An automatic report for the dataset : daily-minimum-temperatures-in-me

James Robert Lloyd
University of Cambridge
Department of Engineering
jrl44@cam.ac.uk

David Duvenaud
University of Cambridge
Department of Engineering
dkd23@cam.ac.uk

Roger Grosse
M.I.T.
Brain and Cognitive Sciences
rgrosse@mit.edu

Joshua B. Tenenbaum
M.I.T.
Brain and Cognitive Sciences
jbt@mit.edu

Zoubin Ghahramani
University of Cambridge
Department of Engineering
zoubin@eng.cam.ac.uk

Abstract

This report was produced automatically by the Gaussian process structure search algorithm. See http://arxiv.org/abs/1302.4922 for a preliminary paper and see https://github.com/jamesrobertlloyd/gpss-research for the latest source code.

1 Executive summary

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

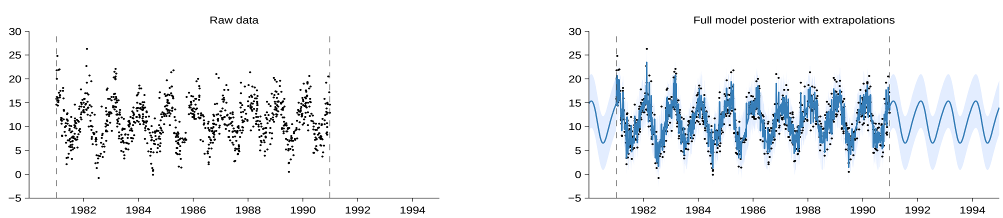

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified four additive components in the data. The first 3 additive components explain 97.3% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The 4 additive components explain 100.0% of the variation in the data. After the first 2 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 constant.
- An exactly periodic function with a period of 1.0 years.
- A rapidly varying smooth function.
- Uncorrelated noise.

#	R^2 (%)	ΔR^2 (%)	Residual R^2 (%)	Cross validated MAE	Reduction in MAE (%)
-	-	-	-	11.21	-
1	0.0	0.0	0.0	3.34	70.2
2	54.6	54.6	54.6	2.23	33.4
3	97.3	42.6	94.0	2.22	0.1
4	100.0	2.7	100.0	2.22	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.

Model checking statistics are summarised in table 2 in section 4. These statistics have revealed highly statistically significant discrepancies between the data and model in component 3.

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 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. A glossary of terms is provided in section B.

2 Detailed discussion of additive components

2.1 Component 1 : A constant

This component is constant.

This component explains 0.0% of the total variance. The addition of this component reduces the cross validated MAE by 70.2% from 11.2 to 3.3.

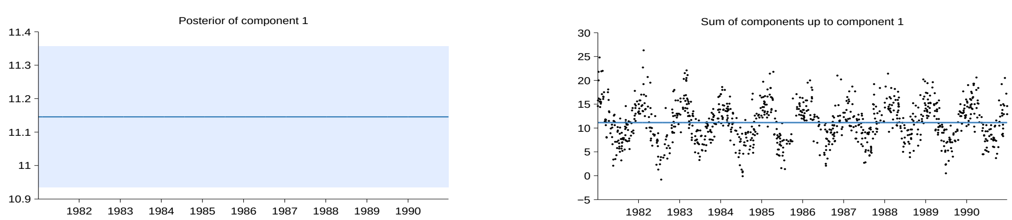

Figure 2: Pointwise posterior of component 1 (left) and the posterior of the cumulative sum of components with data (right)

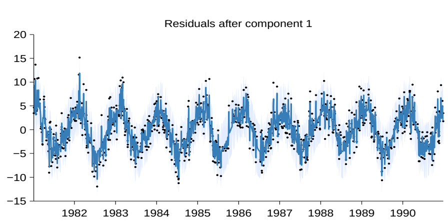

Figure 3: Pointwise posterior of residuals after adding component 1

2.2 Component 2 : An exactly periodic function with a period of 1.0 years

This component is exactly periodic with a period of 1.0 years. The shape of this function within each period has a typical lengthscale of 8.1 months.

This component explains 54.6% of the residual variance; this increases the total variance explained from 0.0% to 54.6%. The addition of this component reduces the cross validated MAE by 33.42% from 3.34 to 2.23.

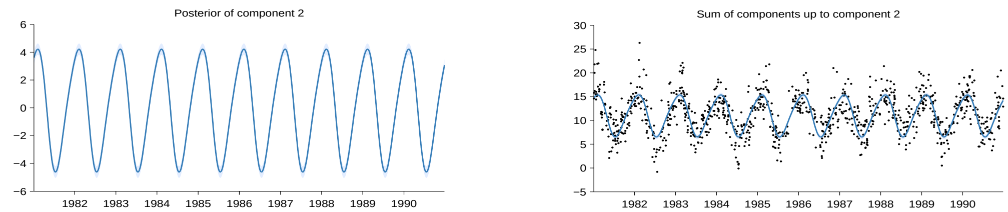

Figure 4: Pointwise posterior of component 2 (left) and the posterior of the cumulative sum of components with data (right)

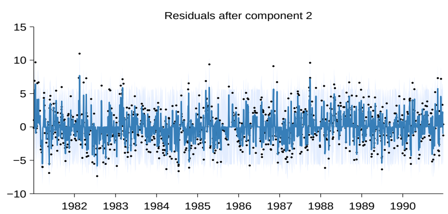

Figure 5: Pointwise posterior of residuals after adding component 2

2.3 Component 3 : A rapidly varying smooth function

This component is a rapidly varying but smooth function with a typical lengthscale of 30.9 hours.

This component explains 94.0% of the residual variance; this increases the total variance explained from 54.6% to 97.3%. The addition of this component reduces the cross validated MAE by 0.07% from 2.23 to 2.22. 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.

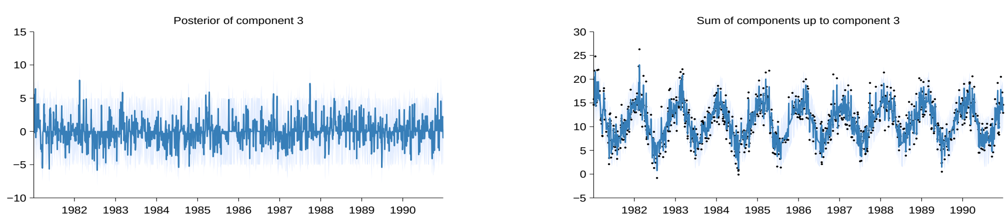

Figure 6: Pointwise posterior of component 3 (left) and the posterior of the cumulative sum of components with data (right)

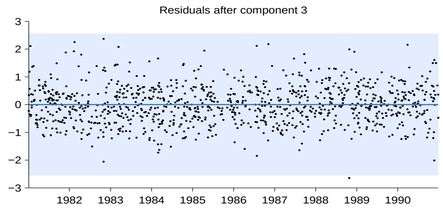

Figure 7: Pointwise posterior of residuals after adding component 3

2.4 Component 4 : Uncorrelated noise

This component models uncorrelated noise.

This component explains 100.0% of the residual variance; this increases the total variance explained from 97.3% to 100.0%. The addition of this component reduces the cross validated MAE by 0.00% from 2.22 to 2.22. 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.

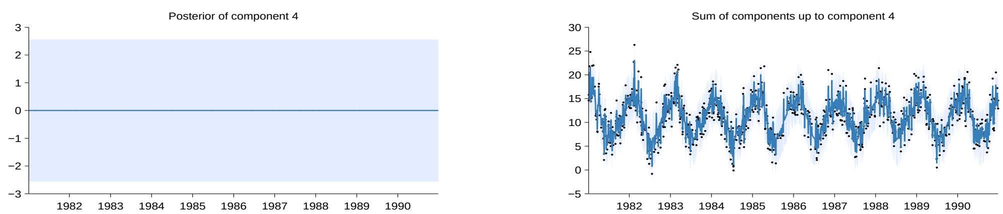

Figure 8: Pointwise posterior of component 4 (left) and the posterior of the cumulative sum of components with data (right)

3 Extrapolation

Summaries of the posterior distribution of the full model are shown in figure 9. 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.

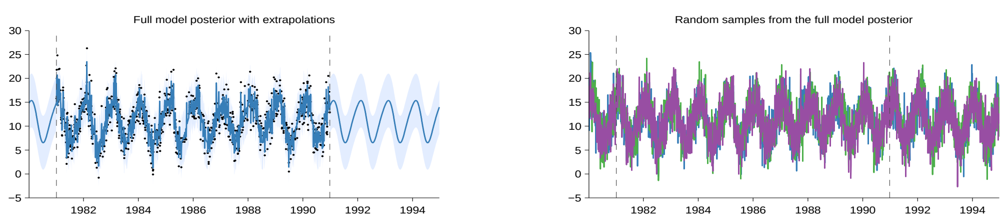

Figure 9: Full model posterior with extrapolation. Mean and pointwise variance (left) and three random samples (right)

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 constant

This component is assumed to stay constant.

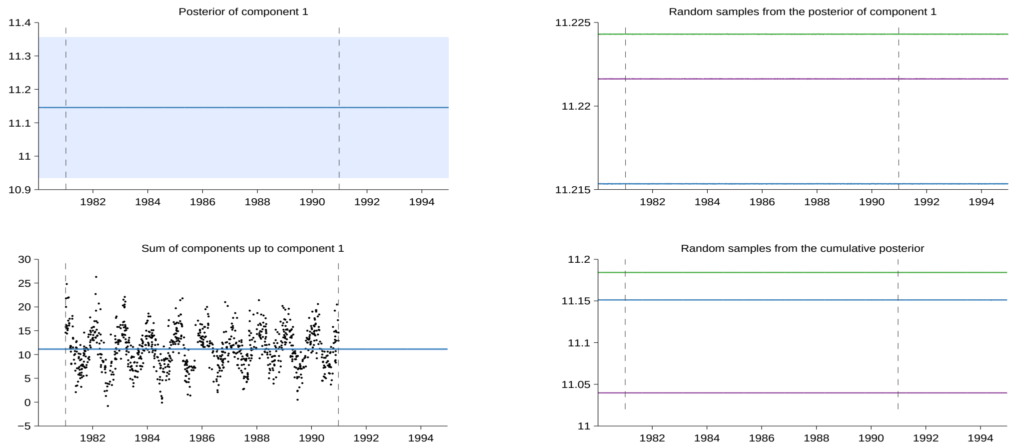

Figure 10: Posterior of component 1 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.2 Component 2 : An exactly periodic function with a period of 1.0 years

This component is assumed to continue exactly periodically with a period of 1.0 years.

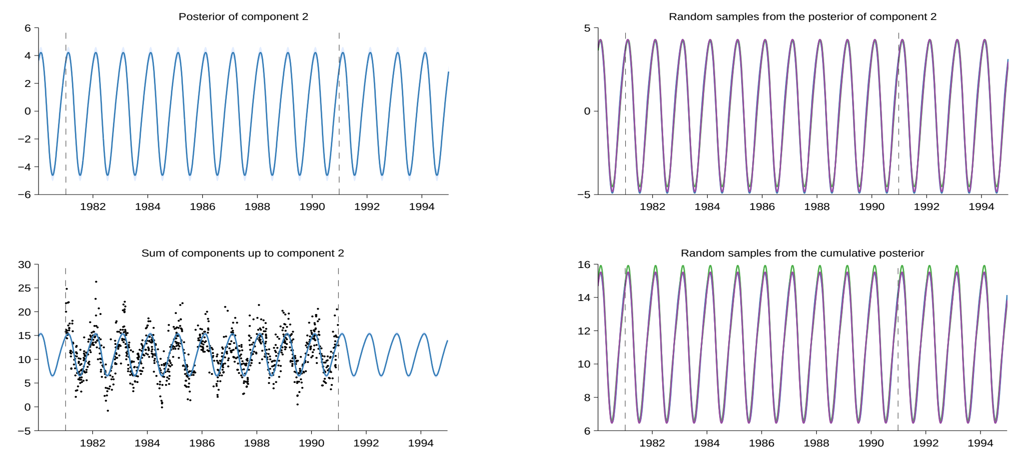

Figure 11: Posterior of component 2 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.3 Component 3 : A rapidly varying smooth function

This component is assumed to continue smoothly but its distribution is assumed to quickly return to the prior. The prior distribution places mass on smooth functions with a marginal mean of zero and a typical lengthscale of 30.9 hours. [This is a placeholder for a description of how quickly the posterior will start to resemble the prior].

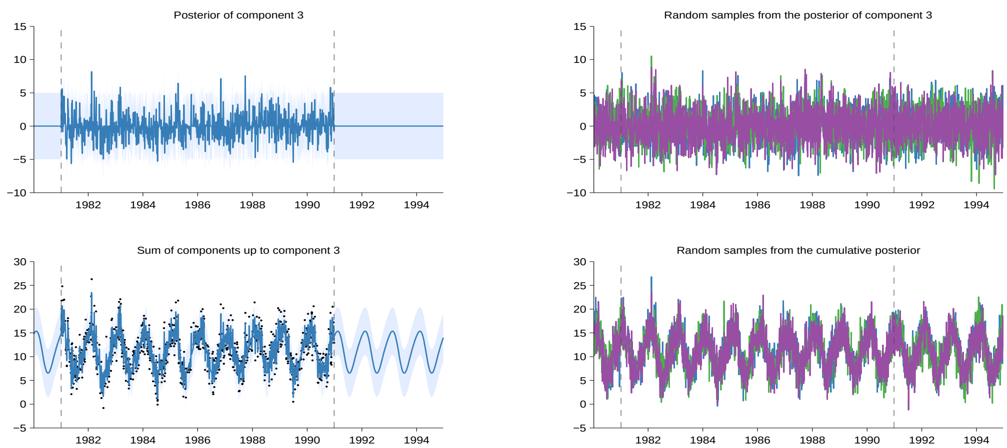

Figure 12: Posterior of component 3 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

3.4 Component 4 : Uncorrelated noise

This component assumes the uncorrelated noise will continue indefinitely.

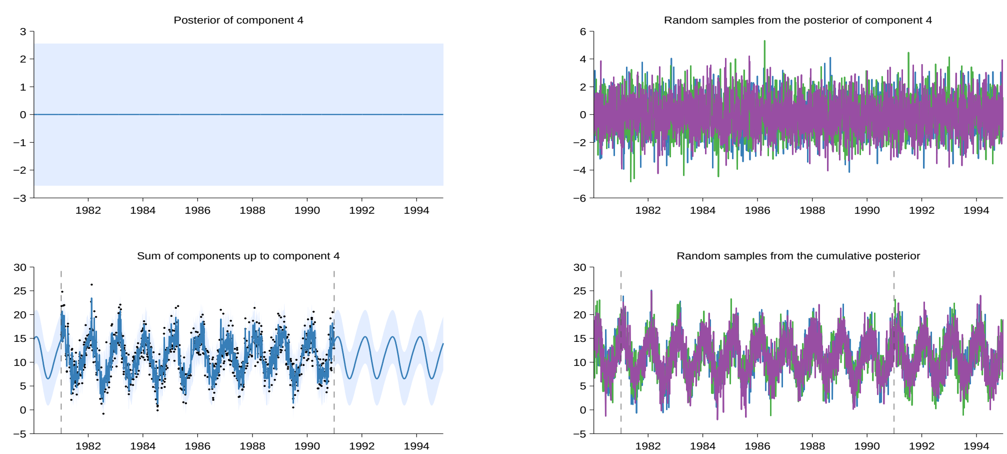

Figure 13: Posterior of component 4 (top) and cumulative sum of components (bottom) with extrapolation. Mean and pointwise variance (left) and three random samples from the posterior distribution (right).

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	0.499	0.485	0.724	0.510	0.150	0.850
2	0.444	0.525	0.443	0.511	0.575	0.506
3	**0.000**	0.219	**1.000**	0.016	0.096	0.832
4	0.233	0.383	0.962	0.028	0.427	0.555

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 Highly statistically significant discrepancies

4.1.1 Component 3 : A rapidly varying smooth function

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

- The minimum value of the ACF is unexpectedly low. This discrepancy has an estimated p-value of **0.000**.
- The maximum value of the periodogram is unexpectedly high. This discrepancy has an estimated p-value of **0.000**.
- The frequency of the maximum value of the periodogram is unexpectedly low. This discrepancy has an estimated p-value of 0.032.

The large maximum value of the periodogram can indicate periodicity that is not being captured by the model.

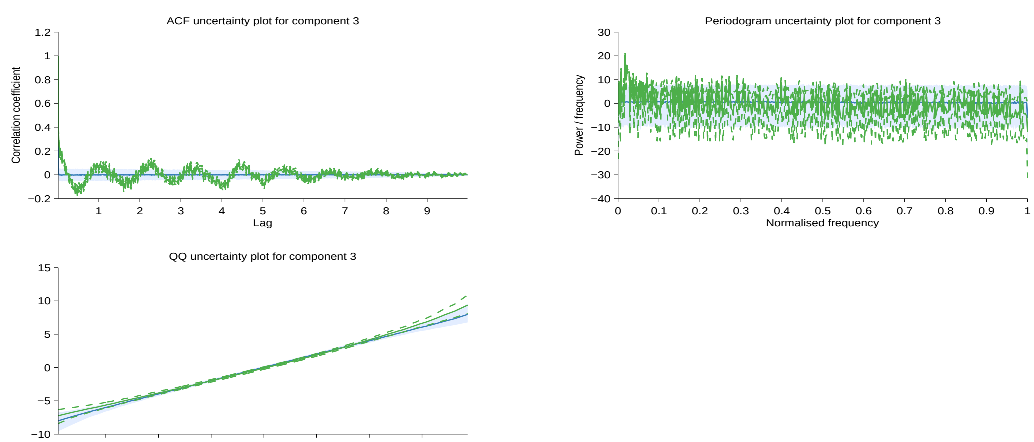

Figure 14: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 3. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2 Model checking plots for components without statistically significant discrepancies

4.2.1 Component 1 : A constant

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

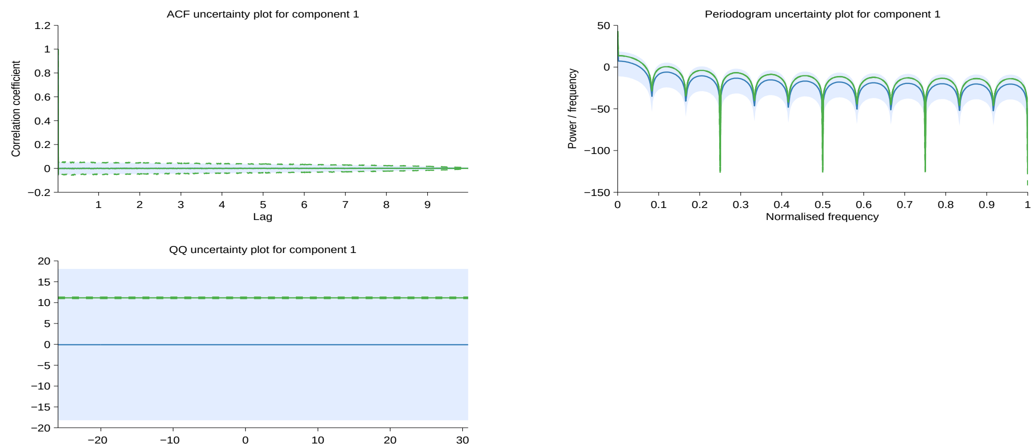

Figure 15: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 1. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.2 Component 2 : An exactly periodic function with a period of 1.0 years

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

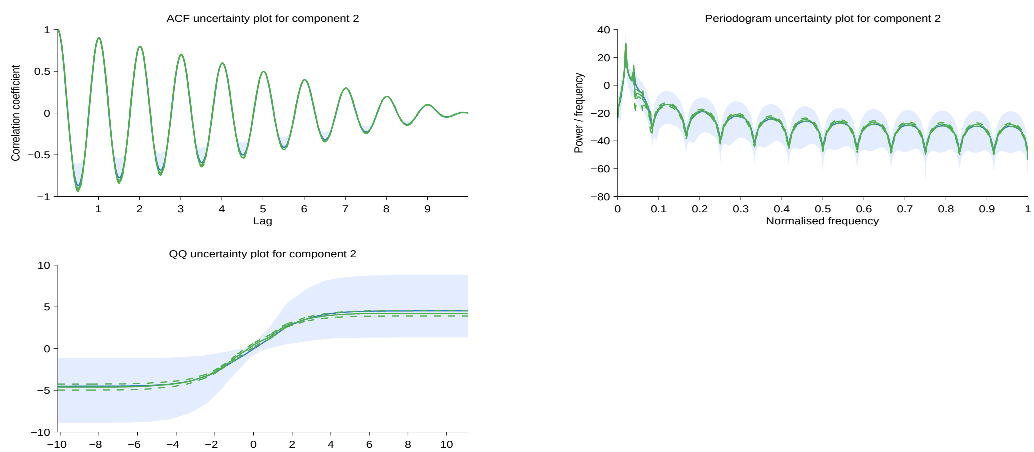

Figure 16: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 2. The green line and green dashed lines are the corresponding quantities under the posterior.

4.2.3 Component 4 : Uncorrelated noise

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

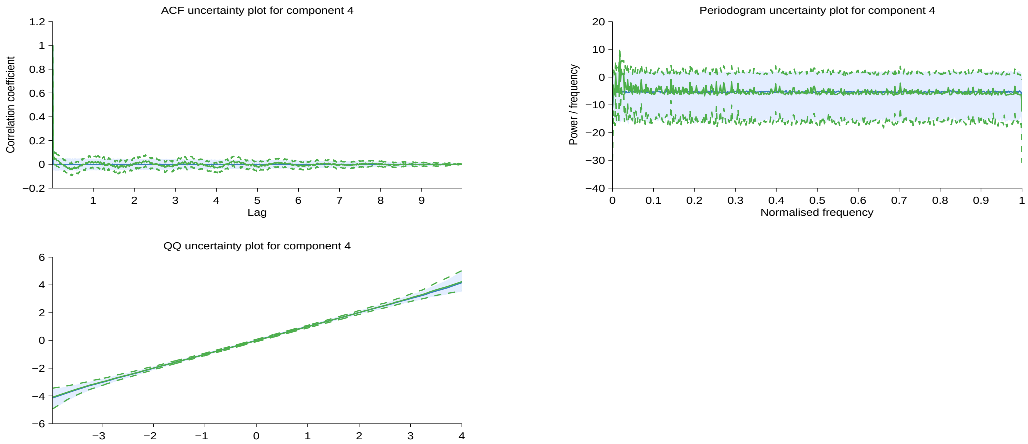

Figure 17: ACF (top left), periodogram (top right) and quantile-quantile (bottom left) uncertainty plots. The blue line and shading are the pointwise mean and 90% confidence interval of the plots under the prior distribution for component 4. The green line and green dashed lines are the corresponding quantities under the posterior.

A Residual style quantities

This appendix contains plots of residual-like quantities. Their utility is still being investigated so there are currently no explanations of their calculation or interpretation.

A.1 Leave one out

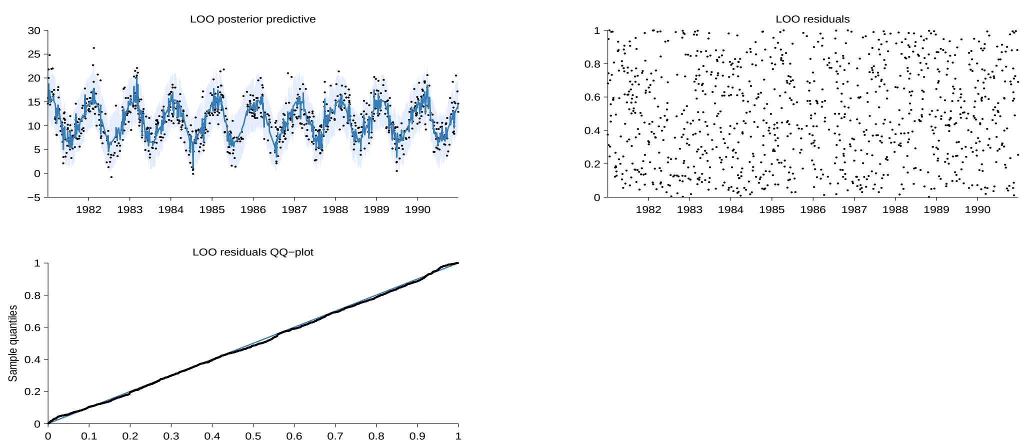

Figure 18: LOO posterior predictive. Distribution (left), standardised residuals (right) and qq-plot (below)

A.2 Leave chunk out

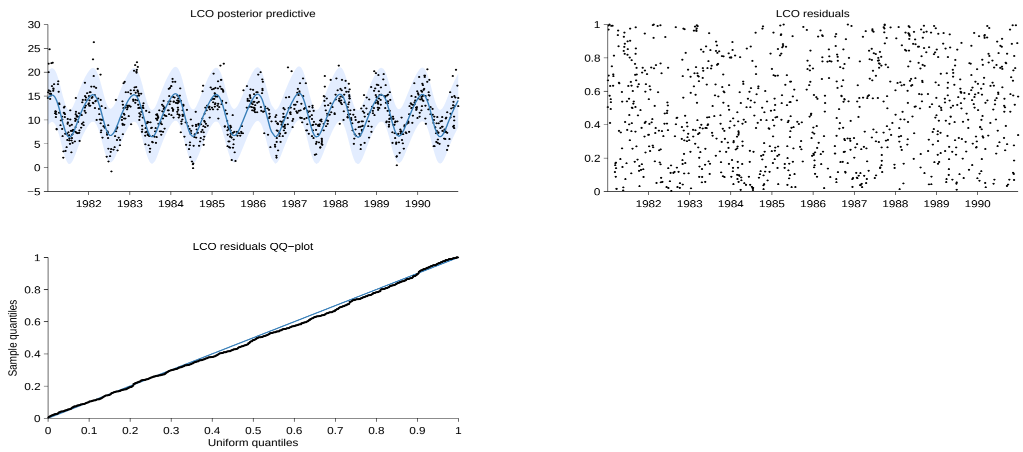

Figure 19: LCO posterior predictive. Distribution (left), standardised residuals (right) and qq-plot (below)

A.3 Next data point

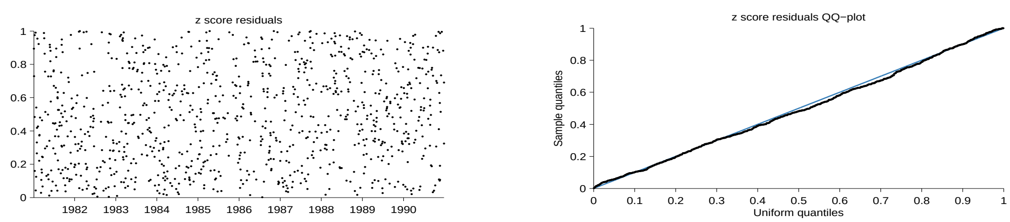

Figure 20: Inverse Cholesky thing. Standardised values (left) and qq-plot (right)

B Glossary of terms

- *lengthscale* - A description of what a lengthscale is