# Sakura Blossom Prediction Model for Japan*

**Forecasting Sakura Blossom Using Bayesian Spline Regression**

Shanjie Jiao

November 25, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

## Table of contents



*Code and data are available at: https://github.com/Jie-jiao05/Sakura-Blossom-Prediction-Model.

# 1 Introduction

Sakura not merely a ornamental plants but also hold profound cultural significance. In Japanese literature, poetry, and art, sakura blossoms carry deep emotional and symbolic meaning, with the aesthetic concept of "mono no aware" being particularly notable. Due to their short blooming period, sakura blossoms are often seen as a metaphor for the impermanence and fleeting beauty of life, evoking deep reflection and appreciation for the essence of existence.

Beyond their cultural significance, sakura blossoms also have a significant positive impact on Japan's economy. "Ohanami" (sakura blossom viewing) is a traditional celebration of spring that attracts a large number of domestic and international visitors every year during the blooming season from April to May. According to research by Katsuhiro Miyamoto, a professor at Kansai University, the 2024 cherry blossom season is projected to contribute up to ¥1.14 trillion (approximately $7.7 billion) to Japan's economy (Kaneko 2024). This event not only supports the post-pandemic recovery of the tourism sector but also positively impacts related industries such as catering and retail.

Given the importance of sakura blooming times for tourism planning and economic activities, accurately forecasting these dates is essential. This study aims to utilize linear regression and Bayesian spline methods to systematically analyze the effects of temperature and geographical location on sakura blooming times. By developing a predictive model, the study seeks to provide scientific insights for sakura enthusiasts worldwide, as well as for tourism and related industries, facilitating more precise planning of viewing activities and resource allocation. Furthermore, analyzing sakura blossom data can also expore on the impact of global warming on blooming periods."

result part

The structure of this paper is as follows: Section Section 2 details the data sources and the methodologies employed, including data scraping and manipulation techniques. Section **?@sec-model** outlines the development of prediction models, specifically Linear Regression

and Bayesian Spline Models, which are further analyzed in Section **?@sec-result**. In Section **?@sec-dis**, the impact of global warming on the sakura blossom period, along with real-life implementation and limitations of the study, will be discussed, providing insights for further improvement.

# 2 Data

## 2.1 Overview

We used the statistical programming language R (R Core Team 2023) to perform all analyses of the modern and historical sakura blossom data. The data were extracted from Alex Cookson's (Cookson 2020) and combined with temperature data scraped from the Japan Meteorological Agency (Agency 2024).

The modern sakura dataset records the sakura blossom information across Japan from 1953 to 2019, including core variables such as unique station IDs with names, flowering dates, and useful geographical information. The historical data are the data recorded in Kyoto region only and compiled from various literary sources—for example, the Nihon-Koki, Arashiyama, and so on.

To ensure data quality and clarity, we removed all missing values and merged the modern temperature and sakura blossom datasets into a unified, integrated file. Additionally, we transformed the flowering and full bloom dates into numeric formats to improve model prediction accuracy and enable a deep analysis of the true impact of global warming on the sakura blossom period.

For performing the analysis, we utilized several R packages. Tidyverse(Wickham et al. 2019), Dplyr(Hadley Wickham and Romain François and Lionel Henry and Kirill Müller and Davis Vaughan 2023), Here(Müller 2020), Readr(Wickham, Hester, and Bryan 2024), Lubridate(Grolemund and Wickham 2011), Vest(Wickham 2024)) for data cleaning and scraping.

This research is constructed under the guidance of Dr.Rohan Alexander. (Alexander 2023)

## 2.2 Measurement

Our dataset, sourced from Alex Cookson's work (Cookson 2020), integrates temperature data scraped from the Japan Meteorological Agency (Agency 2024). The merged dataset contains 5,387 observations, aggregating average temperatures for the corresponding regions and flowering months. Figure 1 It includes detailed records on flowering dates, full bloom dates, and geographic locations. By compiling flowering times and geographic information for sakura blossoms across Japan since 1953 to 2019, this dataset provides comprehensive foundational

data for studying the timing patterns and potential influencing factors of sakura blossom flowering.

| ID | Location | Latitude | Longitude | Year | Month | Flower Day | Full Bloom Day | Mean Temp |
|---|---|---|---|---|---|---|---|---|
| 47401 | Wakkanai | 45.41500 | 141.6789 | 1953 | May | 141 | 150 | 6.9 |
| 47406 | Rumoi | 43.94611 | 141.6319 | 1953 | May | 128 | 133 | 9.8 |
| 47407 | Asahikawa | 43.75694 | 142.3722 | 1953 | May | 131 | 136 | 10.5 |
| 47409 | Abashiri | 44.01778 | 144.2797 | 1953 | May | 144 | 146 | 7.2 |
| 47412 | Sapporo | 43.06000 | 141.3286 | 1953 | May | 127 | 134 | 11.3 |
| 47413 | Iwamizawa | 43.21167 | 141.7858 | 1953 | May | 129 | 131 | 10.6 |

Figure 1: Sample of Modern Sakura Data

For historical sakura data, since the earliest data in this dataset can be traced back to 812, however the accuracy of temperature measurements in the early years is questionable, and data being recorded only in the Kyoto area, there may be some unavoidable bias. Therefore, when building the prediction model, we will only use modern sakura data for fitting, and historical data will only serve as a comparison to help us understand the historical situation.

The outcome variable in this study represents the flowering time of sakura blossoms. As part of the data refinement process, to enhance the reliability of the predictions, we converted the flowering and full bloom dates from the standard "yyyy-mm-dd" format into numerical values, enabling a more precise model fit. Additionally, since the dataset only includes sakura blossom data in Japan, the conclusions drawn from this study are limited to providing a reference for the flowering times of sakura blossoms within Japan and do not consider the influence of different varieties of sakura.

Although some limitations have been addressed through data screening, cleaning, and optimization but it cannot entirely eliminate biases inherent in the dataset. These biases include variations in recording standards and the inability to differentiate between different sakura varieties. Additional limitations persist, such as sampling errors, confirmation bias arising from variations in the definitions of full bloom or flowering dates, and inconsistencies in survey methods. Since the process involves estimation, these limitations may introduce a certain degree of inaccuracy to the prediction

## 2.3 Outcome variables

The main outcome variables in this study are the "flowering day" and "full bloom day," which represent the specific dates (converted into numeric form) when sakura enter the flowering and full bloom stages, respectively. A statistical summary of the "flowering day" and "full bloom

day" is presented in Table 2, while Figure 2 illustrates the general distribution of these two variables. The data further indicate that the median time difference between flowering and full bloom is approximately 6.22 days. Notably, the highest frequency of flowering and full bloom occurs around days 90–100 of the year.

Table 2: Statistic Summary of Flowering and Full Blossom Day

| Statistic | Flowering.Day | Full.Bloom.Day |
|---|---|---|
| 1st Qu. | 87.0000 | 95.0000 |
| 3rd Qu. | 107.0000 | 112.0000 |
| Max. | 151.0000 | 160.0000 |
| Mean | 98.9625 | 105.1819 |
| Median | 94.0000 | 100.0000 |
| Min. | 20.0000 | 60.0000 |

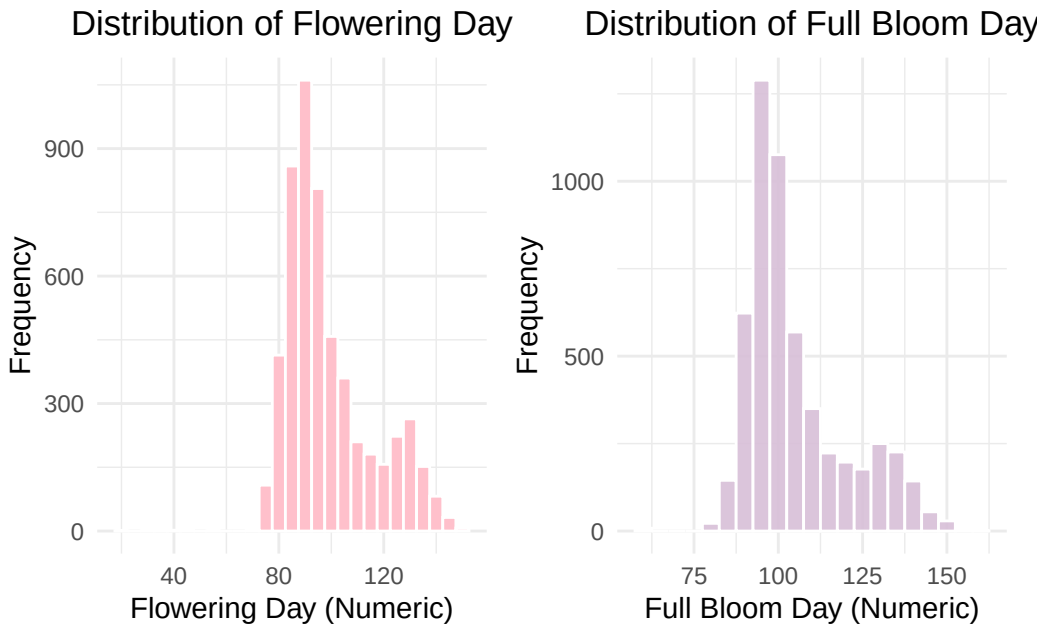



Figure 2: Distribution of Flowering and Full Blossom Day

## 2.4 Predictor variables

In this study, sakura blooming dates are influenced by multiple environmental and geographical factors, leading to the selection of several key predictor variables for analysis. ###

Average Temperature of the Flowering Month The first variable is the average temperature of the flowering month (month_mean_temp). As Dr. Casper J. van der Kooi, Peter G. Kevan, and Matthew H. Koski emphasize in their article "The thermal ecology of flowers" published in PubMed Central, "temperature mediates flower growth and development, pollen and ovule viability, and influences pollinator visitation" (Kooi, Kevan, and Koski 2019). Since temperature directly affects plant physiological processes and ecological interactions, it is considered one of the most critical predictors in this study.

### 2.4.1 Geographical Information (Latitude and Longitude)

The second variable is geographical information, including latitude and longitude, which provides precise spatial details about the recording locations in different regions. In this dataset, a total of 96 unique locations were recorded, covering regions across Japan Figure 3. Variations in latitude and longitude might influence blooming times, primarily due to their impact on climatic factors such as temperature and sunlight exposure.

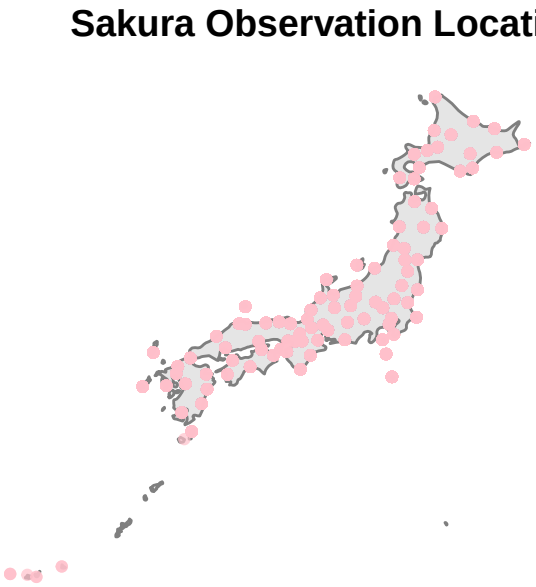



Figure 3

### 2.4.2 Years under Global Warming

Lastly, considering the trend of global climate warming in recent decades, the variable "year" is also included. By retrieving temperature data from 1953 to 2023 from the Japan Meteorological

Agency (Agency 2024), we generated Figure 4, revealing that the temperature in Japan has risen by approximately 2.73 degrees Celsius compared to 1953. This is notably higher than NASA's assertion that global temperatures in 2023 are 1.36 degrees Celsius warmer than the late 19th century (1850–1900) (NASA 2023). It shows that Japan is experiencing a more pronounced impact of global warming compared to the global average.

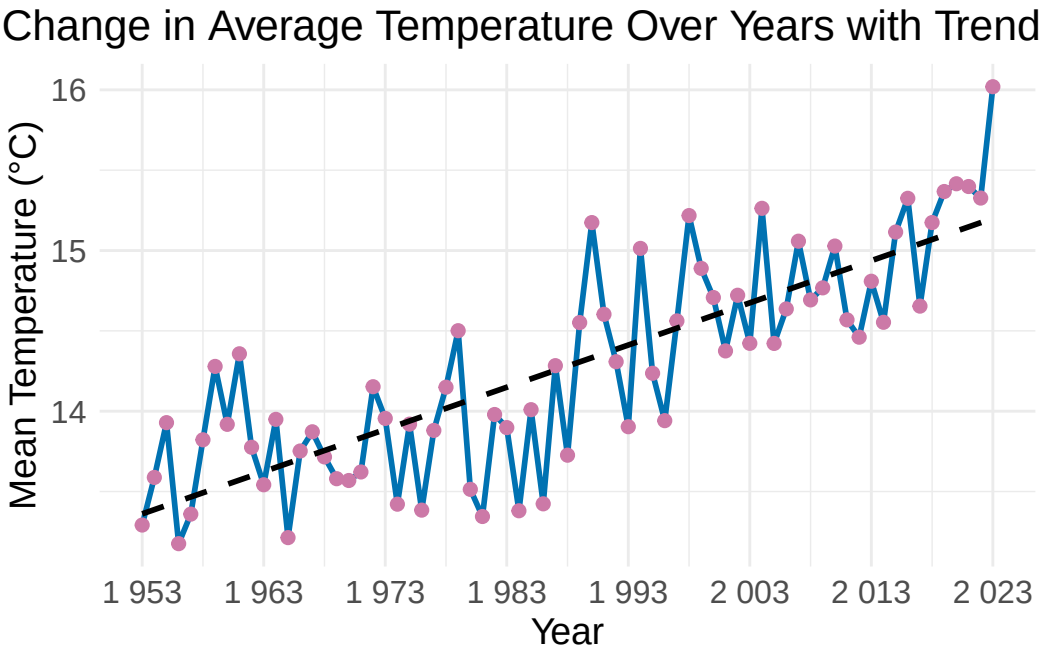



Figure 4

### 2.5 Correlation between Predictor Variables

#### 2.5.1 Latitude and Longitude with Temperature

Figure 5 demonstrates how temperatures vary across different locations based on their geographical coordinates. A positive relation could be observed, with lower temperatures observed at higher latitudes, such as in northern Japan, and gradually increasing temperatures as the coordinates approach regions closer to the equator.

## 3 Model

The goal of our modeling is to predict the precise timing of sakura blooming and full blooming across different regions of Japan each year. To achieve this, the model incorporates geographi-

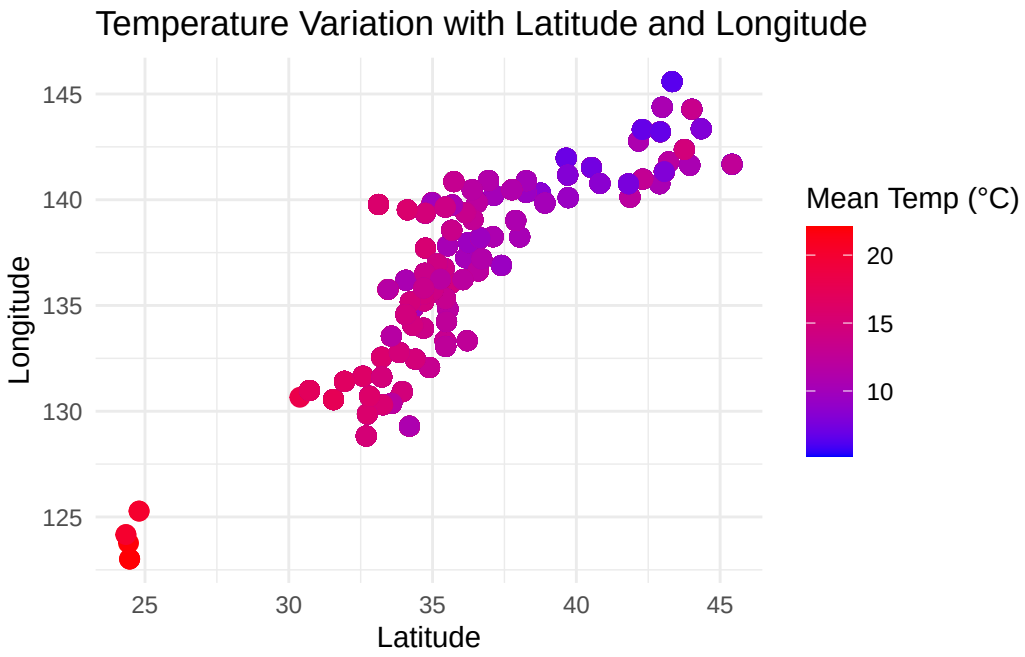



Figure 5

cal factors, accounting for variations in sakura blooming timing due to temperature differences arising from diverse geographical locations.

Given the inherent instability, uncertainty, and unpredictability of the temperature variable, we adopt a two-step modeling approach. First, we employ a linear regression model to establish a baseline prediction, providing initial insights into the relationship between the predictors and blooming dates. Next, we leverage a Bayesian model to address the complexities and uncertainties that the linear model cannot capture. The Bayesian approach allows us to incorporate the uncertainty into the model, enhancing its ability to predict the timing of sakura blooming with greater accuracy and reliability.

Background details and diagnostics are included in Appendix B.

## 3.1 Model set-up

Define $y_i$ as the specific day of sakura flowering date in numeric form. And we begin with a linear regression model than fit it with a more in depth Bayesian model. ### Linear Regression Model

Where: - $y_i$ is the flowering day for the $i$-th observation (measured in numeric form). - $\beta_0$ is the intercept of the model. - $\beta_1$ is the coefficient for the month_mean_temperature (measured

in Celsius). - $\beta_2$ is the coefficient for latitude. - $\beta_3$ is the coefficient for longitude. - $\epsilon_i$ is the error term for the $i$-th observation

#### 3.1.1 Bayesian Model

Where: - $y_i$ is the observed flowering day for the $i$-th observation. - $\mu_i$ is the expected flowering day for the $i$-th observation. - $\beta_0, \beta_1, \beta_2, \beta_3$ are the regression coefficients, assumed to follow a normal prior with mean 0 and variance $10^2$. - $\sigma^2$ is the variance of the residuals, assumed to follow an inverse-gamma prior with hyperparameters $\alpha$ and $\beta$.

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

#### 3.1.2 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance $\theta$.

## 4 Results

Our results are summarized in Table 3.

## 5 Discussion

### 5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

### 5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

Table 3: Explanatory models of flight time based on wing width and wing length

| | First model |
|---|---|
| (Intercept) | 1.12 |
| | (1.70) |
| length | 0.01 |
| | (0.01) |
| width | −0.01 |
| | (0.02) |
| Num.Obs. | 19 |
| R2 | 0.320 |
| R2 Adj. | 0.019 |
| Log.Lik. | −18.128 |
| ELPD | −21.6 |
| ELPD s.e. | 2.1 |
| LOOIC | 43.2 |
| LOOIC s.e. | 4.3 |
| WAIC | 42.7 |
| RMSE | 0.60 |

### 5.3 Third discussion point

### 5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

# Appendix

# A Additional data details

# B Model details

## B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

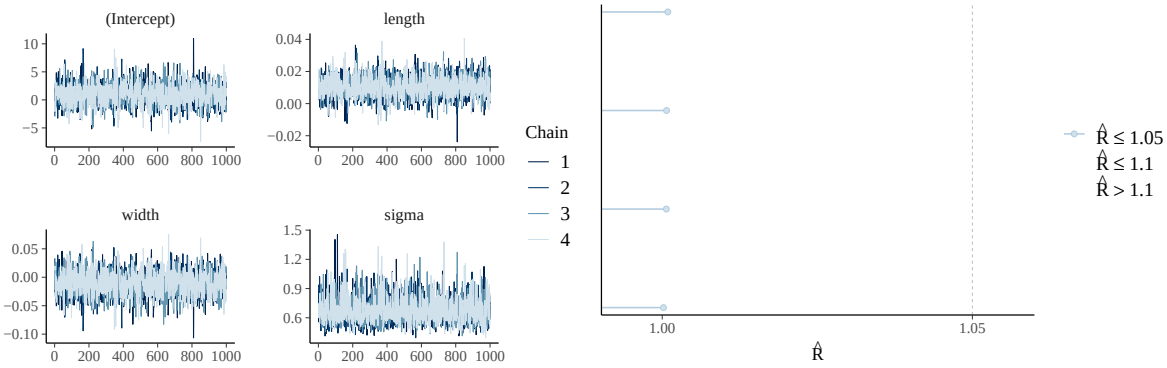
In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

## B.2 Diagnostics

Figure 6a is a trace plot. It shows... This suggests...

Figure 6b is a Rhat plot. It shows... This suggests...




(a) Trace plot

(b) Rhat plot

Figure 6: Checking the convergence of the MCMC algorithm

## References


Agency, Japan Meteorological. 2024. *Japan Meteorological Agency | Tables of Monthly Climate Statistics*. *Jma.go.jp*. https://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3_en.php?block_no=47401.

Alexander, Rohan. 2023. "Telling Stories with Data." Telling Stories with Data. https://tellingstorieswithdata.com/.

Cookson, Alex. 2020. *data/sakura-flowering at master · tacookson/data*. *GitHub*. https://github.com/tacookson/data/tree/master/sakura-flowering.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hadley Wickham and Romain François and Lionel Henry and Kirill Müller and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Kaneko, Karin. 2024. *Economic impact of hanami expected to double this year*. *The Japan Times*. https://www.japantimes.co.jp/news/2024/03/15/japan/society/hanami-economic-impact/.

Kooi, Casper J. van der, Peter G. Kevan, and Matthew H. Koski. 2019. "The thermal ecology of flowers." *Annals of Botany* 124 (3): 343–53. https://doi.org/10.1093/aob/mcz073.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

NASA. 2023. *Global Surface Temperature | NASA Global Climate Change*. *Climate Change: Vital Signs of the Planet*. NASA. https://climate.nasa.gov/vital-signs/global-temperature/?intent=121.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2024. *rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.