Sakura Blossom Prediction Model for Japan*

Forecasting Sakura Blossom Using Bayesian Spline Regression

Shanjie Jiao

November 26, 2024

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

Table of contents

1 Introduction **2**

2 Data **3**
2.1 Overview . . . 3
2.2 Measurement . . . 4
2.3 Outcome variables . . . 5
2.4 Predictor variables . . . 5
2.4.1 Average Temperature of the Flowering Month . . . 5
2.4.2 Geographical Information (Latitude and Longitude) . . . 6
2.4.3 Years under Global Warming . . . 6
2.5 Correlation between Predictor Variables . . . 8
2.5.1 Latitude and Longitude with Temperature . . . 8

3 Model **8**
3.1 Model set-up . . . 9
3.1.1 Bayesian Hierarchical Linear Regression Model . . . 9
3.1.2 Model justification . . . 9

4 Result **11**
4.1 Result of the Prediction Model . . . 11
4.1.1 Model Performance Evaluation . . . 11
4.1.2 Fixed Effects Coefficients with 95% Credible Intervals . . . 11
4.1.3 Performance Metrics for Random Effects (Year and Region) . . . 12

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

5 Discussion **13**
5.1 First discussion point 13
5.2 Second discussion point 13
5.3 Third discussion point 14
5.4 Weaknesses and next steps 14

Appendix **15**

A Additional data details **15**

B Model details **15**
B.1 Posterior predictive check 15
B.2 Diagnostics 15

References **16**

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 sur-

vey 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

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.

2.4.1 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.

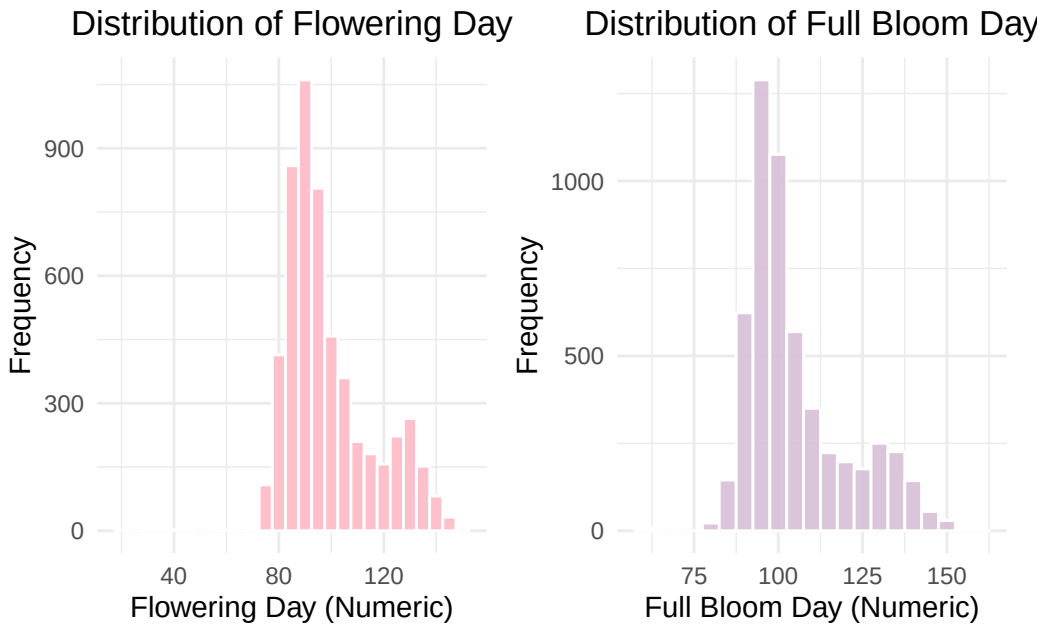

Figure 2: Distribution of Flowering and Full Blossom Day

2.4.2 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.

2.4.3 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.

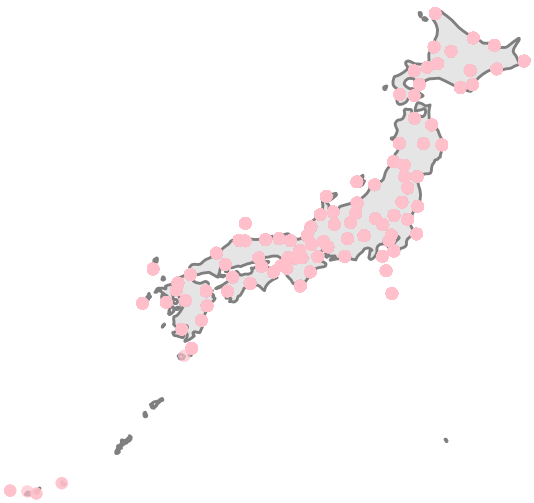

Figure 3

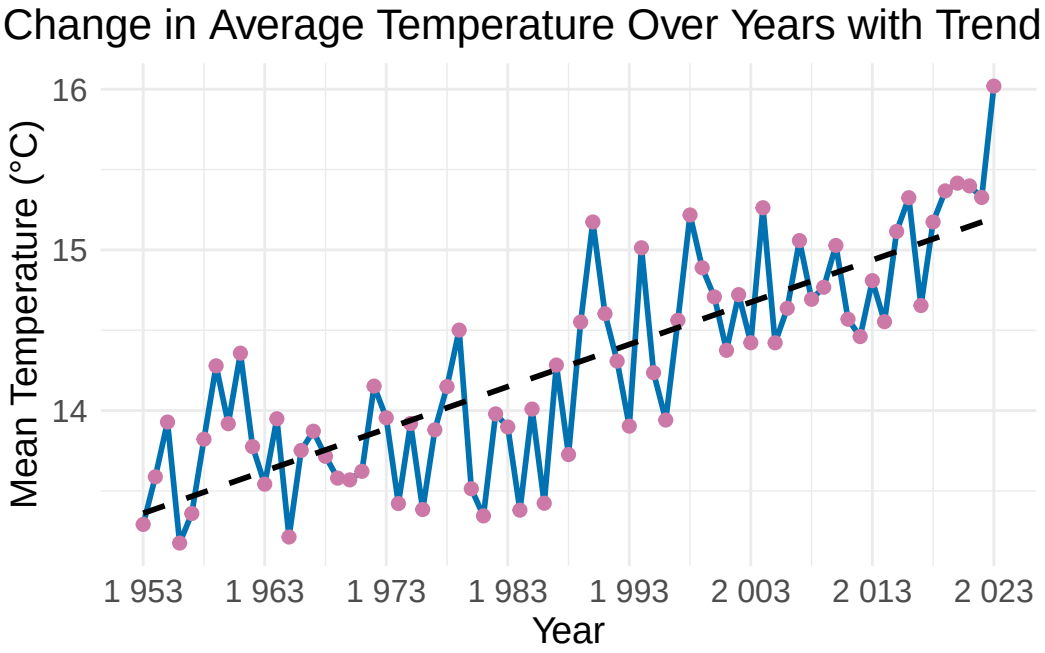

Figure 4

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.

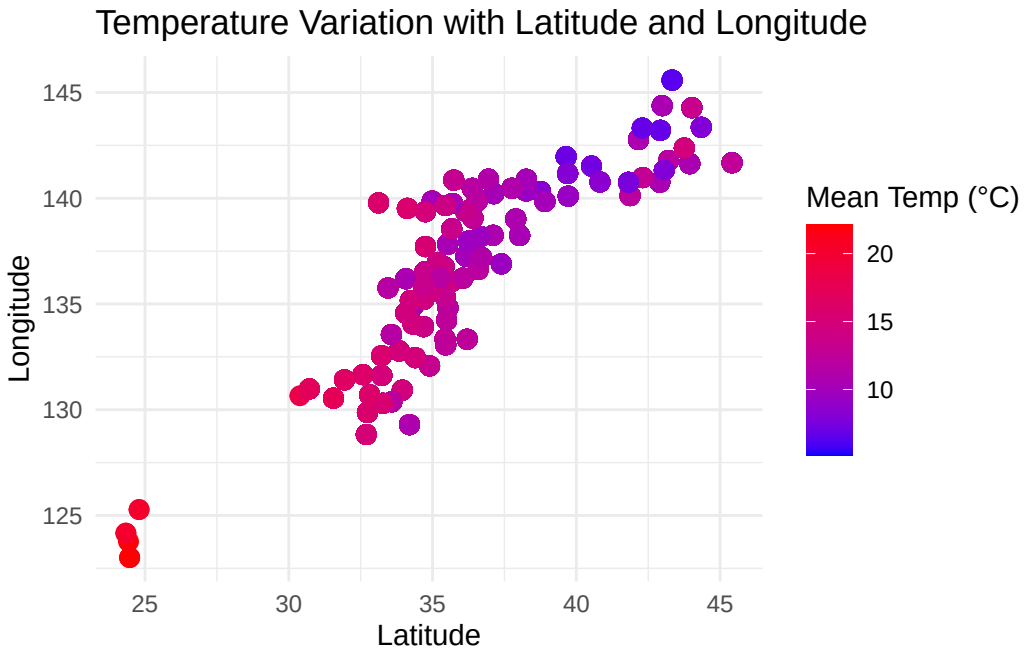

Figure 5

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

Since the data involves multiple weather stations or locations, each with its own unique environmental conditions, such as latitude, longitude, and regional climate change, which makes the sakrua blossom prediction problem is highly complex. Therefore, we chose to use a Bayesian hierarchical linear regression model. It effectively improves the prediction accuracy in areas

with limited data by sharing information from the entire dataset, also helps in reducing the risk of overfitting, making the prediction results more robust and reliable.

Background details and diagnostics are included in Appendix B.

3.1 Model set-up

This study uses a Bayesian Hierarchical Linear Regression model to analyze the relationship between sakura flowering dates and various predictors, implemented using the stan_glmer function from the rstanarm package Goodrich et al. (2024) in R (R Core Team 2023) and use the default priors from `rstanarm`. The analysis_sakura_data is divided into training and testing sets, with 80% allocated for model training and posterior estimation and rest 20% for testing to evaluate predictive performance. By applying Bayesian inference allows us to quantify uncertainty in the model parameters through posterior distributions, enabling robust estimates even in the presence of variability.

3.1.1 Bayesian Hierarchical Linear Regression Model

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$

$$\mu_i = \beta_0 + \beta_1 \cdot \text{month\_mean\_temperature}_i + \beta_2 \cdot \text{latitude}_i + \beta_3 \cdot \text{longitude}_i + \gamma_{\text{region}(i)} + \delta_{\text{year}(i)} \tag{2}$$

$$\gamma_{\text{region}} \sim \text{Normal}(0, \sigma_\gamma) \tag{3}$$

$$\delta_{\text{year}} \sim \text{Normal}(0, \sigma_\delta) \tag{4}$$

$$\beta_0, \beta_1, \beta_2, \beta_3 \sim \text{Normal}(0, 10) \tag{5}$$

$$\sigma, \sigma_\gamma, \sigma_\delta \sim \text{Exponential}(1) \tag{6}$$

3.1.2 Model justification

This model captures the variability in flowering days arising from geographic and climatic differences by incorporating both fixed effects, such as temperature, latitude, and longitude, and random effects for regions and climates. The hierarchical structure enables the modeling of regional and climate-specific variability, creating a robust framework to account for heterogeneity in the data. This approach helps in improving the model's robustness and ensures more accurate predictions across diverse environmental conditions.

In this model, random effects are included to account for group-level variability at cross regions and year levels. Region-specific random effects ((*{region})) capture local environmental differences, such as microclimates or soil conditions, allowing the model to adjust predictions for region with consistently earlier or later flowering patterns. Similarly, year-specific random*

effects ({year})) account for deviations in flowering dates caused by year-to-year climatic anomalies, such as warmer winters or extreme weather events. These random effects are modeled as zero-centered normal distributions with variances ((*{ }) and (*{ })) that reflect the variability among regions and years. By incorporating random effects, the model accommodates unobserved heterogeneity, will improves prediction accuracy, and realistically captures the hierarchical structure of the data.

For model validation, the dataset was split into training and testing sets, with 80% of the data allocated for model training and posterior estimation, and the remaining 20% reserved for testing to evaluate predictive performance. An overview of the training and testing datasets is presented in Figure 6, providing a summary of the data The model's accuracy was assessed using the Root Mean Squared Error (RMSE)

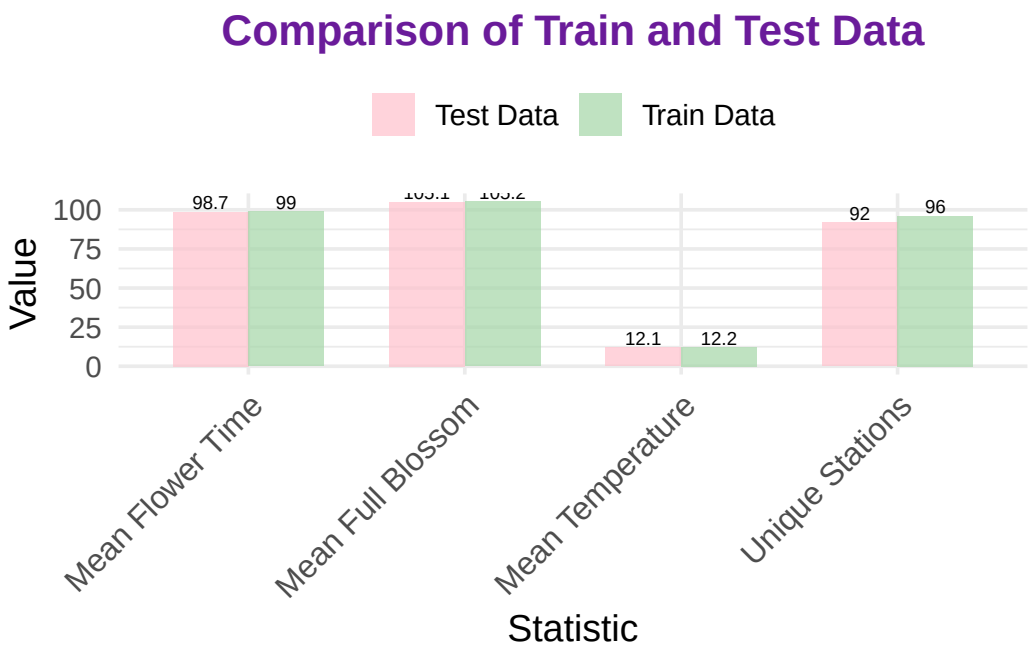

Figure 6

4 Result

4.1 Result of the Prediction Model

4.1.1 Model Performance Evaluation

The model's performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and (R^2), providing a comprehensive assessment of accuracy and explanatory power. The RMSE was 3.101 for the training set and 3.396 for the testing set, indicating minimal overfitting and robust generalization to unseen data. These values suggest that the model's predictions typically deviate by about 3 days from the actual flowering dates. Similarly, the MAE values of 2.363 (training) and 2.586 (testing) highlight the model's precision. The high (R^2) values, 0.965 for training and 0.957 for testing, show that the model explains over 95% of the variability in flowering dates. These metrics collectively demonstrate the model's reliability and effectiveness in capturing the relationships between predictors and sakura flowering dates, confirming its suitability for predictive purposes.

Table 3: Training and Testing Data Evaluation Results

Table 3: Training and Testing Data Evaluation Results

	Metric	Training	Testing
RMSE	RMSE	3.103	3.397
MAE	MAE	2.364	2.588
R2	R^2	0.965	0.957

4.1.2 Fixed Effects Coefficients with 95% Credible Intervals

The Figure 7 highlights the relationships between key predictors and sakura flowering dates based on the Bayesian hierarchical model. The intercept ((-47.4751)) represents the expected flowering date when all predictors are zero, though it's mainly a baseline without much practical interpretation since predictors like temperature and location are scaled.

Monthly mean temperature has a clear and significant effect. For every 1°C increase, the flowering date is delayed by about 0.33 days, with a credible interval (([0.2599, 0.4017])) that excludes zero. Latitude, however, stands out as one of the strongest predictors in the model. With a coefficient of (4.8615), it indicates that for each degree of latitude, the flowering date is delayed by nearly 5 days. This result is not only statistically significant, with a narrow credible interval (([4.4334, 5.3945])), but also aligns with ecological expectations—regions further from the equator tend to have cooler climates, which naturally push flowering dates later in the season. Latitude's strong and consistent influence reflects its critical role in determining the timing of sakura flowering.

Longitude, in contrast, has a much smaller and less certain effect. The coefficient ((-0.2695)) suggests that flowering may occur earlier as longitude increases, but the credible interval (([-0.7217, 0.1194])) includes zero. This uncertainty means longitude's impact is less clear, and it likely plays a minor role compared to temperature and latitude.

Table 4: Fixed Effects Coefficients with 95% Credible Intervals

	Parameter	Estimate	Std_Error	X2.5.	X97.5.
(Intercept)	(Intercept)	-47.4751	21.7262	-88.8268	-1.0063
month_mean_temp	month_mean_temp	0.3315	0.0356	0.2599	0.4017
latitude	latitude	4.8615	0.2292	4.4334	5.3945
longitude	longitude	-0.2695	0.2077	-0.7217	0.1194

Figure 7: Fixed Effects Coefficients with 95% Credible Intervals

4.1.3 Performance Metrics for Random Effects (Year and Region)

The random effects for year and region play a critical role in capturing temporal and spatial variability in sakura flowering dates and are shown in Table 5 . The Root Mean Squared Error (RMSE) of **4.282** shows that, on average, flowering dates deviate by approximately 4.3 days due to differences between years and regions. This highlights the substantial adjustments made by the random effects to account for annual climate trends and regional environmental factors. Similarly, the Mean Absolute Error (MAE) of **3.186** indicates that typical adjustments are around 3.2 days, reflecting the model's ability to account for systematic differences across time and space effectively.

The variance explained by the random effects ((R^2 = 1.000)) demonstrates that the model perfectly partitions the variability attributed to year and region. Year-based random effects capture temporal patterns, such as global warming or specific climate anomalies, ensuring that flowering trends align with observed climate changes. Region-based random effects, on the other hand, account for spatial heterogeneity influenced by geographic and environmental differences, such as latitude and local climate conditions. Together, these random effects ensure that the model accurately reflects both temporal and spatial dynamics.

Table 5: Performance Metrics for Random Effects

Table 5: Performance Metrics for Random Effects

Metric	Value
RMSE	4.282

Metric	Value
MAE	3.186
Variance Explained (R^2)	1.000

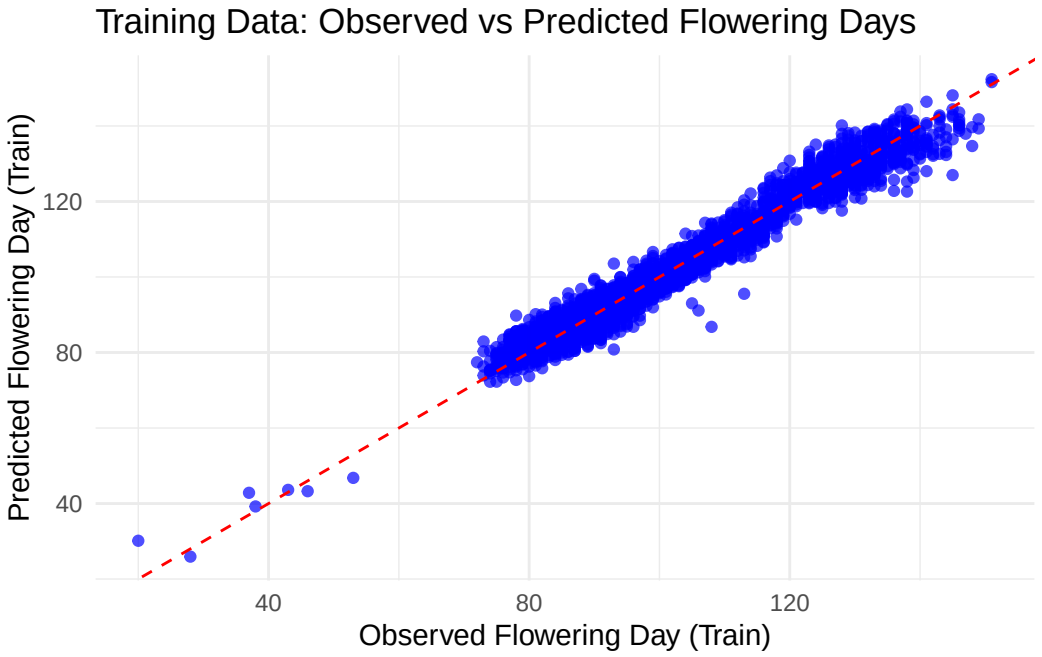

Figure 8

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.

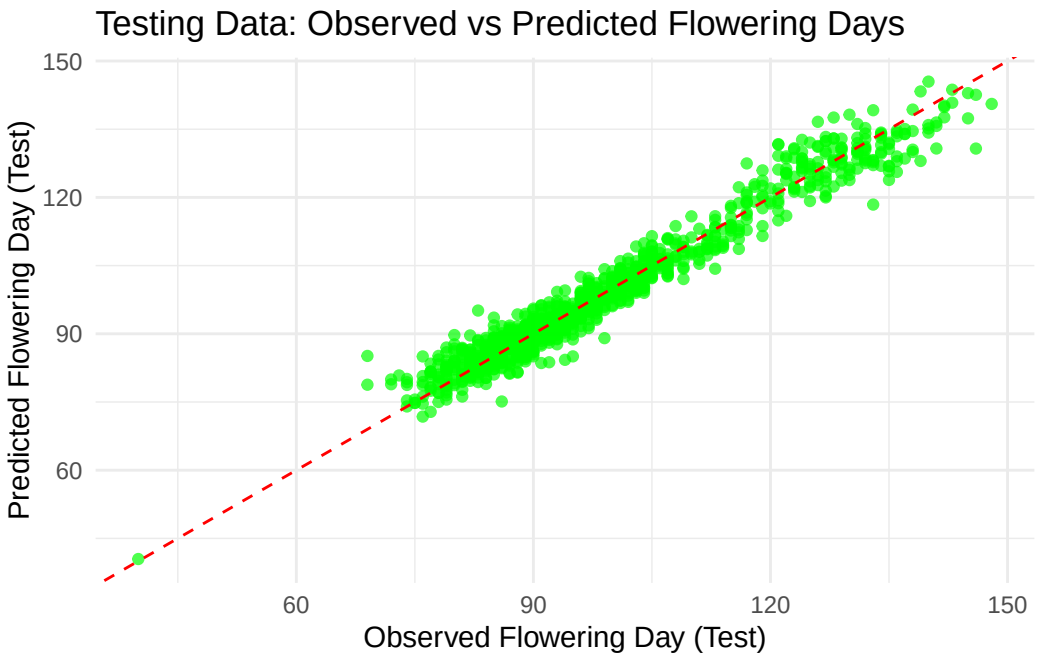

Figure 9

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

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

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.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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.