An automatic report for the dataset : 10-sulphuric

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

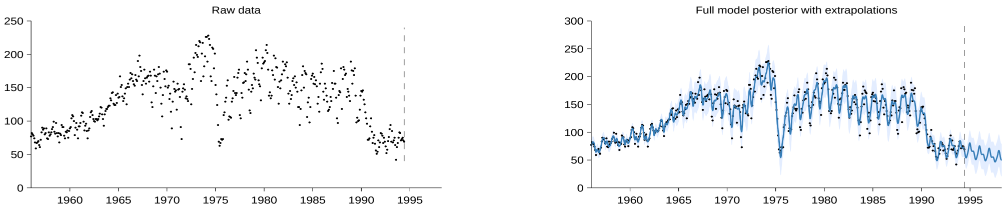

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified nine additive components in the data. The first 4 additive components explain 90.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 8 additive components explain 99.8% of the variation in the data. After the first 6 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A very smooth function.
- A constant. This function applies from 1964 until 1990.
- An approximately periodic function with a period of 1.0 years.
- A smooth function. This function applies from 1969 until 1977.
- A smooth function. This function applies from 1964 until 1969 and from 1977 onwards.
- An exactly periodic function with a period of 2.6 years. This function applies until 1964.
- Uncorrelated noise. This function applies until 1964.
- Uncorrelated noise. This function applies from 1964 until 1990.
- Uncorrelated noise. This function applies from 1990 onwards.

Model checking statistics are summarised in table 2 in section 4. These statistics have revealed statistically significant discrepancies between the data and model in component 1.

The rest of the document is structured as follows. In section 2 the forms of the additive components are described and their posterior distributions are displayed. In section 3 the modelling assumptions

#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	131.34	-
1	34.6	34.6	34.6	37.69	71.3
2	61.8	27.3	41.7	18.79	50.1
3	71.9	10.1	26.4	16.29	13.3
4	90.5	18.6	66.3	15.66	3.9
5	93.1	2.6	27.6	15.54	0.8
6	93.3	0.2	2.6	15.43	0.7
7	93.5	0.2	3.2	15.43	0.0
8	99.8	6.3	96.9	15.43	0.0
9	100.0	0.2	100.0	15.43	0.0

Table 1: Summary statistics for cumulative additive fits to the data. The residual coefficient of determination (R^2) values are computed using the residuals from the previous fit as the target values; this measures how much of the residual variance is explained by each new component. The mean absolute error (MAE) is calculated using 10 fold cross validation with a contiguous block design; this measures the ability of the model to interpolate and extrapolate over moderate distances. The model is fit using the full data and the MAE values are calculated using this model; this double use of data means that the MAE values cannot be used reliably as an estimate of out-of-sample predictive performance.

of each component are discussed with reference to how this affects the extrapolations made by the model. Section 4 discusses model checking statistics, with plots showing the form of any detected discrepancies between the model and observed data.

2 Detailed discussion of additive components

2.1 Component 1 : A very smooth function

This component is a very smooth function.

This component explains 34.6% of the total variance. The addition of this component reduces the cross validated MAE by 71.3% from 131.3 to 37.7.

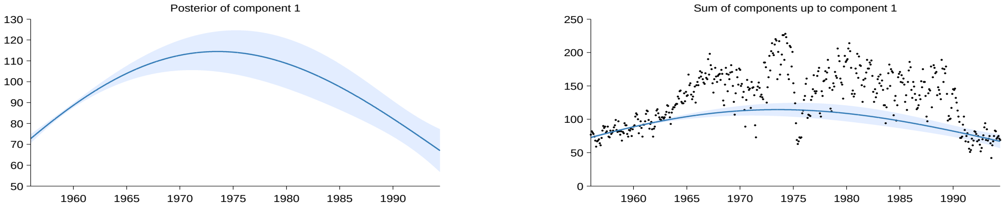

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)

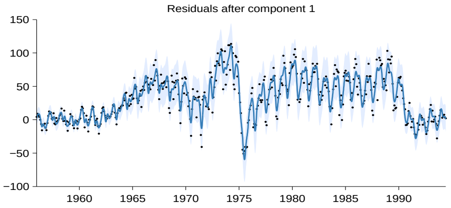

Figure 3: Pointwise posterior of residuals after adding component 1

2.2 Component 2 : A constant. This function applies from 1964 until 1990

This component is constant. This component applies from 1964 until 1990.

This component explains 41.7% of the residual variance; this increases the total variance explained from 34.6% to 61.8%. The addition of this component reduces the cross validated MAE by 50.15% from 37.69 to 18.79.

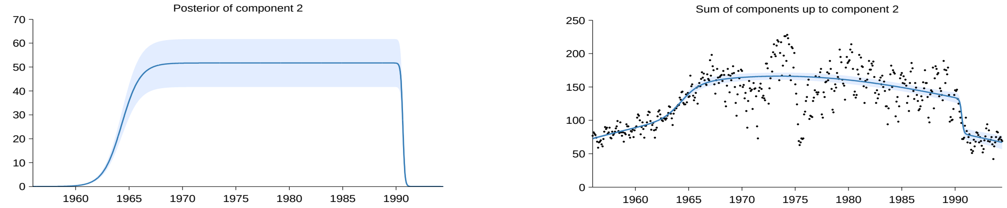

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

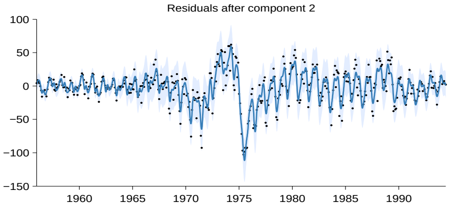

Figure 5: Pointwise posterior of residuals after adding component 2

2.3 Component 3 : An approximately periodic function with a period of 1.0 years

This component is approximately periodic with a period of 1.0 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 14.7 years. The shape of this function within each period has a typical lengthscale of 2.0 months.

This component explains 26.4% of the residual variance; this increases the total variance explained from 61.8% to 71.9%. The addition of this component reduces the cross validated MAE by 13.33% from 18.79 to 16.29.

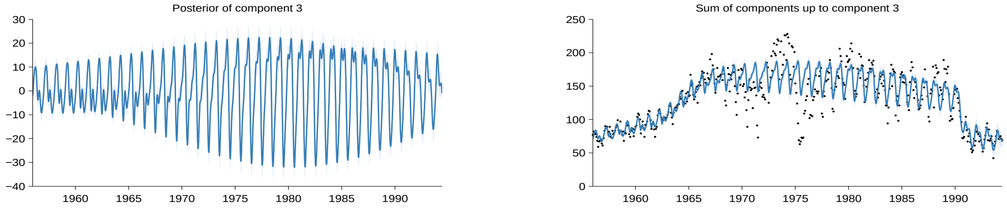

Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

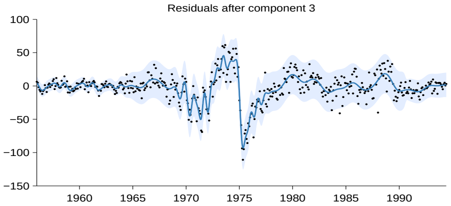

Figure 7: Pointwise posterior of residuals after adding component 3

2.4 Component 4 : A smooth function. This function applies from 1969 until 1977

This component is a smooth function with a typical lengthscale of 3.2 months. This component applies from 1969 until 1977.

This component explains 66.3% of the residual variance; this increases the total variance explained from 71.9% to 90.5%. The addition of this component reduces the cross validated MAE by 3.85% from 16.29 to 15.66.

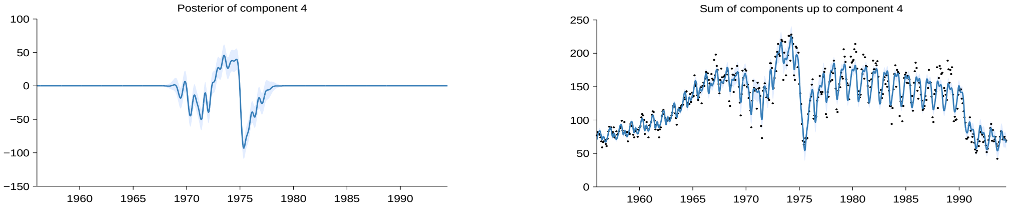

Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

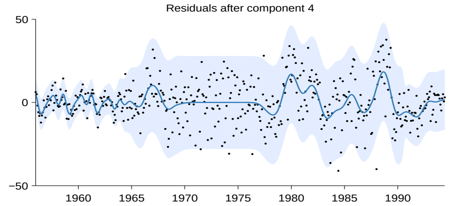

Figure 9: Pointwise posterior of residuals after adding component 4

2.5 Component 5 : A smooth function. This function applies from 1964 until 1969 and from 1977 onwards

This component is a smooth function with a typical lengthscale of 7.2 months. This component applies from 1964 until 1969 and from 1977 onwards.

This component explains 27.6% of the residual variance; this increases the total variance explained from 90.5% to 93.1%. The addition of this component reduces the cross validated MAE by 0.77% from 15.66 to 15.54.

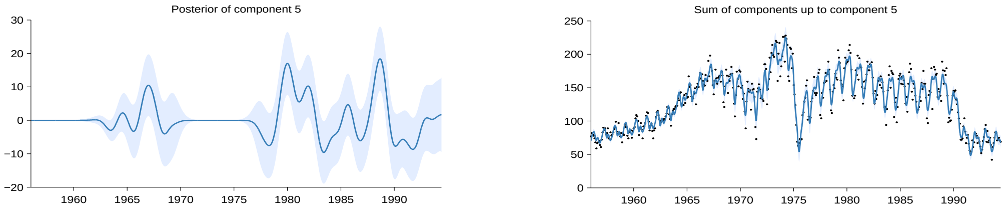

Figure 10: Pointwise posterior of component 5 (left) and the posterior of the cumulative sum of components with data (right)

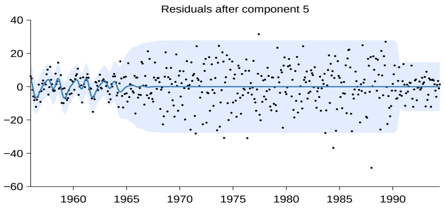

Figure 11: Pointwise posterior of residuals after adding component 5

2.6 Component 6 : An exactly periodic function with a period of 2.6 years. This function applies until 1964

This component is exactly periodic with a period of 2.6 years. The shape of this function within each period has a typical lengthscale of 3.6 months. This component applies until 1964.

This component explains 2.6% of the residual variance; this increases the total variance explained from 93.1% to 93.3%. The addition of this component reduces the cross validated MAE by 0.67% from 15.54 to 15.43.

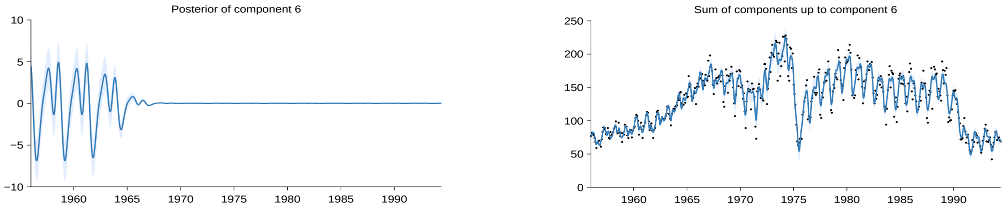

Figure 12: Pointwise posterior of component 6 (left) and the posterior of the cumulative sum of components with data (right)

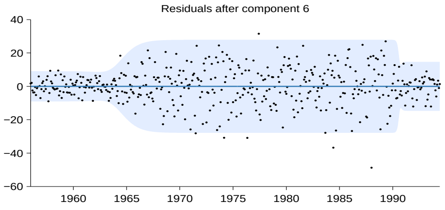

Figure 13: Pointwise posterior of residuals after adding component 6

2.7 Component 7 : Uncorrelated noise. This function applies until 1964

This component models uncorrelated noise. This component applies until 1964.

This component explains 3.2% of the residual variance; this increases the total variance explained from 93.3% to 93.5%. The addition of this component reduces the cross validated MAE by 0.00% from 15.43 to 15.43. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

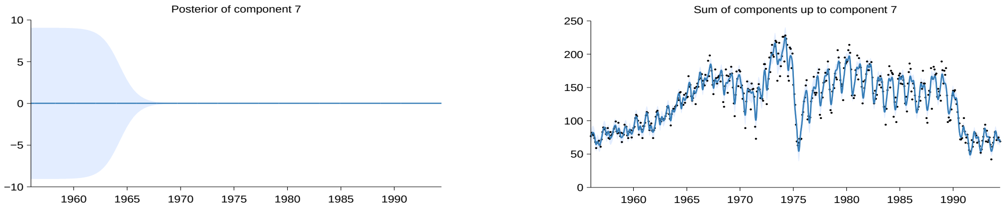

Figure 14: Pointwise posterior of component 7 (left) and the posterior of the cumulative sum of components with data (right)

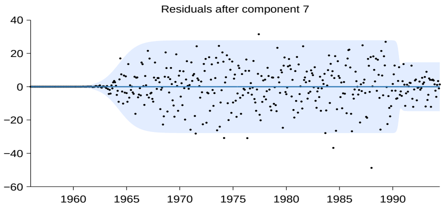

Figure 15: Pointwise posterior of residuals after adding component 7

2.8 Component 8 : Uncorrelated noise. This function applies from 1964 until 1990

This component models uncorrelated noise. This component applies from 1964 until 1990.

This component explains 96.9% of the residual variance; this increases the total variance explained from 93.5% to 99.8%. The addition of this component reduces the cross validated MAE by 0.00% from 15.43 to 15.43. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

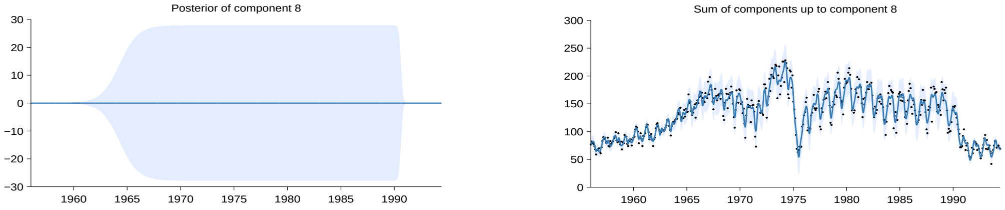

Figure 16: Pointwise posterior of component 8 (left) and the posterior of the cumulative sum of components with data (right)

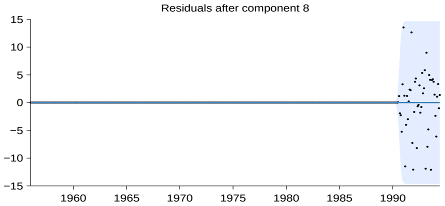

Figure 17: Pointwise posterior of residuals after adding component 8

2.9 Component 9 : Uncorrelated noise. This function applies from 1990 onwards

This component models uncorrelated noise. This component applies from 1990 onwards.

This component explains 100.0% of the residual variance; this increases the total variance explained from 99.8% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 15.43 to 15.43. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

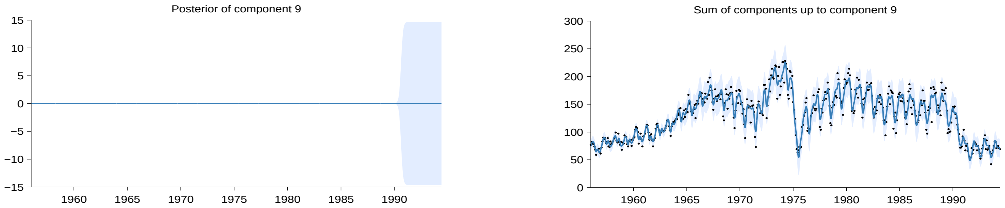

Figure 18: Pointwise posterior of component 9 (left) and the posterior of the cumulative sum of components with data (right)

3 Extrapolation

Summaries of the posterior distribution of the full model are shown in figure 19. The plot on the left displays the mean of the posterior together with pointwise variance. The plot on the right displays three random samples from the posterior.

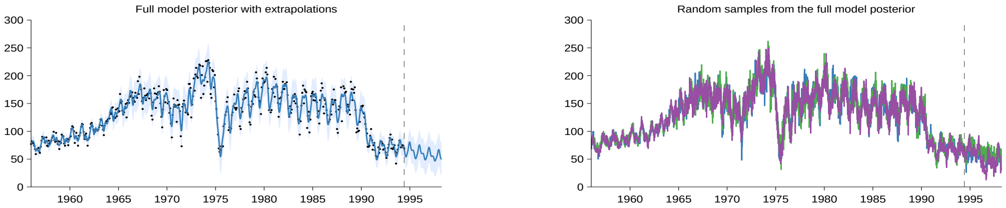

Figure 19: Full model posterior with extrapolation. Mean and pointwise variance (left) and three random samples (right)

Below are descriptions of the modelling assumptions associated with each additive component and how they affect the predictive posterior. Plots of the pointwise posterior and samples from the posterior are also presented, showing extrapolations from each component and the cuulative sum of components.

3.1 Component 1 : A very smooth function

This component is assumed to continue very smoothly but is also assumed to be stationary so its distribution will eventually return to the prior. The prior distribution places mass on smooth functions with a marginal mean of zero and a typical lengthscale of 27.2 years. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

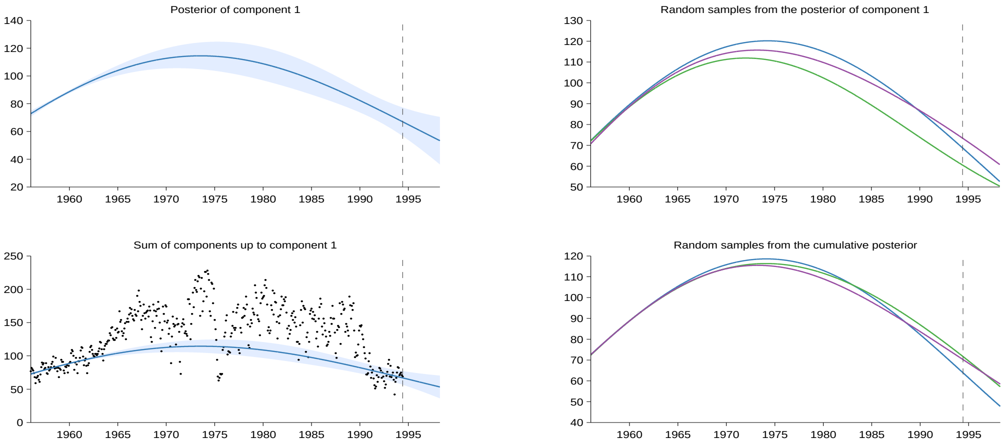

Figure 20: Posterior of component 1 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.2 Component 2 : A constant. This function applies from 1964 until 1990

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

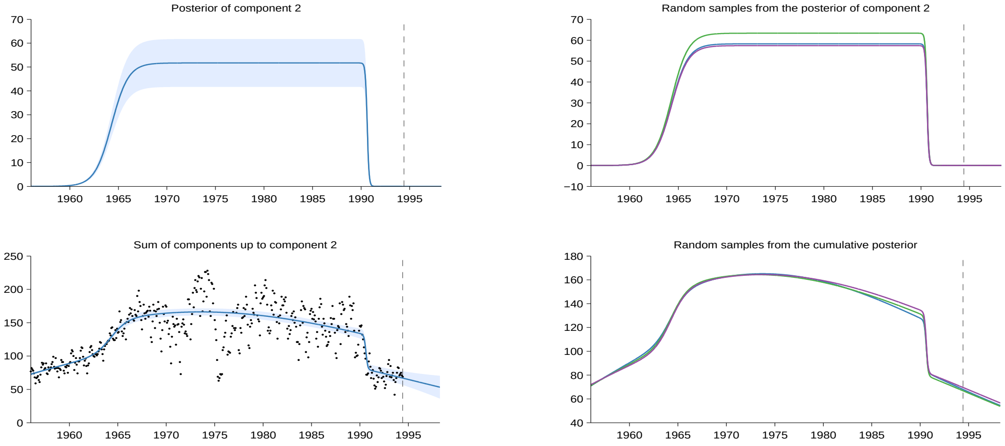

Figure 21: Posterior of component 2 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.3 Component 3 : An approximately periodic function with a period of 1.0 years

This component is assumed to continue to be approximately periodic. The shape of the function is assumed to vary smoothly between periods but will return to the prior. The prior is entirely uncertain about the phase of the periodic function. Consequently the pointwise posterior will appear to lose its periodicity, but this merely reflects the uncertainty in the shape and phase of the function. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

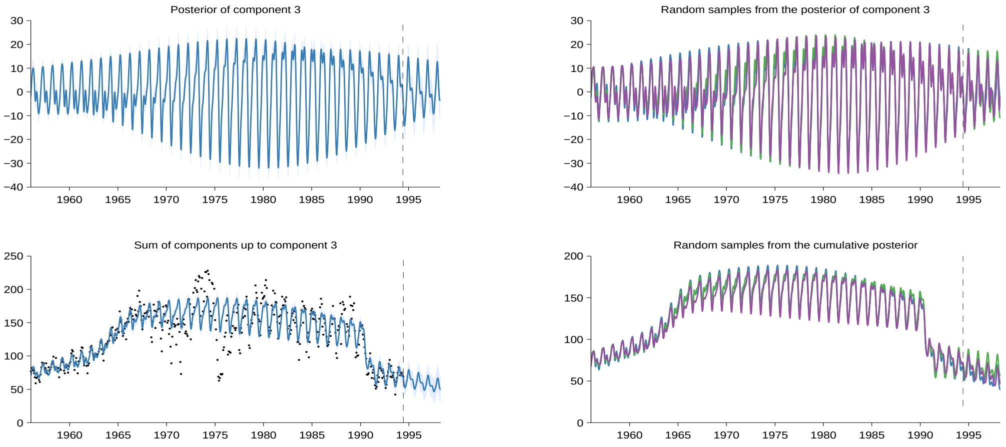

Figure 22: Posterior of component 3 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.4 Component 4 : A smooth function. This function applies from 1969 until 1977

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

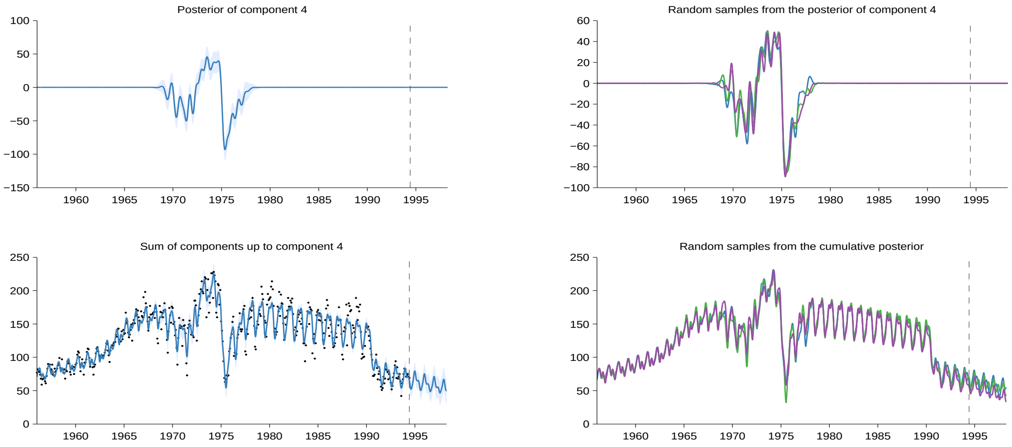

Figure 23: Posterior of component 4 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.5 Component 5 : A smooth function. This function applies from 1964 until 1969 and from 1977 onwards

This component is assumed to continue smoothly but is also assumed to be stationary so its distribution will return to the prior. The prior distribution places mass on smooth functions with a marginal mean of zero and a typical lengthscale of 7.2 months. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

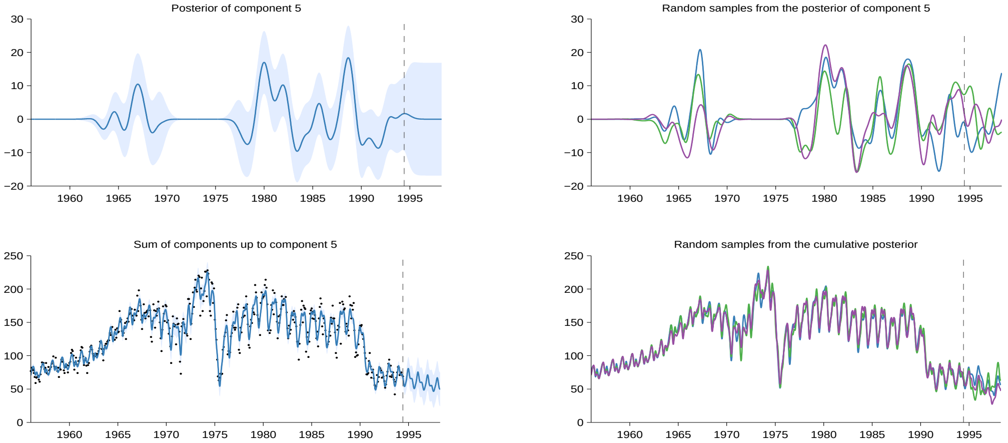

Figure 24: Posterior of component 5 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.6 Component 6 : An exactly periodic function with a period of 2.6 years. This function applies until 1964

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

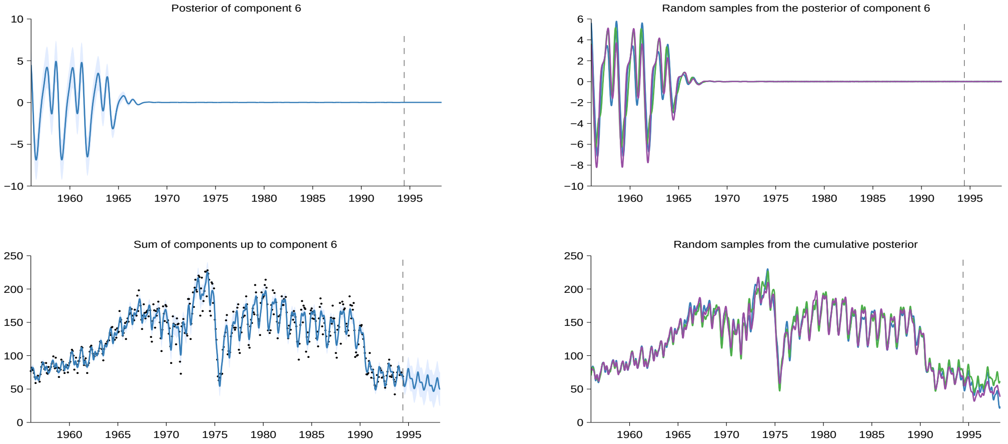

Figure 25: Posterior of component 6 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.7 Component 7 : Uncorrelated noise. This function applies until 1964

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

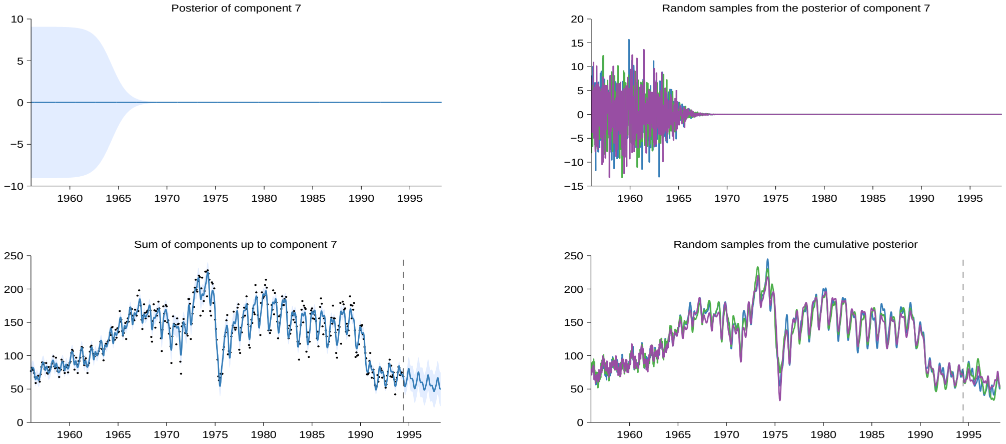

Figure 26: Posterior of component 7 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.8 Component 8 : Uncorrelated noise. This function applies from 1964 until 1990

This component is assumed to stop before the end of the data and will therefore be extrapolated as zero.

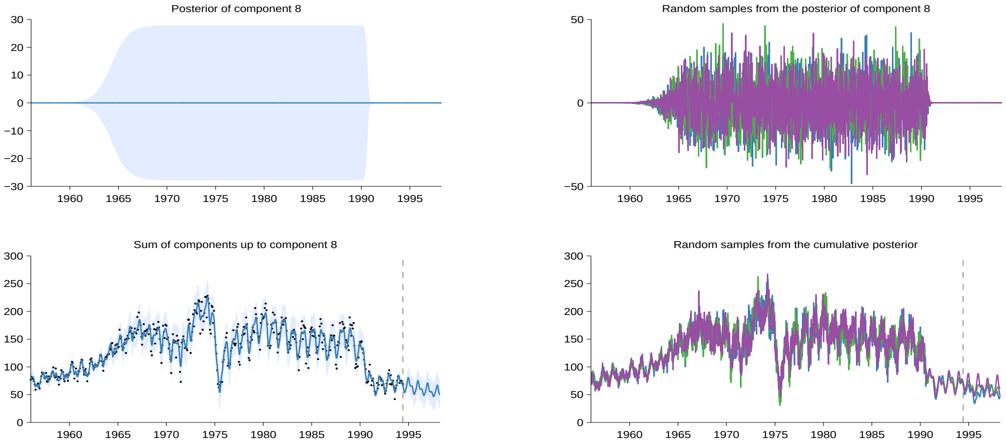

Figure 27: Posterior of component 8 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.9 Component 9 : Uncorrelated noise. This function applies from 1990 onwards

This component assumes the uncorrelated noise will continue indefinitely.

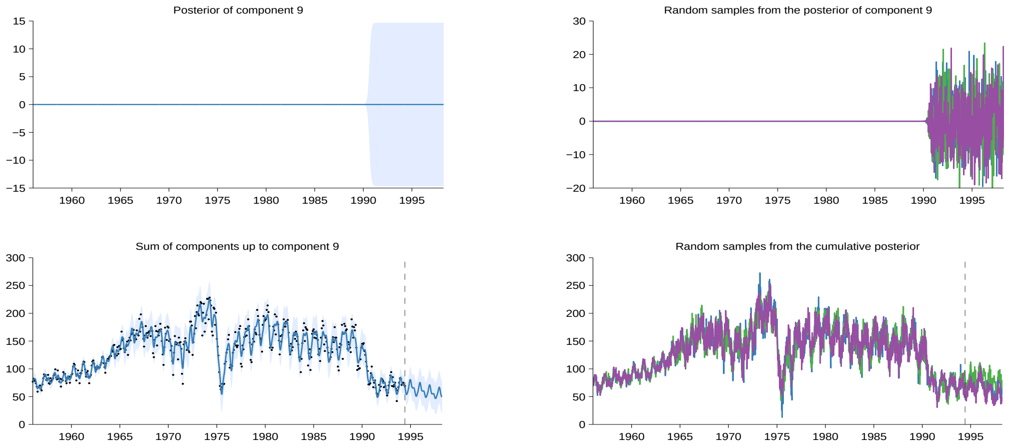

Figure 28: Posterior of component 9 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

4 Model checking

Several posterior predictive checks have been performed to assess how well the model describes the observed data. These tests take the form of comparing statistics evaluated on samples from the prior and posterior distributions for each additive component. The statistics are derived from autocorrelation function (ACF) estimates, periodograms and quantile-quantile (qq) plots.

Table 2 displays cumulative probability and p-value estimates for these quantities. Cumulative probabilities near 0/1 indicate that the test statistic was lower/higher under the posterior compared to the prior unexpectedly often i.e. they contain the same information as a p-value for a two-tailed test and they also express if the test statistic was higher or lower than expected. p-values near 0 indicate that the test statistic was larger in magnitude under the posterior compared to the prior unexpectedly often.

	ACF		Periodogram		QQ	
#	min	min loc	max	max loc	max	min
1	0.088	0.073	0.898	0.371	0.048	0.895
2	0.534	0.478	0.584	0.492	0.206	0.766
3	0.447	0.499	0.746	0.659	0.933	0.516
4	0.098	0.493	0.902	0.373	0.933	0.209
5	0.376	0.543	0.690	0.526	0.305	0.766
6	0.817	0.419	0.501	0.497	0.865	0.535
7	0.494	0.492	0.490	0.487	0.559	0.590
8	0.517	0.524	0.502	0.522	0.790	0.100
9	0.496	0.489	0.508	0.478	0.676	0.542

Table 2: Model checking statistics for each component. Cumulative probabilities for minimum of autocorrelation function (ACF) and its location. Cumulative probabilities for maximum of periodogram and its location. p-values for maximum and minimum deviations of QQ-plot from straight line.

The nature of any observed discrepancies is now described and plotted and hypotheses are given for the patterns in the data that may not be captured by the model.

4.1 Moderately statistically significant discrepancies

4.1.1 Component 1 : A very smooth function

The following discrepancies between the prior and posterior distributions for this component have been detected.

- The qq plot has an unexpectedly large positive deviation from equality ($x = y$). This discrepancy has an estimated p-value of 0.048.

The positive deviation in the qq-plot can indicate heavy positive tails if it occurs at the right of the plot or light negative tails if it occurs as the left.

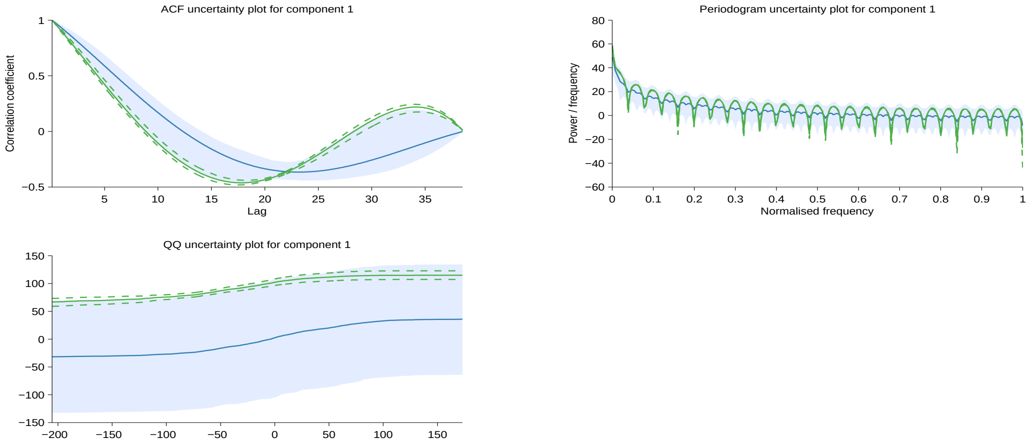

Figure 29: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 1. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2 Model checking plots for components without statistically significant discrepancies

4.2.1 Component 2 : A constant. This function applies from 1964 until 1990

No discrepancies between the prior and posterior of this component have been detected

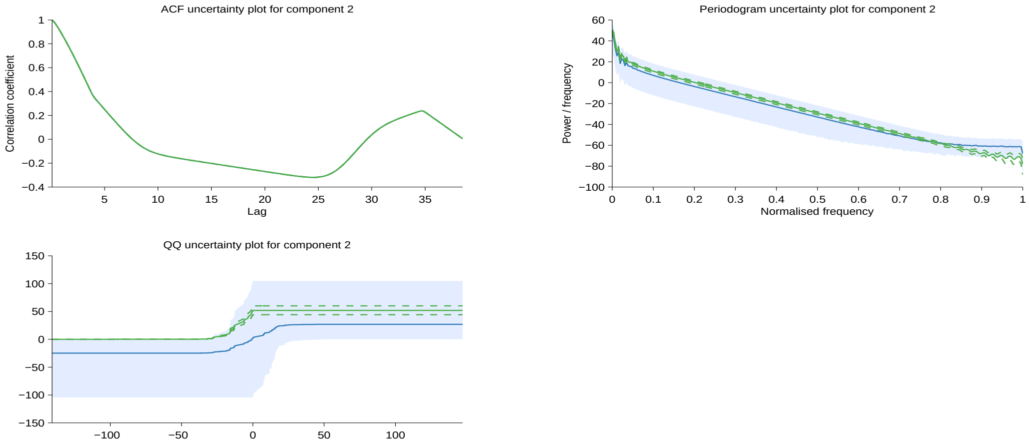

Figure 30: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 2. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.2 Component 3 : An approximately periodic function with a period of 1.0 years

No discrepancies between the prior and posterior of this component have been detected

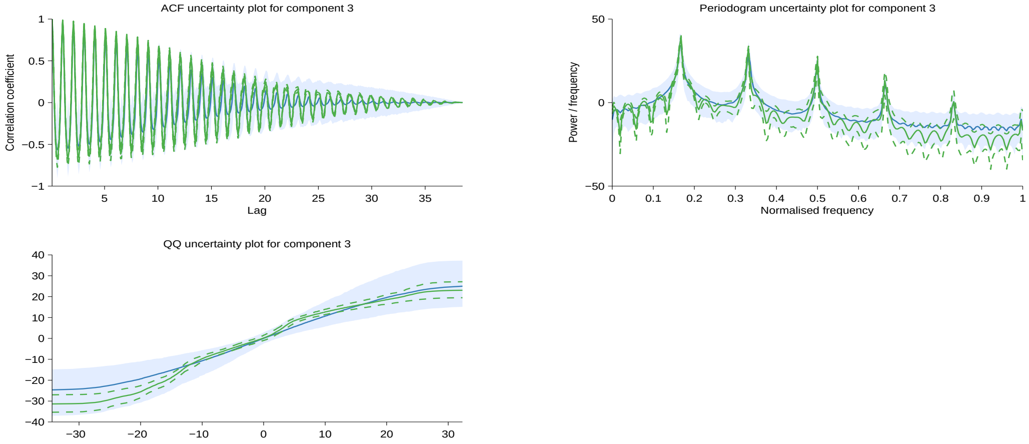

Figure 31: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 3. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.3 Component 4 : A smooth function. This function applies from 1969 until 1977

No discrepancies between the prior and posterior of this component have been detected

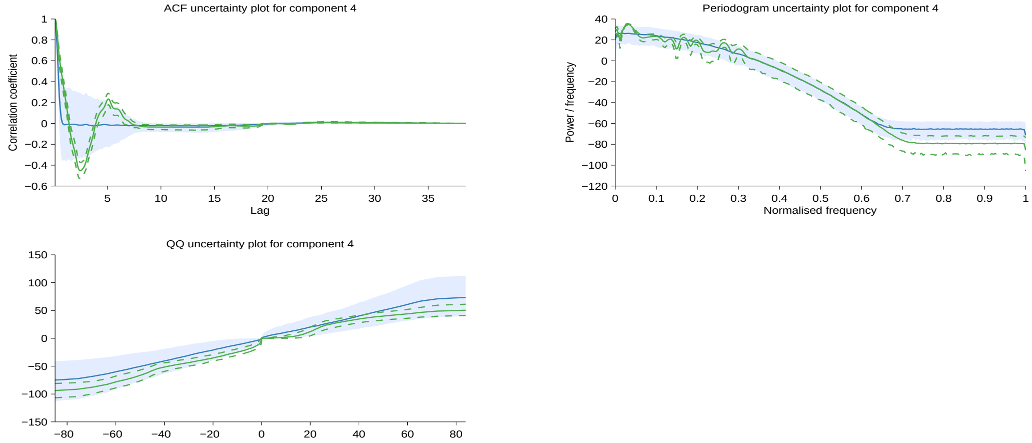

Figure 32: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 4. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.4 Component 5 : A smooth function. This function applies from 1964 until 1969 and from 1977 onwards

No discrepancies between the prior and posterior of this component have been detected

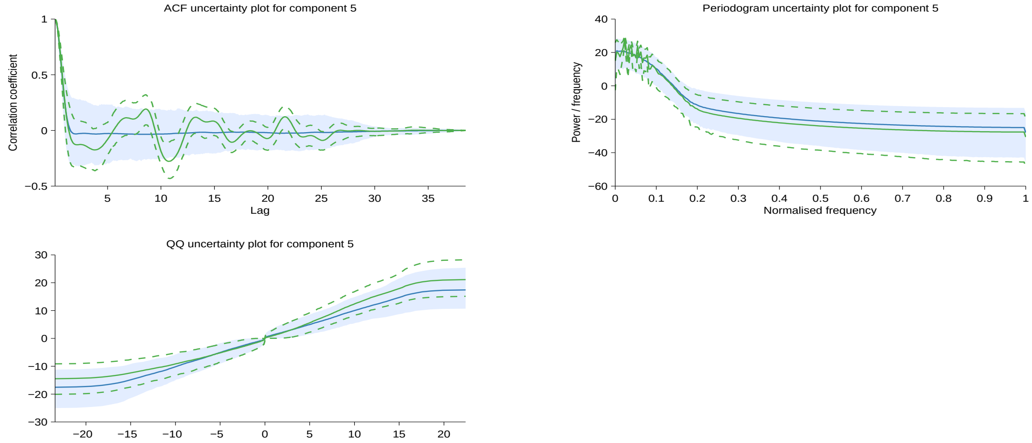

Figure 33: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 5. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.5 Component 6 : An exactly periodic function with a period of 2.6 years. This function applies until 1964

No discrepancies between the prior and posterior of this component have been detected

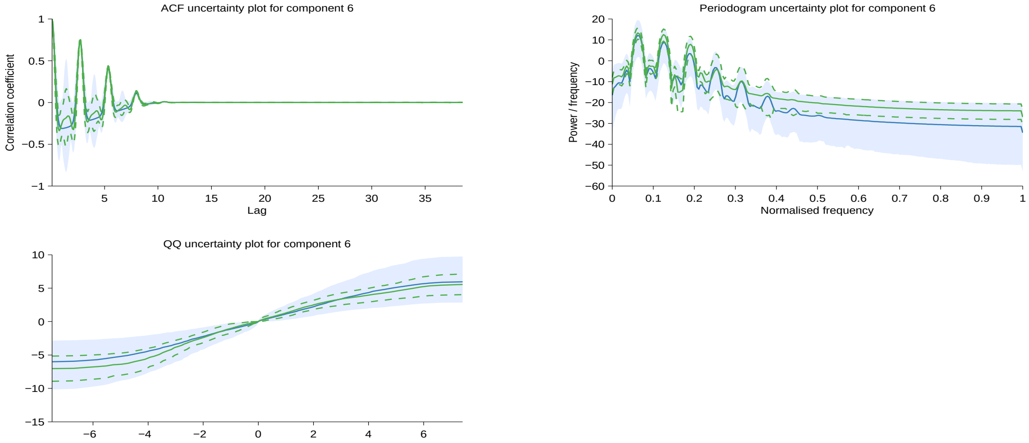

Figure 34: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 6. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.6 Component 7 : Uncorrelated noise. This function applies until 1964

No discrepancies between the prior and posterior of this component have been detected

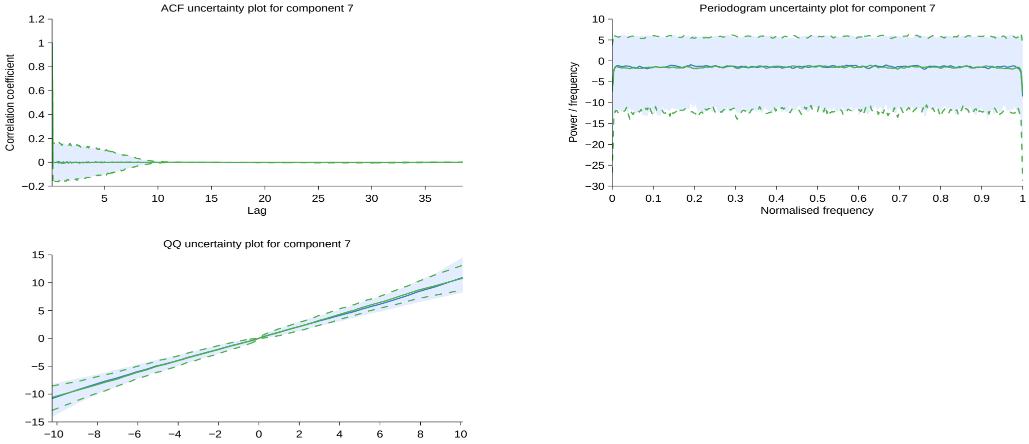

Figure 35: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 7. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.7 Component 8 : Uncorrelated noise. This function applies from 1964 until 1990

No discrepancies between the prior and posterior of this component have been detected

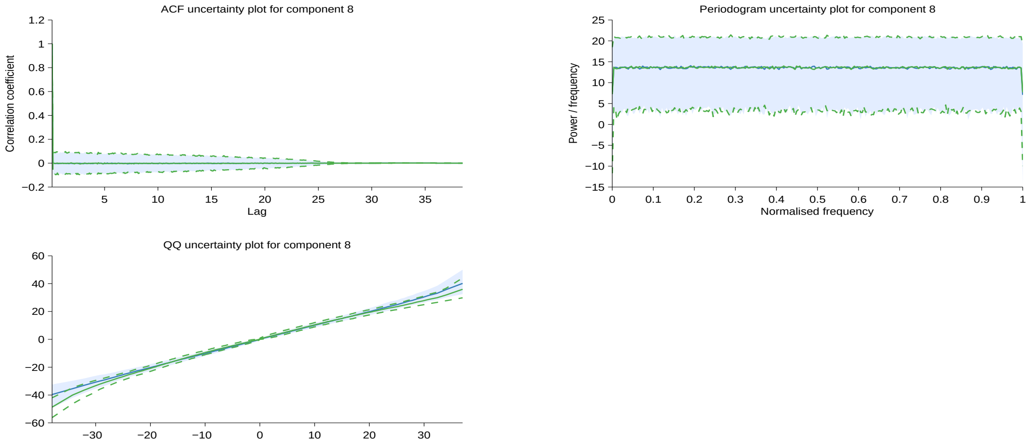

Figure 36: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 8. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.8 Component 9 : Uncorrelated noise. This function applies from 1990 onwards

No discrepancies between the prior and posterior of this component have been detected

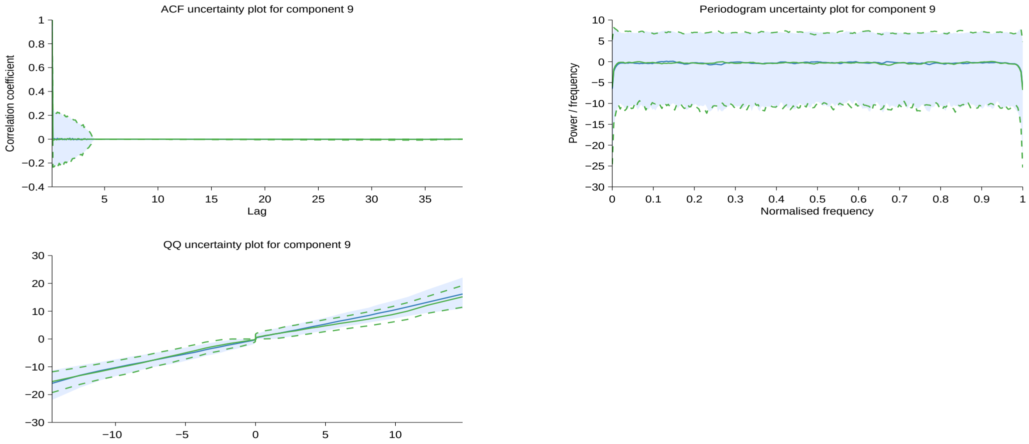

Figure 37: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 9. The green line and green dashed lines are the corresponding quantities under the posterior.