Sakura Blossom Prediction Model for Japan*

Forecasting Sakura Blossom Using Bayesian Hierarchical Regression

Shanjie Jiao

November 28, 2024

This study develops a Bayesian hierarchical linear regression model to predict sakura blooming dates across Japan by analyzing key predictors such as temperature, latitude, and longitude. The model demonstrates strong predictive accuracy, capturing over 95% of the variability in flowering dates and revealing clear regional and climatic patterns. The findings provide valuable insights into the effects of climate change on sakura phenology, aiding tourism planning and ecological conservation. Future improvements could incorporate dynamic modeling techniques, additional environmental variables, and broader datasets to enhance the model's precision and adaptability.

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."

The Bayesian hierarchical model build in this research predicts sakura flowering dates with high accuracy, achieving RMSE values of 3.1 (training) and 3.4 (testing) while explaining over 95% of variability. Notably, temperature and latitude emerge as significant predictors, revealing earlier blooming in southern regions and highlighting strong sensitivity to climatic and geographic factors. Also as emperature emerges as the most significant predictor, with an increase of 1°C in mean monthly temperature advancing flowering dates by approximately 0.33 days. Latitude shows a strong positive effect, with higher latitudes leading to later flowering dates. In contrast, longitude has a negligible impact.

The structure of this paper is as follows: Section 2 details the data sources and the methodologies employed, including data scraping and manipulation techniques. Section 3 outlines the development of prediction models, specifically Linear Regression and Bayesian Spline Models, which are further analyzed in Section 4. In Section 5, the impact of global warming on the sakura blossom period, along with real-life implementation and limitations of the study, will be discussed, also how sakura flowering condition from history to present, and will providing insights for further improvement. Further dataset and model detail will be presented in Section A.1 and Section A.2 separately.

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, maps(Richard A. Becker, Ray Brownrigg. Enhancements by Thomas P Minka, and Deckmyn. 2024), ggplot2(Wickham 2016),knitr(Xie 2023), arrow(Richardson et al. 2024), rstanarm(Cepeda et al. 2023), plotly(Sievert 2020), patchwork(Pedersen 2024), tidyr(Wickham, Vaughan, and Girlich 2024), bayesplot(Gabry et al. 2023), gridExtra(Auguie 2017) for data building and model preparation .

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. As shown in Table 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.

Table 1: Sample of Modern Sakura Data

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

Table 1: Sample of Modern Sakura Data

ID	Location	Latitude	Longitude	Year	Month	Flower Day	Full Bloom Day	Mean Temp
47412	Sapporo	43.06000	141.3286	1953	May	127	134	11.3
47413	Iwamizawa	43.21167	141.7858	1953	May	129	131	10.6

For historical sakura data is provided by Prof. Yasuyuki Aono (Yasuyuki 2015), 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 1 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

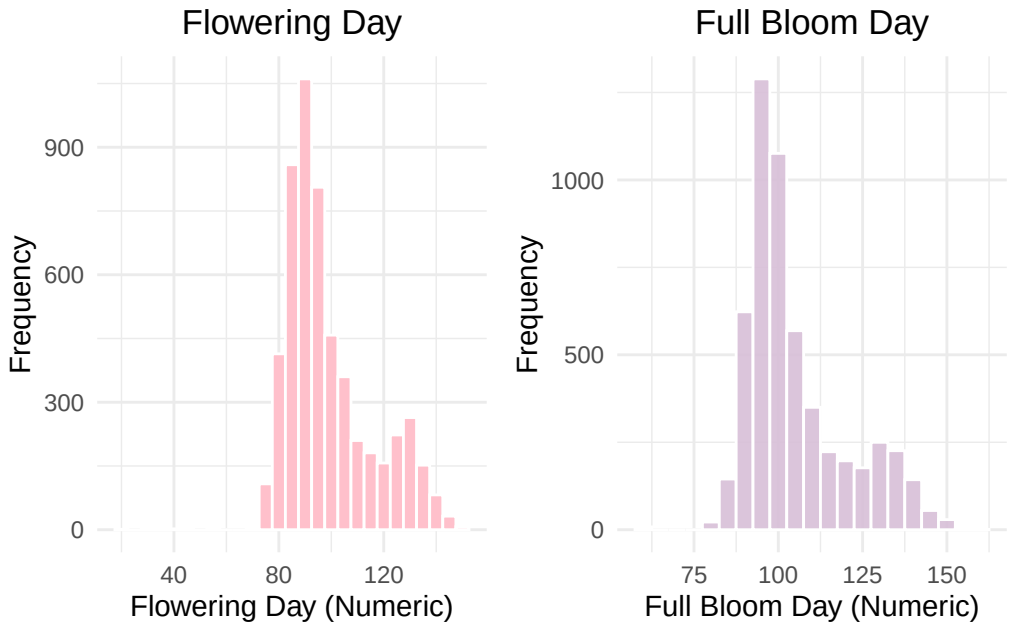


Figure 1: Distribution of Flowering and Full Blossom Day

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.

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.

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 2. Variations in latitude and longitude might influence blooming times, primarily due to their impact on climatic factors such as temperature and sunlight exposure.

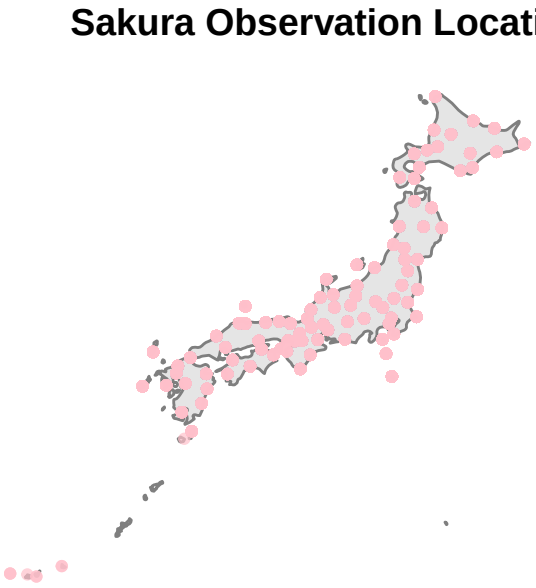


Figure 2: Recorded Geographical Information

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 3, 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.

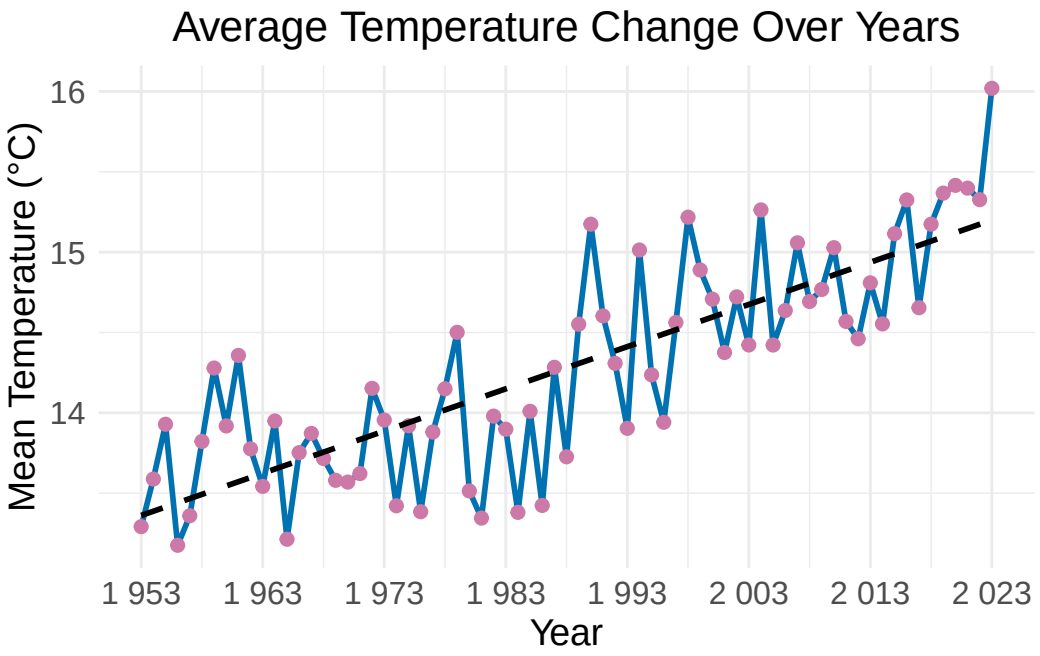


Figure 3: Change in Average Temperature Over Years with Trend Line

2.5 Correlation between Predictor Variables

2.5.1 Latitude and Longitude with Temperature

Figure 4 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 geographical factors, accounting for variations in sakura blooming timing due to temperature differences arising from diverse geographical locations.

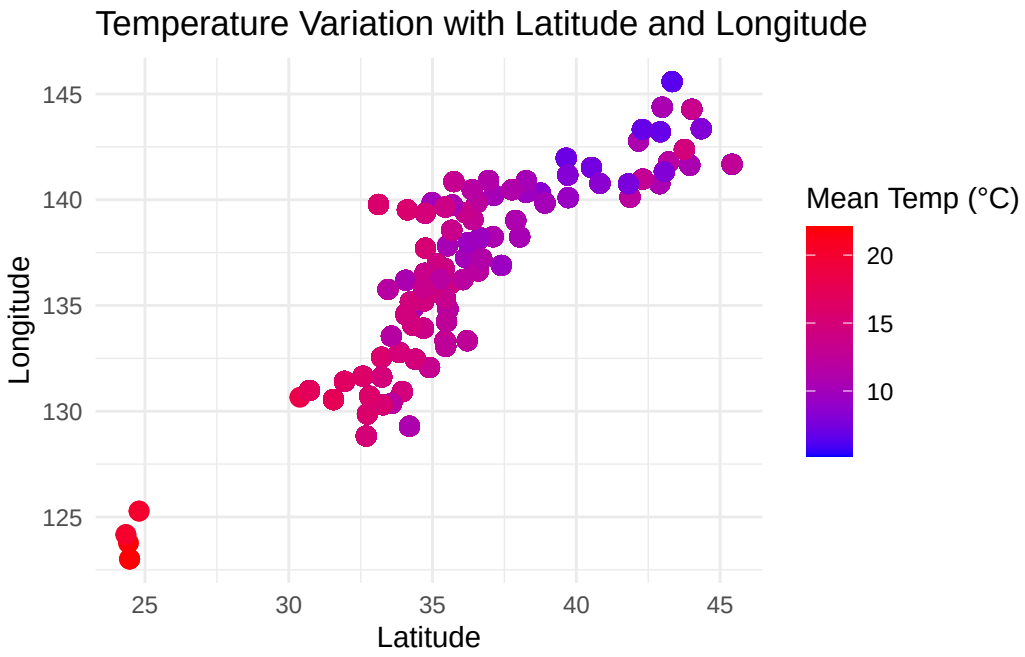


Figure 4: Temperature Variation

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 A.2.

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 Goodrich et al. (2024). 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 Estimand

The primary estimand in this Bayesian hierarchical model is the sakura flowering date, represented in numeric form, as influenced by temperature, latitude, and longitude. The model estimates both the fixed effects of these predictors and the random effects associated with yearly and regional variations, capturing localized deviations and inter-annual trends. Additionally, the model estimates residual variance to account for unexplained variability, ensuring a comprehensive understanding of the factors driving sakura phenology. These estimands enable the quantification of the relationship between climatic and geographic predictors and the timing of sakura blooming, while accounting for spatial and temporal heterogeneity.

3.1.2 Bayesian Hierarchical Linear Regression Model

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \beta_0 + \beta_1 \cdot \text{month\_mean\_temperature}_i + \beta_2 \cdot \text{latitude}_i \\
&\quad + \beta_3 \cdot \text{longitude}_i + \gamma_{\text{region}(i)} + \delta_{\text{year}(i)} && (2)\\
\gamma_{\text{region}} &\sim \text{Normal}(0, \sigma_\gamma) && (3)\\
\delta_{\text{year}} &\sim \text{Normal}(0, \sigma_\delta) && (4)\\
\beta_0, \beta_1, \beta_2, \beta_3 &\sim \text{Normal}(0, 10) && (5)\\
\sigma, \sigma_\gamma, \sigma_\delta &\sim \text{Exponential}(1) && (6)
\end{aligned}
$$

3.1.3 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 γ_{region} capture local environmental differences, such as microclimates or soil conditions, allowing the model to adjust predictions for regions with consistently earlier or later flowering patterns. Similarly, year-specific random effects δ_{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 σ_γ and σ_δ that reflect the variability among regions and years. By incorporating random effects, the model accommodates unobserved heterogeneity, 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 5, providing a summary of the data The model's accuracy was assessed using the Root Mean Squared Error (RMSE)

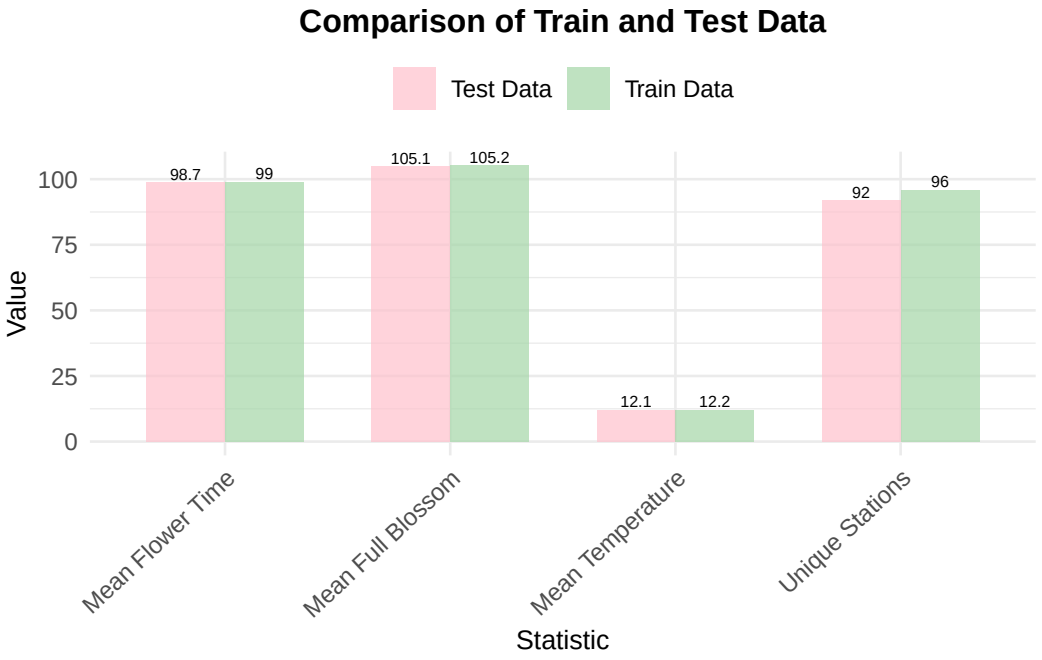


Figure 5: Training and Testing Data Evaluation Results

4 Result

4.1 Result of the Analysis Data

In Figure 6 is the result of correlation between each predictor variable, including latitude, longitude, and monthly mean temperature, and flowering day. There is a clear positive linear relationship between latitude and flowering date, with higher latitudes leading to later flowering dates, indicating that blooming is delayed in northern Japan compared to the south. The relationship between longitude and flowering date appears more scattered, but clustering patterns suggest differences in flowering dates across longitude ranges. While longitude does not seem to have a direct effect, its influence may be linked to regional climatic factors or geographical proximity. Lastly, there is a strong negative nonlinear relationship between monthly mean temperature and flowering date, where higher temperatures result in earlier

flowering, reflecting the biological response of sakura blossoms to warmer spring temperatures that accelerate blooming.

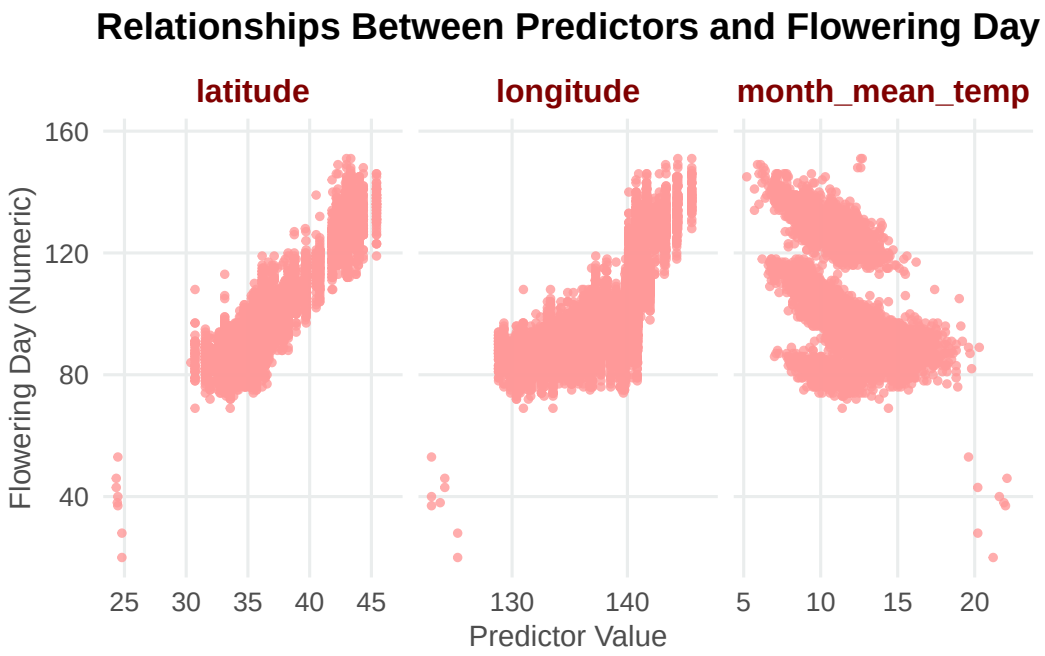


Figure 6: Predictors and Outcome Variable Relations

4.2 Result of the Prediction Model

4.2.1 Performance Overview Analysis

The scatter plot Figure 7 illustrates a strong alignment between observed and predicted sakura flowering dates for both training and testing datasets. Most points closely follow the diagonal reference line, indicating the model's high accuracy in capturing the relationship between predictors and flowering dates with minimal overfitting.

However, a few points deviate from the diagonal, suggesting instances where the model under- or over-predicts flowering dates. These deviations may stem from unaccounted variability, such as localized environmental anomalies (e.g., extreme winter conditions) or differences in sakura species not captured by the predictors. Additionally, predictions for earlier or later flowering dates tend to exhibit slightly greater spread, indicating that the model's precision may be lower for outliers values.

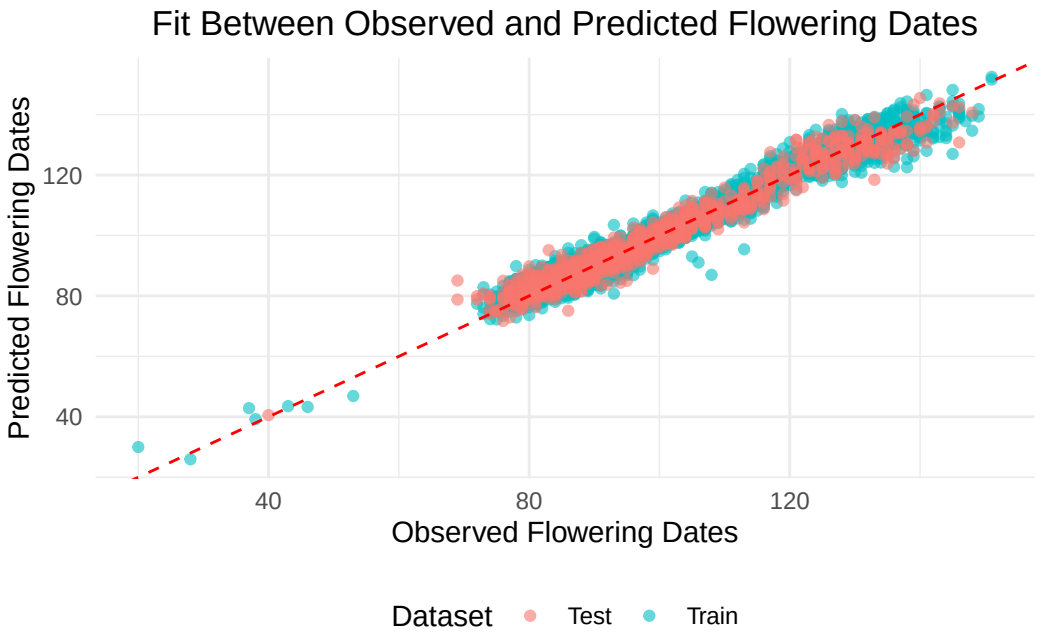


Figure 7: Observed vs. Predicted Flowering Dates

4.2.2 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

	Metric	Training	Testing
RMSE	RMSE	3.103	3.396
MAE	MAE	2.365	2.586
R2	R^2	0.965	0.957

4.2.3 Posterior: Fixed Effects Coefficients with 95% Credible Intervals

The Table 4 highlights the relationships between key predictors and sakura flowering dates based on the Bayesian hierarchical model and the distribution of posterior are shown in Figure 8. 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.

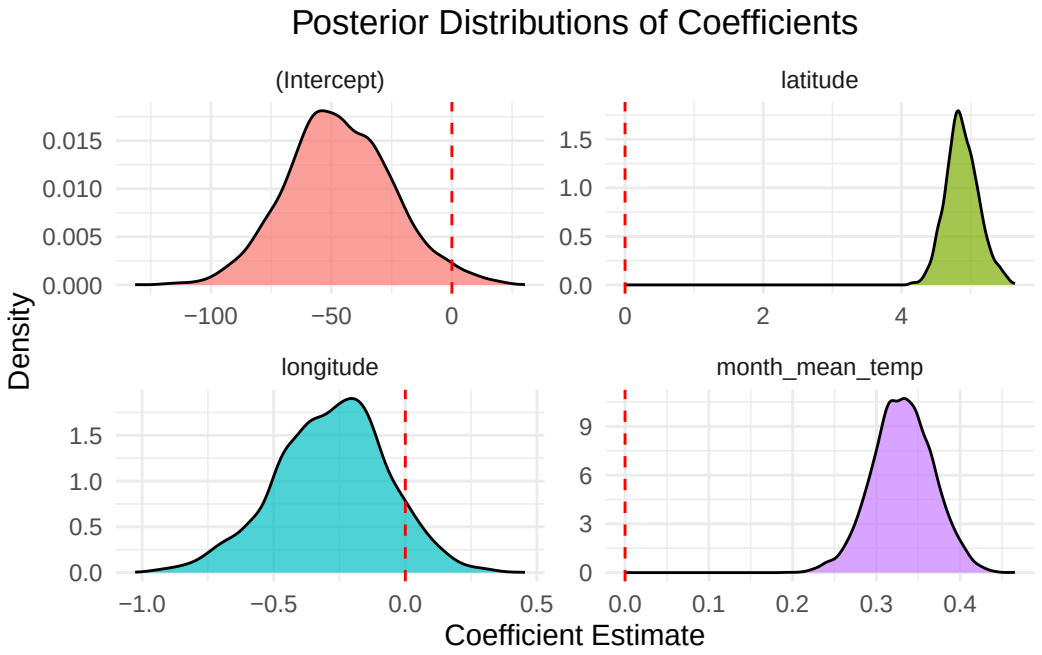


Figure 8: Posterior Distributions

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

4.2.4 Posterior: 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

Metric	Value
RMSE	4.282
MAE	3.186
Variance Explained (R^2)	1.000

5 Discussion

5.1 Interpretation and Implications

Temperature emerged as the most significant predictor, with higher monthly mean temperatures accelerating flowering dates. This aligns with established ecological theories linking

warming trends to phenological shifts in plant species. Latitude also showed a strong positive effect, reflecting the gradient of delayed flowering in colder, northern regions compared to southern ones. Interestingly, longitude had a less pronounced effect, suggesting that the east-west climatic variations within Japan are relatively minor compared to the north-south differences. These findings emphasize the importance of geographic and climatic factors in shaping flowering patterns and suggest that warming trends will continue to drive earlier sakura blooming in future decades.

5.2 History and Modern Sakura

Since the history_data only documents the blooming status of cherry blossoms in Kyoto and the tem_data records March temperatures, in Figure 9, we analyze data from sakura_data and tem_data to focus on Kyoto's blooming trends and climate in March.

The left panel illustrates that historically, cherry blossoms in Kyoto typically bloomed between the 90th and 120th days of the year, corresponding to late March to early April. In contrast, modern flowering days have shifted earlier, often occurring before day 100.

Meanwhile, the right panel shows that for much of history, March temperatures fluctuated between 2.5°C and 10°C. However, in recent years, temperatures have risen sharply, now ranging from 5°C to 12°C. This rapid temperature increase aligns with the earlier blooming days depicted in the left panel, highlighting a clear and significant relationship between rising temperatures and cherry blossom blooming in Kyoto. This linkage underscores the impact of climate change on seasonal cycles and phenology.

5.3 Limitation

The model demonstrates strong predictive performance but has several limitations that should be addressed. By assuming that all regions respond uniformly to climate variables, the dataset may oversimplify the complex ecological processes influencing cherry blossom flowering. The important factors such as soil moisture variations, urban heat effects, and species-specific traits are not explicitly accounted for, potentially overlooking significant regional nuances. Furthermore, the absence of geographic predictors like elevation and detailed precipitation data likely contributes to larger prediction errors for outliers, particularly in regions with extreme or atypical flowering conditions. Additionally, the dataset does not differentiate between sakura varieties, despite substantial variation in flowering times driven by species-specific biological and environmental factors. This lack of certain specices may reduce the model's precision, especially in areas with diverse sakura species and environmental contexts.

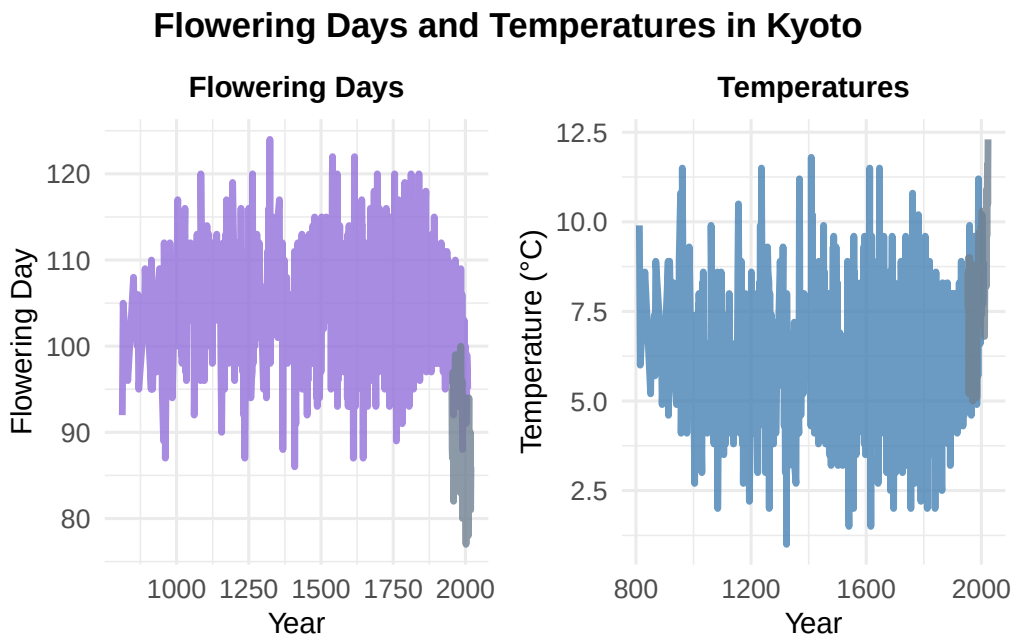


Figure 9: History and Modern

5.4 Implications of the Sakura Prediction Model for Climate Sensitivity and Tourism Planning

The establishment of this prediction model will significantly enhance understanding of sakura's sensitivity to climate, as well as inform tourism and cultural planning associated with sakura. Due to the influence of latitude and temperature, sakura blooms at different times across Japan, with warmer southern regions blooming earlier than cooler northern areas. These regional differences in flowering dates offer an opportunity to better distribute tourist influx in densely visited areas, alleviating pressure on urban infrastructure and facilities. Accurate predictions of flowering times can lead to more effective resource allocation, enabling tourists to plan their trips to Japan at the optimal time. This allows for a more tailored travel experience, with visitors enjoying sakura in different regions at different times, while simultaneously boosting regional economic development. Tourism bureaus can utilize these predictions to adjust event schedules, aligning activities with the expected bloom dates to provide an optimal visitor experience. Additionally, regional collaboration could encourage staggered travel itineraries, where tourists follow the sakura bloom progression from south to north, effectively dispersing demand across time and space. Northern regions with later blooming may benefit from an extended tourist season, whereas central regions such as Kyoto might face challenges with overcrowding if their peak bloom overlaps with other popular destinations.

Beyond its primary function of predicting flowering times, the model incorporates precise fac-

tors such as temperature and geographic location, offering valuable insights into the ecological dynamics of sakura. The marked sensitivity of flowering dates to temperature underscores the characteristic responses of plant phenology to climate change, often referred to as a "natural barometer." By systematically analyzing the relationships between climate factors and flowering dates, the model reveals how sakura responds to seasonal temperature fluctuations, such as earlier blooming due to rising temperatures, as well as regional adaptations influenced by latitude. Furthermore, this model provides a powerful tool for evaluating the long-term impacts of climate change. For example, by simulating future climate scenarios under varying greenhouse gas emission trajectories, the model can predict shifts in sakura phenology over the coming decades. This is essential for understanding the broader implications of climate change on plant phenology, both in Japan and globally. Varieties with heightened sensitivity to temperature changes may face greater survival pressures as climate change intensifies, and insights from the model can guide conservation strategies, helping to identify priority regions and species that require focused protection efforts.

5.5 Future Directions for Improving the Sakura Prediction Model

For a more precise prediction, several improvements should be implemented. First, incorporating additional predictors such as elevation and microclimate factors would enhance accuracy by accounting for localized influences on blooming, particularly in regions with diverse topography or urban heat effects. Including soil conditions and precipitation data would provide a deeper understanding of environmental factors affecting sakura phenology, especially in areas vulnerable to drought or irregular rainfall. Second, expanding the spatial and temporal coverage of the dataset is crucial. Integrating records from other sakura-growing regions, such as South Korea, China, and the United States, would enable cross-regional comparisons and distinguish universal trends from region-specific climate responses. Extending the dataset with more historical and continuously updated records would also facilitate a thorough analysis of long-term flowering trends. Lastly, adopting dynamic or time-series modeling techniques would improve the model's ability to capture year-to-year variability and respond to sudden climate anomalies, such as extreme weather events. Methods like dynamic Bayesian networks or ARIMA models could enhance predictions in increasingly erratic climatic conditions. Collectively, these advancements would make the model more accurate, flexible, and insightful for understanding sakura phenology and its interactions with climate change.

A Appendix

A.1 Additional data details

A.1.1 Survey design and sampling techniques

The survey design for this study integrates historical and modern sakura flowering datasets to provide a robust framework for analyzing long-term trends and regional variability. The historical data, spanning 812 CE to 2015, focuses on Kyoto, where flowering dates have been reconstructed using historical documents, such as diaries and records of hanami events. These reconstructions were validated against observed data starting in 1881, ensuring reliability. Sampling for historical data relies on curated sources from specific studies and publications, offering consistent insights into the relationship between March temperatures and flowering dates over centuries.

But several sampling bias are not neglectble that may impact its accuracy and interpretation. Extreme climatic events, sakura specious, unusual flowering years can disproportionately influence the dataset, introducing outlier bias that skews trends and complicates the relationship between temperature and flowering dates. Additionally, the dataset's exclusive focus on Kyoto reflects selection bias, as it fails to capture regional variations in flowering patterns across Japan. This narrow scope, combined with the reliance on historical records from elite or literate groups, limits its broader applicability.

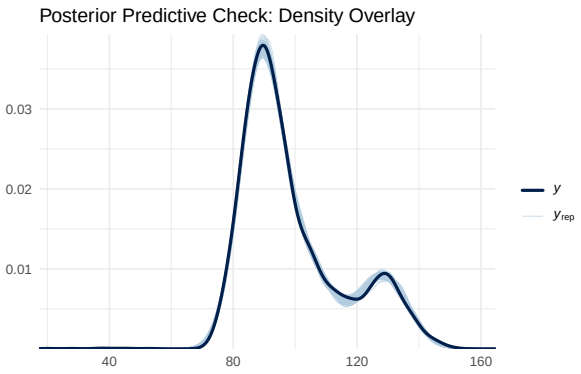
The modern dataset, covering 1953 to 2019, uses data collected by the Japan Meteorological Agency (Agency 2024). The sampling design includes a network of meteorological stations spread across Japan, capturing sakura flowering and full bloom dates from Kyushu in the south to Hokkaido in the north. Monthly mean air temperature data is also included to account for regional climatic variability. Sampling stations are geographically distributed, ensuring representation of diverse climatic zones and topographies. This spatially extensive design captures the progression of the sakura zensen (sakura blossom front) and provides granular insights into phenological differences across regions.Biases in how flowering dates are defined can also lead to confirmation bias, as recorded dates may reflect individual recorders' subjective definitions or expectations of what constitutes "flowering." This inconsistency can result in deviations that align more with personal or cultural interpretations than objective phenological events, further complicating the reliability of historical records. It is also the reason, when we build the model, we only use the data from modern_sakura_data, to ensure high accuracy and reliability for prediction.

In conclusion, the historical sampling focuses on long-term climatic trends in a single location (Kyoto), while modern sampling emphasizes spatial variability and regional climate sensitivity. Together, these approaches enable a comprehensive analysis of how sakura flowering responds to both historical climate trends and modern regional dynamics, ensuring robust and representative findings.

A.2 Model details

A.3 Posterior predictive check

In Figure 10 we implement a posterior predictive check and the distribution of posterior are show in Figure 8. This shows the density overlay plot demonstrates a strong alignment between the observed sakura flowering dates and the model's posterior predictions. The observed density, represented by the dark blue line, closely matches the predictive densities (lighter blue lines) across most of the distribution, indicating that the model captures the central tendency and major patterns effectively. The primary peak around 80 is well-represented, showing the model's strength in predicting the most common flowering dates. Additionally, the secondary peak around 120 aligns reasonably well, suggesting the model's ability to handle more complex patterns, such as bimodal distributions. However, slight discrepancies are observed in the tails of the distribution, particularly below 40 and above 140, where the predictive densities deviate from the observed data. These divergences indicate potential challenges in capturing extreme values or outliers, which may be addressed by including additional predictors, such as microclimatic conditions or species-specific traits.



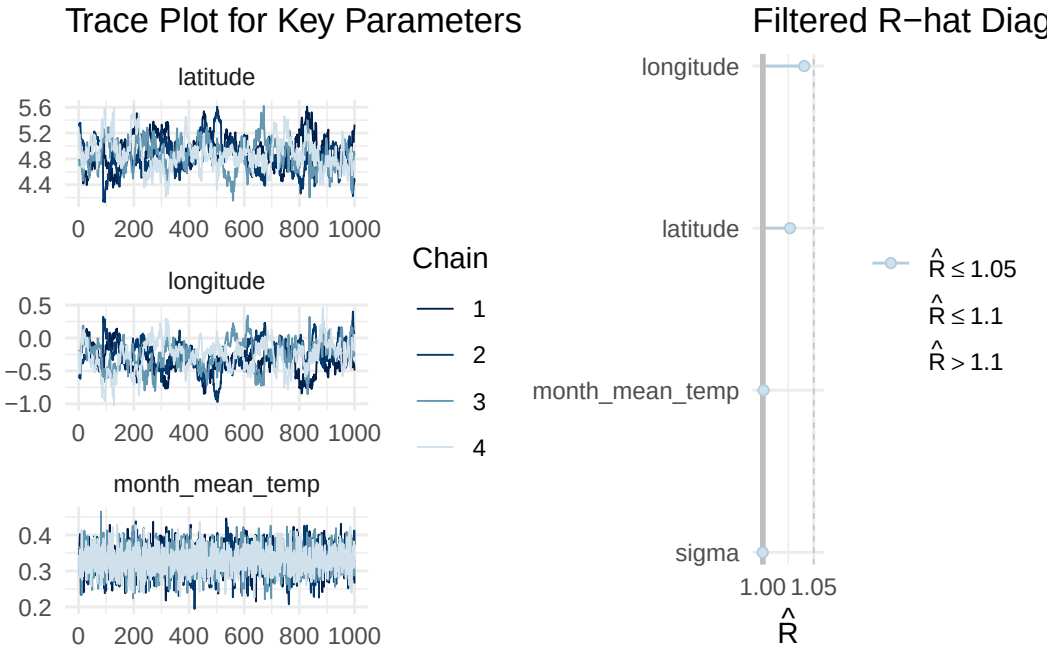
(a) Posterior prediction check

Figure 10: Examining how the model fits, and is affected by, the data

A.4 Diagnostics

Figure 11 The MCMC diagnostics provide strong evidence of convergence and reliability in the posterior estimates. The trace plots for key parameters, including latitude, longitude, and month_mean_temp, demonstrate good mixing across all four chains, with no discernible trends or drift over iterations. Each parameter fluctuates within a stable range, reflecting consistent exploration of the posterior distribution. For instance, latitude ranges between approximately 4.5 and 5.6, while longitude and month_mean_temp remain similarly stable. The R-hat diagnostics further confirm convergence, with all values close to 1.0, indicating that

the chains are well-mixed and have reached equilibrium. These results collectively validate the robustness of the MCMC algorithm, ensuring that the posterior estimates accurately reflect the relationships between the predictors and sakura flowering dates. This provides a solid foundation for subsequent inferences and model interpretation.



(a) Trace Plot for Key Parameters

Figure 11: MCMC Convergence Diagnostics

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/.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics.* https://CRAN.R-project.org/package=gridExtra.

Cepeda, Gabriel A., Jonah Gabry, Ben Goodrich, Andrew Gelman, Aki Vehtari, and Stan Development Team. 2023. *rstanarm: Bayesian Applied Regression Modeling via Stan.* https://mc-stan.org/rstanarm/.

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

Gabry, Jonah, Ben Goodrich, Aki Vehtari, Michael Betancourt, and Stan Development Team. 2023. *bayesplot: Plotting for Bayesian Models.* https://mc-stan.org/bayesplot/.

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.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots.* https://CRAN.R-project.org/package=patchwork.

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

Richard A. Becker, Original S code by, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. 2024. *Maps: Draw Geographical Maps.* https://CRAN.R-project.org/package=maps.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow.

2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with r, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 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.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Yasuyuki, Aono. 2015. *Cherry blossom phenology and temperature reconstructions at Kyoto*. . http://atmenv.envi.osakafu-u.ac.jp/aono/kyophenotemp4/.