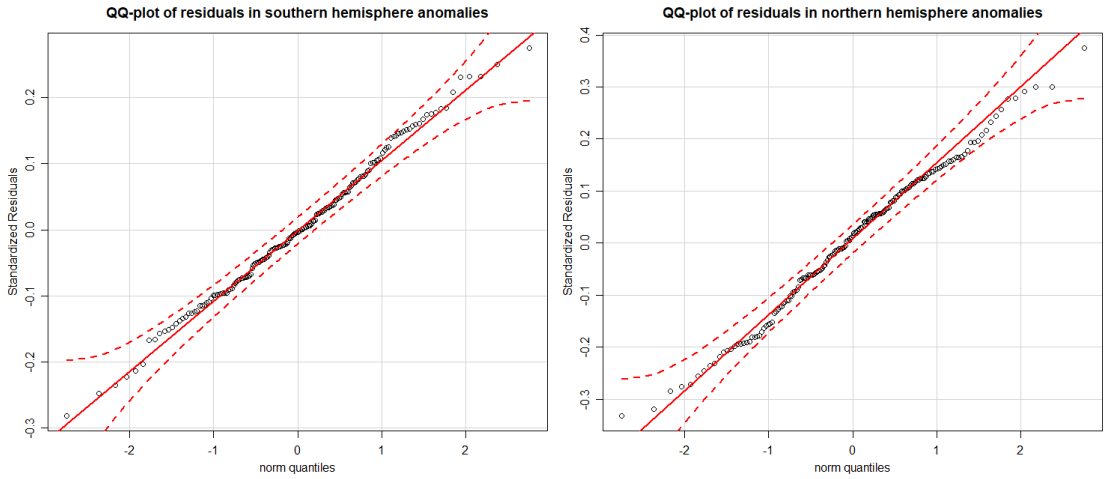
The predictions are no longer constant, this is because a trend is added to the latent variables for each of the anomalies. The residuals are now analysed:



(a) Southern hemisphere (b) Northern hemisphere

Figure 23: Two plots showing the residuals, ACF and PACF and Ljung-Box of the residuals.

The noise looks random and indeed gives the following confidence intervals when making a sign-test for south and north ([0.3987119, 0.5545238]), which both contain $P(S) = 0.5$ in the interval. There seems to be significant lags in the PACF and ACF, which means not all information is contained in the model, the model might be too simplistic. Finally the Ljung-Box test could again be better for the southern residuals and also some of the first few residuals for the north could be improved.



(a) Southern hemisphere (b) Northern hemisphere

Figure 24: Two plots showing the QQ-plot of the standardized residuals.

The residuals are still well contained within the confidence interval and the residuals are assumed normally distributed, which is good.

Question 4.8: Comparison

Looking at the log-likehood of each model:

Table 5: Log-Likelihood of each model.

	Model 1	Model 2	Model 3	Model 4
Log-Likelihood	204.9797	238.1264	268.5282	268.717

For model 1, it had the lowest log-likelihood. It was the unoptimised model assuming independent random walks between the variables. Clearly this model is too simple for trustworthy predictions. Optimising the parameters made it significantly better in model 2, with a log-likelihood of 238.1264. However it's still a very simple model and the predictions shouldn't be trusted. Furthermore the reconstruction of model 2 seemed to be a bit off, with more values not within the confidence interval. Model 3 incorporated the correlation between the anomalies, this made the log-likelihood a lot better. However the model still assumes a constant prediction, which is not very trustworthy when looking at the trend of the time series. Finally model 4 had the highest log-likelihood, it extends model 3 with including the trend. Even though model 4 has the highest log-likelihood, it is very close to model 3, however model 4 has more parameters, so if comparing the two models using AIC, model 3 would be best. Still it's clear model 4 predicts a lot better. The model now takes the trend into account and the predictions seem to be more trustworthy for a short horizon prediction. The prediction intervals become very large when predicting a few years out and therefore one should avoid long-term predictions. Model 4 is the best model out of the other three models for predicting in this case. The residuals seemed to be equally good in all 4 models. There is still room for improvement, this will now be discussed.

To improve the state space model several things can be done. One of the things is to make a MARIMA model and formulate it to a state space model. It could be seen in all of the residual plots of the ACF and PACF that some of the autocorrelation is still significant in the PACF and ACF. Furthermore looking at the CCF of the residuals for the last model including the correlation and trend, there also still seem to be some significant correlation between the two hemispheres left:

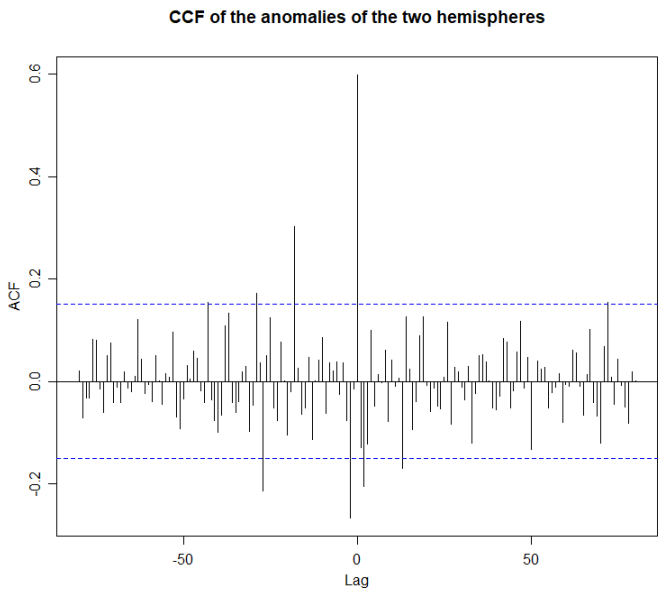


Figure 25: CCF of the temperature anomaly of the northern and southern hemisphere.

This means some information is not explained. The observational equation is just dependent on the most recent observation $t-1$, a higher lag order would be interesting to try. Furthermore certain residuals are still significant in the Ljung-Box which amplify the superstition of not capturing all the information in the time series. By using table 6.1 in [1], one would be able to guess a correct order. A MARIMA model would also change the $\mathbf{A}_t$ matrix and include coefficients instead of just ones. This means not all latent variables $\mathbf{X}_t$ are weighted equally. This may make the model perform better. It also leads to the second idea.
Assuming that the two anomalies are following an independent random walk, might be too simplistic. A more complex design matrix $\mathbf{A}_t$ for the system equation might be beneficial in making better predictions. One could imagine when temperature anomalies rises in the north it will also rise in the south, which would require elements in the off-diagonal of the matrix $\mathbf{A}_t$. This is taken into account for the system noise, but not for when making the predictions as can be seen in (10.76) in [1].
A third suggestion would be to include a deterministic input vector u_t for the state-space model. To only look at the anomaly of temperature to give an optimal prediction might be too simplistic. There might be underlying factors. One could imagine the CO_2 levels might explain the positive trend, so a higher CO_2 level would mean a higher prediction of temperature anomaly. Other factors such as sun hours, methane emissions and other factors alike, would be interesting to include in the model for a better performance.
A final suggestion is to have an individual trend for each of the hemispheres. From the time series, it can be seen that the trend of the northern hemisphere anomaly seems to be slightly higher than the southern hemisphere. Therefore making individual trends instead of common trends, might give a better prediction. It might not make a huge difference in performance, but still worthy of a mention.

References

[1] Henrik Madsen, *Time Series Analysis*, Chapman & Hall/CRC, 2008, Chapter 10.

Appendix

```
################ASSIGNMENT 4#################
library("forecast")
library(MASS)
library(tseries)
library(car)
library(marima)
library("FKF")
par(mgp=c(2,0.8,0), mar=c(3,3,2,1),las=0,
    pty="m", xpd=F)
setwd("")
dat = read.csv('A4_annual.txt', header=TRUE,
               sep ="\t",dec = ".",colClasses =
                 c("integer",rep("numeric",2)))

ljung.test<-function(x, nlag){
  res <- x
  pval <- sapply(1:nlag, function(i) Box.test(
    res, i, type = "Ljung-Box")$p.value)
  plot(1L:nlag, pval, xlab = "lag",
       ylab = "p value",main="Ljung-Box", ylim = c(0,1))
  abline(h = 0.05, lty = 2, col = "blue")
}

sign.test<-function(x){
  res <- x
  (N.sign.changes <- sum( res[-(1)] *
                            res[-length(res)]<0 ))
  print(binom.test(N.sign.changes,
                   (length(res))-1))
}

diagtool<- function(residuals,tlag, modnr){
  par(mfrow=c(4,1), mar=c(3,3,3,1), mgp=c(2,0.8,0))
  plot(residuals , type="l",
       main = paste("P-values for Ljung-Box statistic for Model",
       modnr))
  abline(0,0, col=2)
  acf(residuals ,lag=tlag, main="ACF")
  pacf(residuals ,lag=tlag, main="PACF")
  ljung.test(residuals,tlag)
  sign.test(residuals)
  par(mfrow=c(1,1))
```

```
}

plot.kf.rec <- function(x, string){
  par(mgp=c(2,0.8,0), mar=c(3,3,2,1))
  N<- 169
  year.plot <- c(1850:2018)
  if(string == "south"){
    plot(x$att[1,], type='l', ylim=c(-0.8, 0.8),
         ylab="Temperature Anomaly (C)",
         xlab = "Year",  xaxt='n', col=2,
         main="Reconstruction of Southern Hemisphere")
    lines(dat$sh, col=1)
    axis(1, c(seq(1,N,length.out=12)),
         year.plot[c(seq(1,N,length.out=12))])
    upper = x$att[1,] + 1.96*sqrt(x$Ptt[1, 1,])
    lower = x$att[1,] - 1.96*sqrt(x$Ptt[1, 1,])
    lines(1:169,upper, col=2, lty=2)
    lines(1:169,lower, col=2, lty=2)
    legend("topleft", legend=c("Observed Southern
                                Hemisphere",
                               "Reconstructed Southern
                                Hemisphere",
                               "95% Confidence Interval"),
           col=c(1,2,2), lty=c(1,1,2))
  }else{
    plot(x$att[2,], type='l', ylim=c(-0.8, 1.3),
         ylab="Temperature Anomaly (C)",
         xlab = "Year",  xaxt='n', col=2,
         main="Reconstruction of Northern Hemisphere")
    lines(dat$nh, col=1)
    axis(1, c(seq(1,N,length.out=12)),
         year.plot[c(seq(1,N,length.out=12))])
    upper = x$att[2,] + 1.96*sqrt(x$Ptt[2, 2,])
    lower = x$att[2,] - 1.96*sqrt(x$Ptt[2, 2,])
    lines(1:169,upper, col=2, lty=2)
    lines(1:169,lower, col=2, lty=2)
    legend("topleft", legend=c("Observed Northern Hemisphere",
                               "Reconstructed Northern
                                Hemisphere",
                               "95% Confidence Interval"),
           col=c(1,2,2), lty=c(1,1,2))
  }
}

plot.kf.pred <- function(x, string){
  par(mgp=c(2,0.8,0), mar=c(3,3,2,1))
  N<- 169
  year.plot <- c(1850:2050)
  if(string == "south"){
    plot(x$at[1,], type='l', ylim=c(-0.85, 1.6),
```

```
         ylab="Temperature␣Anomaly␣(C)",
         xlab = "Year",  xaxt='n', col=2,
         main="Prediction␣of␣Southern␣Hemisphere")
    lines(dat$sh, col=1)
    axis(1, c(seq(1,N+32,length.out=12)),
         year.plot[c(seq(1,N+32,length.out=12))])
    upper = x$at[1,] + 1.96*sqrt(x$Pt[1, 1,])
    lower = x$at[1,] - 1.96*sqrt(x$Pt[1, 1,])
    lines(1:170,upper[1:170], col=2, lty=2)
    lines(1:170,lower[1:170], col=2, lty=2)
    lines(170:201,upper[170:201], col=3, lty=2)
    lines(170:201,lower[170:201], col=3, lty=2)
    lines(170:201,x$at[1,170:201], col=3)
    legend("topleft", legend=c("Observed␣Southern␣Hemisphere",
                               "1-Step␣Prediction␣Southern
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣Hemisphere",
                               "95%␣Prediction␣Interval",
                               "1-Step␣Forecast",
                               "95%␣Prediction␣Interval"),
           col=c(1,2,2,3,3), lty=c(1,1,2,1,2))
  }else{
    plot(x$at[2,], type='l', ylim=c(-0.8, 2),
         ylab="Temperature␣Anomaly␣(C)",
         xlab = "Year",  xaxt='n', col=2,
         main="Prediction␣of␣Northern␣Hemisphere")
    lines(dat$nh, col=1)
    axis(1, c(seq(1,N+32,length.out=12)),
         year.plot[c(seq(1,N+32,length.out=12))])
    upper = x$at[2,] + 1.96*sqrt(x$Pt[2, 2,])
    lower = x$at[2,] - 1.96*sqrt(x$Pt[2, 2,])
    lines(1:170,upper[1:170], col=2, lty=2)
    lines(1:170,lower[1:170], col=2, lty=2)
    lines(170:201,upper[170:201], col=3, lty=2)
    lines(170:201,lower[170:201], col=3, lty=2)
    lines(170:201,x$at[2,170:201], col=3)
    legend("topleft", legend=c("Observed␣Northern␣Hemisphere",
                               "1-Step␣Prediction␣Northern
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣Hemisphere",
                               "95%␣Prediction␣Interval",
                               "1-Step␣Forecast",
                               "95%␣Prediction␣Interval"),
           col=c(1,2,2,3,3), lty=c(1,1,2,1,2))
  }
}

print.pred <- function(x){

  upper = x$at[1,] + 1.96*sqrt(x$Pt[1, 1,])
  lower = x$at[1,] - 1.96*sqrt(x$Pt[1, 1,])
  upper2 = x$at[2,] + 1.96*sqrt(x$Pt[2, 2,])
```

```
  lower2 = x$at[2,] - 1.96*sqrt(x$Pt[2, 2,])
  print("south")
  print(x$at[1,169+2])
  print(x$at[1,169+12])
  print(x$at[1,169+22])
  print(x$at[1,169+32])
  print("North")
  print(x$at[2,169+2])
  print(x$at[2,169+12])
  print(x$at[2,169+22])
  print(x$at[2,169+32])

  print("south")
  print(paste("[",lower[169+2],",",upper[169+2],"]"))
  print(paste("[",lower[169+12],",",upper[169+12],"]"))
  print(paste("[",lower[169+22],",",upper[169+22],"]"))
  print(paste("[",lower[169+32],",",upper[169+32],"]"))

  print("North")
  print(paste("[",lower2[169+2],",",upper2[169+2],"]"))
  print(paste("[",lower2[169+12],",",upper2[169+12],"]"))
  print(paste("[",lower2[169+22],",",upper2[169+22],"]"))
  print(paste("[",lower2[169+32],",",upper2[169+32],"]"))

}

#############Q4.1###############
N<-length(dat$year)
summary(dat)
cor(dat$sh,dat$nh)

par(mgp=c(2,0.8,0), mar=c(3,3,3,1))
par(mfrow=c(2,2))
acf(dat$sh, lag=50, main="ACF␣of␣southern␣hemisphere")
acf(dat$nh, lag=50,main="ACF␣of␣northern␣hemisphere")
pacf(dat$sh, lag=50, main="PACF␣of␣southern␣hemisphere")
pacf(dat$nh, lag=50, main="PACF␣of␣northern␣hemisphere")
par(mfrow=c(1,1))
ccf(dat$sh, dat$nh, main="CCF␣of␣northern␣and␣southern
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣hemisphere")

boxplot(dat$nh,dat$sh, ylab="Temperature␣Anomaly␣(C)",
        names=c("Northern␣Hemisphere", "Southern␣Hemisphere"),
        main="Box-Plot␣of␣the␣two␣hemispheres")
plot(dat$sh, type='l', xlab = "Year", ylab="Temperature
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣Anomaly␣(C)",
     main="Southern␣and␣Northern␣Hemisphere␣Temperature
␣␣␣␣␣␣␣␣␣␣␣␣Anomaly",
     xaxt='n', ylim=c(-0.7, 1.1))
axis(1, c(seq(1,N,length.out=12)),
```

```
     dat$year[c(seq(1,N,length.out=12))])
lines(dat$nh, type='l', col=2)
legend("topleft", legend=c("Nothern Hemisphere",
                           "Southern Hemisphere"),
       col=c(2,1), lty=c(1,1))


##############Q4.2###############
A <- matrix(c(1,0,0,1),nrow=2)
C <- matrix(c(1,0,0,1),nrow=2)

#############Q4.3################
Sigma1 <- diag(0.01,nrow=2)
Sigma2 <- diag(0.01,nrow=2)
kf1 <- fkf(a0=c(-0.4, -0.3), P0 = matrix(c(0.01,0,0,0.01),
                                         nrow=2),
           dt = matrix(c(0,0), nrow=2),
           Tt = A,
           ct=matrix(c(0,0), nrow=2),
           Zt = C, HHt = Sigma1,
           GGt = Sigma2,
           yt = cbind(rbind(dat$sh, dat$nh)))
diagtool(kf1$vt[1,],150,"South")
diagtool(kf1$vt[2,],150,"North")
kf1$logLik

kf1.pred <- fkf(a0=c(-0.4, -0.3),
                P0 = matrix(c(0.01,0,0,0.01),nrow=2),
           dt = matrix(c(0,0), nrow=2),
           Tt = A, ct=matrix(c(0,0), nrow=2),
           Zt = C, HHt = Sigma1,
           GGt = Sigma2,
           yt = cbind(rbind(dat$sh, dat$nh),
                      rbind(rep(NA,31),rep(NA,31))))
plot.kf.pred(kf1.pred, "south")
plot.kf.rec(kf1,"south")
plot.kf.pred(kf1.pred, "north")
plot.kf.rec(kf1,"north")

print.pred(kf1.pred)


#############Q4.4################

Optim.fun = function(vars){
  var0 <- matrix(c(exp(vars[3]), 0,0,exp(vars[4])), nrow=2)
  C <- matrix(c(1,0,0,1),nrow=2)
  A <- matrix(c(1,0,0,1),nrow = 2)
  Sigma2 <- matrix(c(exp(vars[7]),0,0,exp(vars[8])),nrow=2)
  Sigma1 <- matrix(c(exp(vars[5]), 0,0,exp(vars[6])), nrow=2)
```

```
  kf <- fkf(a0=c(vars[1], vars[2]), P0 = var0,
            dt = matrix(c(0,0), nrow=2), Tt = A,
             ct=matrix(c(0,0), nrow=2), Zt = C,
            HHt = Sigma1, GGt = Sigma2,
             yt = rbind(dat$sh, dat$nh))
  return (-kf$logLik)
}

optimSol = optim(c(-0.4, -0.3, rep(log(0.01),6)),
                 fn=Optim.fun, method ="L-BFGS-B",
                  control=list(maxit=10000))
sigma01 <- exp(optimSol$par[3])
sigma02 <- exp(optimSol$par[4])
init1 <- optimSol$par[1]
init2 <- optimSol$par[2]
cor1 <- exp(optimSol$par[5])
cor2 <- exp(optimSol$par[6])
cor3 <- exp(optimSol$par[7])
cor4 <- exp(optimSol$par[8])
Sigma2 <- matrix(c(cor3,0,0,cor4),nrow=2)
Sigma1 <- matrix(c(cor1, 0,0,cor2), nrow=2)



kf2 <- fkf(a0=c(init1, init2),
           P0 = matrix(c(sigma01,0,0,sigma02),nrow=2),
                dt = matrix(c(0,0), nrow=2),
           Tt = A,
                ct=matrix(c(0,0), nrow=2),
           Zt = C, HHt = Sigma1, GGt = Sigma2,
                yt = cbind(rbind(dat$sh, dat$nh)))
kf2.pred <- fkf(a0=c(init1, init2),
                P0 = matrix(c(sigma01,0,0,sigma02),nrow=2),
           dt = matrix(c(0,0), nrow=2), Tt = A,
           ct=matrix(c(0,0), nrow=2),
           Zt = C, HHt = Sigma1, GGt = Sigma2,
           yt = cbind(rbind(dat$sh, dat$nh),
                      rbind(rep(NA,31),rep(NA,31))))
diagtool(kf2$vt[1,],100,"South")
diagtool(kf2$vt[2,],100,"North")
qqPlot(kf2$vt[1,], ylab="Standardized␣Residuals",
       main="QQ-plot␣of␣residuals␣in␣southern␣hemisphere
␣␣␣␣␣␣␣anomalies")
qqPlot(kf2$vt[2,], ylab="Standardized␣Residuals",
       main="QQ-plot␣of␣residuals␣in␣northern␣hemisphere
␣␣␣␣␣␣␣anomalies")
kf2$logLik
plot.kf.pred(kf2.pred, "south")
plot.kf.rec(kf2,"south")
```

```
plot.kf.pred(kf2.pred, "north")
plot.kf.rec(kf2,"north")
print.pred(kf2.pred)
##############Q4.5#################

Optim.fun2 = function(vars){
  var0 <- matrix(c(exp(vars[3]), 0,
                   0,exp(vars[4])), nrow=2)
  C <- matrix(c(1,0,0,1),nrow=2)
  A <- matrix(c(1,0,0,1),nrow = 2)
  Sigma2 <- matrix(c(exp(vars[8]),0,
                     0,exp(vars[9])),nrow=2)
  Sigma1 <- matrix(c(exp(vars[5]),
                     exp(vars[6]),exp(vars[6]),
                     exp(vars[7])), nrow=2)

  kf <- fkf(a0=c(vars[1], vars[2]), P0 = var0,
            dt = matrix(c(0,0), nrow=2), Tt = A,
             ct=matrix(c(0,0), nrow=2), Zt = C,
            HHt = Sigma1, GGt = Sigma2,
             yt = rbind(dat$sh, dat$nh))
  return (-kf$logLik)
}

optimSol = optim(c(-0.4, -0.3, rep(log(0.01),7)),
                 Optim.fun2,method="L-BFGS-B",
                 control=list(maxit=10000))


sigma01 <- exp(optimSol$par[3])
sigma02 <- exp(optimSol$par[4])
init1 <- optimSol$par[1]
init2 <- optimSol$par[2]
cor1 <- exp(optimSol$par[5])
cor2 <- exp(optimSol$par[6])
cor4 <- exp(optimSol$par[7])
eps1 <- exp(optimSol$par[8])
eps2 <- exp(optimSol$par[9])
var0 <- matrix(c(sigma01, 0,0,sigma02), nrow=2)
C <- matrix(c(1,0,0,1),nrow=2)
A <- matrix(c(1,0,0,1),nrow = 2)
Sigma2 <- matrix(c(eps1,0,0,eps2),nrow=2)
Sigma1 <- matrix(c(cor1,cor2,cor2,cor4), nrow=2)

kf3 <- fkf(a0=c(init1, init2), P0 = var0,
           dt = matrix(c(0,0), nrow=2),
           Tt = A,
                ct=matrix(c(0,0), nrow=2),
           Zt = C, HHt = Sigma1,
           GGt = Sigma2,
```

```
                  yt = cbind(rbind(dat$sh, dat$nh)))

kf3.pred <- fkf(a0=c(init1, init2),
                P0 = var0,
                dt = matrix(c(0,0), nrow=2),
                Tt = A,
            ct=matrix(c(0,0), nrow=2), Zt = C,
            HHt = Sigma1, GGt = Sigma2,
            yt = cbind(rbind(dat$sh, dat$nh),
                       rbind(rep(NA,31),rep(NA,31))))
par(mgp=c(2,0.8,0), mar=c(3,3,2,1))
diagtool(kf3$vt[1,],100,"South")
diagtool(kf3$vt[2,],100,"North")
qqPlot(kf3$vt[1,], ylab="Standardized␣Residuals",
       main="QQ-plot␣of␣residuals␣in␣southern␣hemisphere
␣␣␣␣␣␣␣anomalies")
qqPlot(kf3$vt[2,], ylab="Standardized␣Residuals",
       main="QQ-plot␣of␣residuals␣in␣northern␣hemisphere
␣␣␣␣␣␣␣anomalies")
kf3$logLik
plot.kf.pred(kf3.pred, "south")
plot.kf.pred(kf3.pred, "north")
plot.kf.rec(kf3,"south")
plot.kf.rec(kf3,"north")

print.pred(kf3.pred)

##############Q4.6#################

#SEE PAPER


##############Q4.7#################
A = matrix(c(1,0,0,0,1,0,1,1,1),nrow=3)
C = matrix(c(1,0,0,1,0,0), nrow=2)

Optim.fun3 = function(vars){
  kf <- fkf(a0=c(vars[1], vars[2], vars[3]),
            P0 = matrix(c(exp(vars[4]), 0,
                          0,0,exp(vars[5]),0,0
                          ,0,exp(vars[6])), nrow=3),
            dt = matrix(c(0,0,0), nrow=3),
            Tt = A,
            ct=matrix(c(0,0), nrow=2),
            Zt = C,
            HHt = matrix(c(exp(vars[7]),
                           exp(vars[8]),0,exp(vars[8]),
                           exp(vars[9]),0,0,0,
                           exp(vars[12])), nrow=3),
            GGt = matrix(c(exp(vars[10]), 0, 0,
```

```
                              exp(vars[11])), nrow=2),
              yt = rbind(dat$sh, dat$nh))
  return (-kf$logLik)
}
optimOut = optim(c(-0.4, -0.3, 0.0001,
                   rep(log(0.01),8),log(4e-06)),
                 Optim.fun3, method="L-BFGS-B",
                 control=list(maxit=10000))

kf4 <-  fkf(a0=c(optimOut$par[1],
                 optimOut$par[2], optimOut$par[3]),
                 P0 = matrix(c(exp(optimOut$par[4]), 0,
                               0,0,
                               exp(optimOut$par[5]),0,0,
                               0,
                               exp(optimOut$par[6])), nrow=3),
                 dt = matrix(c(0,0,0), nrow=3),
                 Tt = A,
                 ct=matrix(c(0,0), nrow=2),
                 Zt = C,
                 HHt = matrix(c(exp(optimOut$par[7]),
                                exp(optimOut$par[8]),0,
                                exp(optimOut$par[8]),
                                exp(optimOut$par[9]),0,0
                                ,0,exp(optimOut$par[12])),
                                nrow=3),
                 GGt = matrix(c(exp(optimOut$par[10]), 0, 0,
                                exp(optimOut$par[11])),
                                nrow=2),
                 yt = rbind(dat$sh, dat$nh))

kf4.pred <- fkf(a0=c(optimOut$par[1], optimOut$par[2],
                     optimOut$par[3]),
            P0 = matrix(c(exp(optimOut$par[4]), 0,
                          0,0,exp(optimOut$par[5]),0,0,
                          0,exp(optimOut$par[6])), nrow=3),
            dt = matrix(c(0,0,0), nrow=3),
            Tt = A,
            ct=matrix(c(0,0), nrow=2),
            Zt = C,
            HHt = matrix(c(exp(optimOut$par[7]),
                           exp(optimOut$par[8]),0,
                           exp(optimOut$par[8]),
                           exp(optimOut$par[9]),
                           0,0,0,exp(optimOut$par[12])),
                         nrow=3),
            GGt = matrix(c(exp(optimOut$par[10]), 0, 0,
                           exp(optimOut$par[11])), nrow=2),
            yt = cbind(rbind(dat$sh, dat$nh),
                       rbind(rep(NA,31),rep(NA,31))))
```

```
diagtool(kf4$vt[1,],100,"South")
diagtool(kf4$vt[2,],100,"North")
qqPlot(kf4$vt[1,], ylab="Standardized␣Residuals",
       main="QQ-plot␣of␣residuals␣in␣southern␣hemisphere
␣␣␣␣␣␣␣anomalies")
qqPlot(kf4$vt[2,], ylab="Standardized␣Residuals",
       main="QQ-plot␣of␣residuals␣in␣northern␣hemisphere
␣␣␣␣␣␣␣anomalies")
kf4$logLik
plot.kf.rec(kf4,"south")
plot.kf.rec(kf4,"north")

plot.kf.pred(kf4.pred, "south")
plot.kf.pred(kf4.pred, "north")

print.pred(kf4.pred)

ccf(kf4$vt[1,], kf4$vt[2,], lag=80,
main="CCF␣of␣the␣anomalies␣of␣the␣two␣hemispheres")

##############Q4.8#################
```