Implementing a Bayesian workflow on historical temperature data to model seasonality

[redacted]

June 26, 2024

Abstract

The seasonal signal of temperature data is challenging to model due to internal climate variability and complex, non-linear patterns. In this study we explore how a probabilistic Bayesian workflow can be used to successfully model the periodic seasonal temperature signal. To this end, a first-order approximation of the Fourier Series is fitted to historical temperature data from Berlin-Dahlem. Two different approaches with respect to the observation data are explored. The first approach uses a 4-year time series of monthly temperature data with one standard deviation calculated from daily temperature observations. The second approach considers 30-year averages for each day of the year (1950-1980). Using the two Fourier coefficients and a shift parameter, we find that the mode of the temperature signal is well captured by the model in both approaches. Furthermore, we explore how the choice of free vs. fixed parameters affects the modelling results, showing that model performance decreases when non-linear parameters are included as free parameters in the model.

1 Introduction

Bayesian data analysis is a powerful tool because it allows to account for uncertainties in both observations and model parameters, integrate prior knowledge, and provide the flexibility to update beliefs as new data becomes available. Therefore, it is well-suited to explore complex natural processes like the seasonal temperature cycle where some prior knowledge is available, e.g. with respect to the periodicity of the cycle, and where an understanding of associated uncertainty ranges is important. Since the seasonal temperature cycle displays a periodic pattern, it can be modelled as a Fourier Series

$$s(x) = \sum_{n=1}^{N}\Big[a_n\cos(\omega_n x+\phi)+b_n\sin(\omega_n x+\phi)\Big]+t \tag{1}$$

where a_n and b_n denote the Fourier coefficients, ω_n the frequency and ϕ the phase shift. The Bayesian workflow consists of three steps: Building a probabilistic model, fitting the model to the data and checking the model. In the following, a probabilistic model based on Equation 1 will be proposed and fit to historical temperature data for Berlin-Dahlem. Subsequently, the model variation and performance will be evaluated by generating uncertainty ranges for parameter values.

2 Methods

Historical daily temperature data for Berlin-Dahlem (weather station at Freie Universitaet Berlin) from 1950 to 2022 was downloaded from the German weather service [1]. Since no error estimates are provided for this data two approaches were explored to create time series with plausible uncertainty ranges. To simplify the analysis, only subsets of the original 72-year time series are used in the following.

For the first approach, we extracted a four year cycle from the dataset spanning the years 1950-1954, calculated monthly temperature averages and used one standard deviation to estimate the uncertainty range. In the second approach 30-years of daily temperature data are used to compute daily averages for each day of the year and corresponding standard deviations (1950-1980). This approach is based on the commonly used notion that 30-year averages can be considered a *climatology* [2].

To set up the probabilistic model, a first-order approximation of Equation 1 is used for the two approaches

$$\mu(a_0, a_1, \omega, \phi, t, x) = \Big[a_0\cos(\omega x+\phi)+a_1\sin(\omega x+\phi)\Big]+t \tag{2}$$

which comprises only one summand of the Fourier Series. Since ω describes the frequency it was calculated as

$$\omega = \frac{2\pi}{P} \tag{3}$$

where P is the period of the seasonal cycle, 12 in the case of monthly data and 365 in the case of the daily data giving values of 0.017 and 0.524 respectively. To begin with ϕ was also treated as a fixed parameter with value zero. In addition, it is assumed that the data y_i are Gaussian distributed around the model

$$y_i \sim \mathcal{N}\left(\mu(a_0, a_1, \omega, \phi, t, x_i), \sigma_y^2\right) \tag{4}$$

Furthermore, priors are estimated for each parameter. For parameter t it is expected that it should be close to the yearly mean temperature of Berlin, hence a prior of

$$t \sim \mathcal{N}(12, 2) \tag{5}$$

is picked. In a similar fashion, priors are defined for the other remaining parameters.

$$a_0, a_1 \sim \mathcal{N}(-6, 5) \tag{6}$$

In these cases μ is a somewhat arbitrary choice due to a lack of prior knowledge about these parameters and hence a slightly larger variance of 5 is assumed. To fit the model, the maximum a-posteriori (MAP) values for each parameter are computed

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|d) \tag{7}$$

with $p(\theta|d)$ the posterior probability distribution which describes the probability of θ after the data d has been observed. These MAP values give the mode of the posterior. To also obtain an estimate regarding the uncertainty ranges, the posterior distribution is sampled using the Markov-Chain Monte Carlo (MCMC) implementation of the emcee python package [3]. The number of walkers was set to $n_{walkers} = 2 \cdot n_{parameters} + 1$ which equates to seven in the case of three free parameters. The number of steps for each walker was set to 7000. Around three times the auto-correlation time initial steps were discarded as burn-in and the chains were thinned by sampling every half auto-correlation time steps.

In a last step, the quality of the model is examined by random sampling of the posterior distribution MCMC chains. First a predictive distribution for the model was created by randomly sampling parameter values from the MCMC chains to generate a "new" model. Following this, Gaussian noise is added to the sampled models by adding random noise. Results at both stages are visualised by quantile ranges and are used to evaluate to what extent data generated from this distribution agrees with the observations.

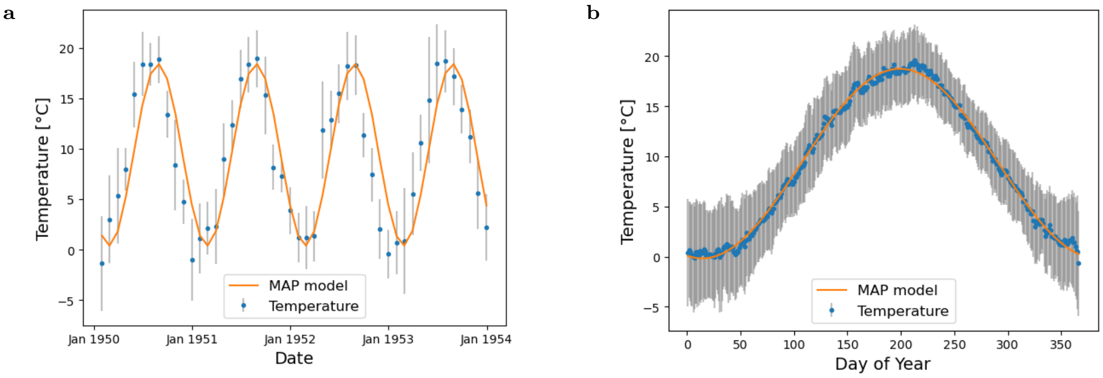

Figure 1: MAP results (orange) fitted to temperature data (blue) and error bars (grey) derived from one standard deviation. Panel 1a shows monthly means for the years 1950-1954 while Panel 1b shows daily means from a 30-year period (1950-1980).

parameter	a_0	a_1	t
Monthly data	-7.594	-4.876	9.409
Yearly data	-2.278	-9.179	9.277

Table 1: MAP parameters.

3 Results and Discussion

Figure 1 shows the results of the fitted model to the data using the MAP values. Qualitatively it can be seen that for both approaches a reasonable fit can be achieved, though for the monthly data there appears to be a slight phase shift. Table 1 lists the corresponding MAP values. While both approaches seem to arrive at similar values for t, the values of a_0 and a_1 differ by an order of a half between the two approaches.

The chains obtained from running MCMC can be seen in Figure 2. Auto-correlation times calculated for all the parameters ranged between 30 and 60. Using trimmed and thinned chains as described in the methods, parameter values could be calculated with an estimate of the standard deviation, these can be seen in Table 2. For all parameters there is minimal uncertainty around the mean, which is in good agreement with the MAP parameters. This can also be seen in the corner plots in Figure 3.

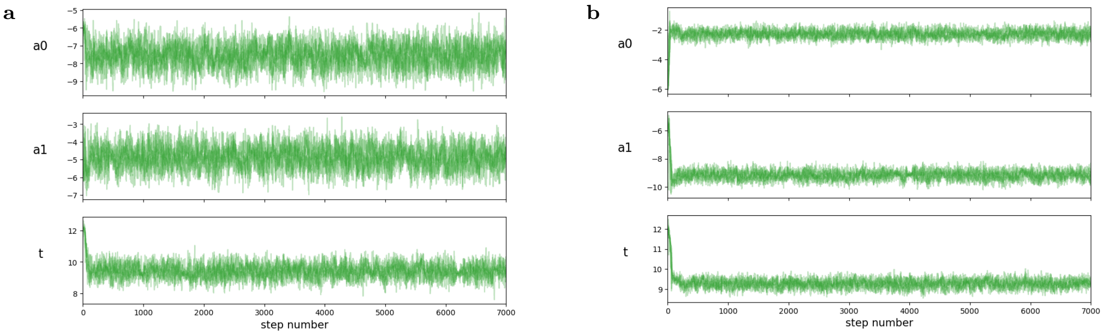

Figure 2: MCMC chains for the three parameters before trimming and thinning. Panel 2a shows monthly means for the years 1950-1954 while Panel 2b shows daily means from a 30 year period (1950-1980).

parameter	a_0	a_1	t
Monthly data	-7.54 ± 0.59	-4.88 ± 0.58	9.41 ± 0.41
Yearly data	-2.28 ± 0.25	-9.17 ± 0.29	9.27 ± 0.19

Table 2: Parameters from posterior distribution sampling, mean ± std.

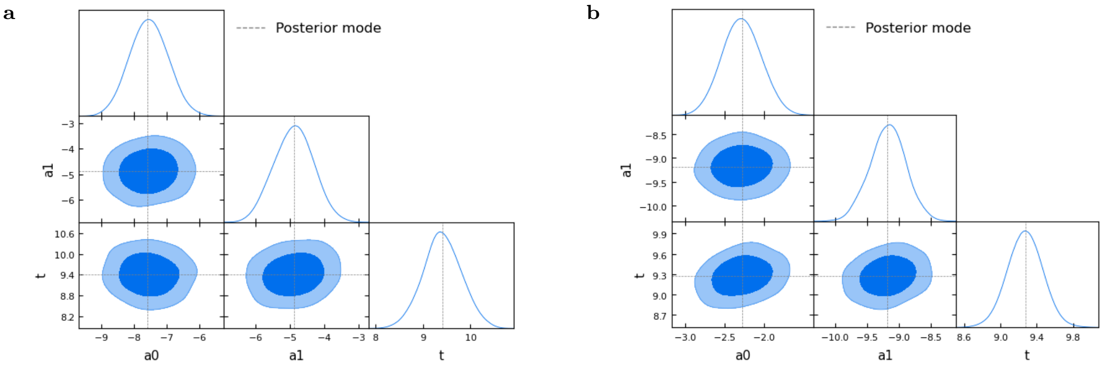

Figure 3: Corner plots for the three parameters. Panel 3a shows monthly means for the years 1950-1954 while Panel 3b shows daily means from a 30 year period (1950-1980)

As described in the methods the model was checked using the generated MCMC chains. Figure 4 shows the model predictive distribution i.e. the variation in the model when parameter values are drawn randomly from our MCMC chains rather than just taking the MAP values. For the case of monthly data it can be seen that, similar to the MAP model, the model predictive distribution remains slightly phase-shifted and that the quantile bands do not overlap with all data points given their corresponding uncertainties. This is in contrast to the results for the climatology case where the observations generally fall within the model predictive distribution quantile bands. Figure 5 displays the posterior predictive distributions for both approaches, where random noise, scaled according to the observed error, is added to the model. In this instance the quantile distributions show how the model is able to capture the uncertainty estimates of the original observations.

It has been noted that the MAP model for the monthly data appeared to have a small phase shift compared to the data. In an attempt to improve this result the parameter ϕ was no longer assumed fixed at zero but

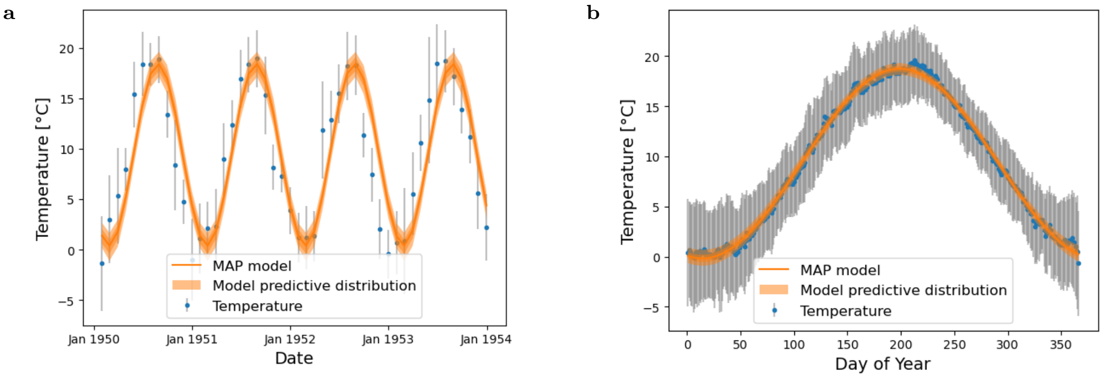

Figure 4: Predictive distributions from the MAP model (orange) with 2.5%, 16%, 84% and 97.5% quantiles are shown on top of the data (blue) and error bars (grey). 4a shows monthly means for the years 1950-1954 while 4b shows daily means from a 30 year period (1950-1980)

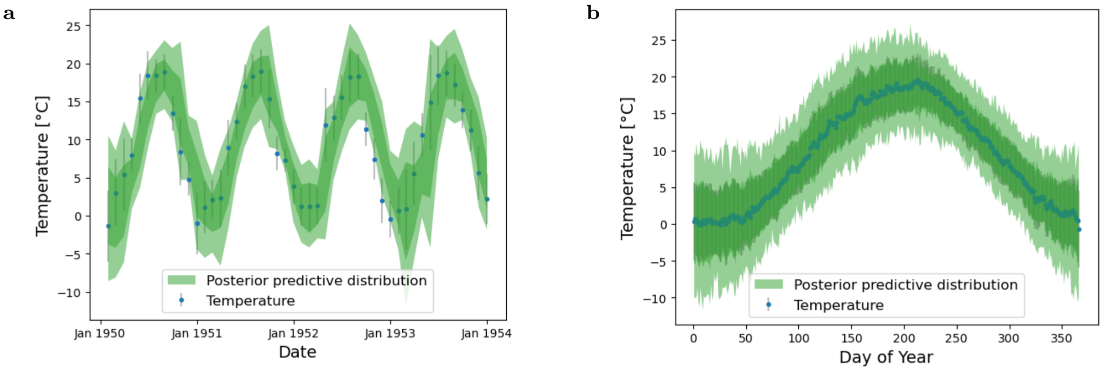

Figure 5: Posterior predictive distributions from sampling the MCMC chains (green) with 2.5%, 16%, 84% and 97.5% quantiles are shown on top of the data (blue) and error bars (grey). 5a shows monthly means for the years 1950-1954 while 5b shows daily means from a 30 year period (1950-1980)

rather modelled as a free parameter. Following the same analysis steps as above, the MAP value for ϕ was estimated at $\phi = 0.214$ and $\phi = 0.20 \pm 0.63$ from sampling the posterior using the same MCMC procedure as described before. Interestingly, though the standard deviation estimate for ϕ is rather small, the estimated standard deviation for other parameters a_0 and a_1 was drastically increased; from 0.59 to 4.63 for a_0 and 0.58 to 2.23 for a_1 respectively.

Including ϕ as a free parameter affected the shape of the corner plot distributions, again in particular the shape of parameters a_0 and a_1 changed (see Figure 6). Additionally, the model predictive distribution quantile band that spans from quantile 0.025 to quantile 0.975 (light orange shading) does not display any seasonality features. Only for a smaller quantile range such as 0.16 to 0.84, is there a strong seasonal pattern. This is also the case for the posterior predictive distribution. Hence, although the MAP model phase-shift disappeared, the overall model performance decreased with regards to sampling the posterior.

From this observation it was of interest to explore the effect of changing another fixed parameter to a free one. Therefore, for the climatology data, it was also tested how modelling ω as a free parameter affects the results (data not shown). Interestingly, when ω is estimated as a free parameter the performance of the model also decreases. The MAP model and parameters remain essentially the same, however, as with changing the ϕ parameter, the standard deviation of estimated parameters increases and MCMC performance decreases. Corner plots show that we end up with a multi-modal posterior distribution with "pockets" in which the MCMC walkers can get trapped and as such auto-correlation times are so large that they throw an error. The effect of this is an increase in the uncertainty around the model means.

These effects are likely due to the fact that ϕ and ω are non-linear parameters of our model. This can give rise to a multi-modal posterior distribution which makes sampling the parameter space using MCMC method less efficient and could lead to Markov chains that are stuck in local minima.

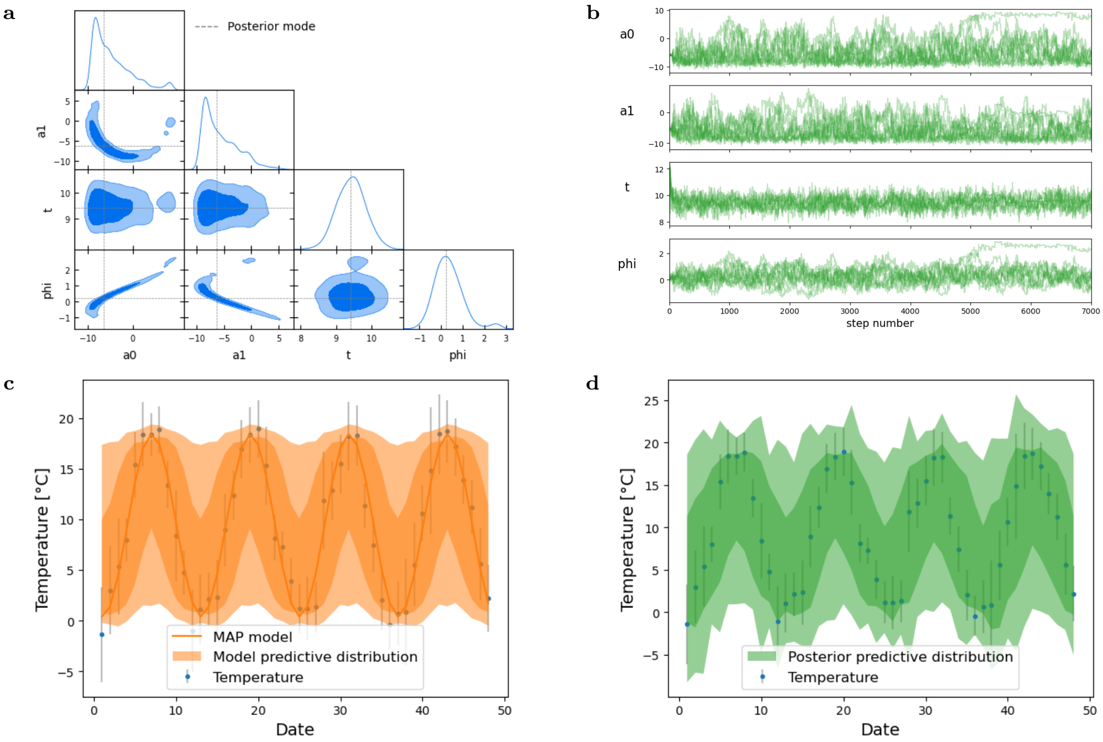

Figure 6: Results from the monthly data analysis where ϕ is an additional free parameter.

4 Conclusion

Our results show that the simplistic probabilistic model of a first-order approximation of the Fourier Series is suitable to capture the mode of seasonal temperature data given uncertainty estimates of one standard deviation obtained from averaging either on a monthly or a climatological scale. This simplistic model works best when only linear parameters are allowed to be free. With regards to the uncertainty structure of the model, the MCMC chains of both approaches displayed the expected pattern of fluctuations around an optimal value. The posterior predictive distribution for both approaches also shows that simulated data should agree with the already observed data. For the climatology data this agreement appears qualitatively better than for the monthly data. One possible explanation for this is that by averaging over 30-years more variations in the data are averaged out making the data more smooth.

In contrast when the non-linear parameters, ϕ and ω, are modeled as free, the model performance decreases, especially when making predictions or trying to estimate uncertainty around parameter values. For example, the quantile band 0.16 to 0.84 shows a strong seasonality but extending it to the quantiles 0.025 to 0.975 removes the seasonal pattern entirely. We conclude that if non-linear parameters are included in the model, the MCMC method may not be suitable to sample from the more complex resulting distributions and alternative sampling methods should be considered. Further research could be done investigating the impact of priors, including more Fourier Series terms in the model and applying the model to other, more recent temperature data from the same region.

References

[1] Deutscher Wetterdienst (DWD). Historische Klimadaten Berlin-Dahlem. https://www.dwd.de/DE/leistungen/klimadatendeutschland/klarchivtagmonat.html. Accessed: 2024-06-20.

[2] WMO. Calculation of monthly and annual 30-year standard normals,. *World Meteorological Organization Tech. Doc. 341, WCDP*, pages 10–11, 1989.

[3] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. emcee: The MCMC Hammer. , 125(925):306, March 2013. doi:10.1086/670067.