Predicting Sakura Florescence in Japan with a Dual-Model Approach*

Forecasting Bloom Duration Using Geographic, Time-based, and Temporal Data

Shaotong (Max) Li

November 17, 2024

This study uses a dual-linear-model approach to predict cherry bloom duration in Japan. The first model predicts bloom duration using temperature, while the second estimates temperature based on latitude, month, and date of blossom. The findings confirm that rising temperatures lead to earlier bloom periods while latitude and seasonal timing significantly shape temperature dynamics, highlighting the importance of understanding these relationships for ecological conservation, cultural event planning and the broader impacts of climate change on phenological events.

1 Introduction

Cherry blossoms, or sakura, are a powerful symbol of Japan, representing both the fleeting beauty of nature and a cherished cultural heritage (Wikipedia contributors 2024). Each spring, the blooming of cherry blossoms is celebrated across the country, marking a time of renewal and festivity (Wikipedia contributors 2024). However, the timing and duration of these blooms are highly sensitive to climatic conditions, which makes predicting sakura florescence increasingly important in the face of climate change (Ocko 2024). Understanding how environmental factors, such as temperature, latitude, and seasonal timing, influence bloom duration can help forecast these events and provide intuition into broader ecological changes (Ocko 2024).

Existing research has demonstrated the relationship between temperature and bloom timing, but many studies focus on specific regions or lack consideration of geographic and temporal variability. In this paper, both historical and modern phenological data of different places in Japan are utilized into a dual-model framework that predicts bloom duration based on temperature and estimates temperature using geographic and temporal variables. Specifically, the

[*] Code and data are available at: Sakura_Florescence_Prediction.

first model predicts bloom duration based on temperature, while the second model estimates temperature as a function of latitude, month, and date.

The objectives of this study are to: (1) predict the duration of cherry blossom blooms based on temperature, (2) model temperature as a function of latitude, month, and day, and (3) assess how these factors interact to influence bloom timing and duration across Japan. By focusing on these aims, the study seeks to improve predictions of cherry blossom blooms while providing a clearer understanding of the complex interactions between time, geographic and temporal factors factors that influence bloom dynamics, particularly how these factors drive temperature changes that, in turn, affect bloom duration.

This research contributes to the fields of phenology and climate science by the dual-model approach that supports planning for cultural and tourism events, informs conservation strategies, and enhances understanding of the broader impacts of climate change on phenological events across Japan. By examining how geographic and temporal factors influence temperature and, in turn, bloom dynamics, this study highlights the potential ripple effects of shifting seasonal patterns on ecosystems, biodiversity, and human activities reliant on predictable biological events.

The remainder of this paper is structured as follows: Section 2 discusses the data sources and preprocessing methods. Section 3 details the dual-model approach, including the temperature estimation and bloom prediction models. Section 4 presents the results, followed by a discussion in Section 5 on the implications of our findings. Finally, Section 6 concludes with intuitions into future research directions and the broader impact of climate change on cherry blossoms.

2 Data

In this project, we used data from the 'Sakura Flowering' branch of the dataset created by tacookson (tacookson n.d.). This dataset provided cherry bloom and temperature records essential for our analysis. Notably, we did not use the full bloom data included in the dataset, as our focus is on the complete florescence period, specifically the bloom duration.We also used Rohan Alexander's GitHub starter folder (Rohan Alexander 2021) as the foundation for our paper and made minor modifications to tailor it to the specific needs of our study.

In this project, we used R(R Core Team 2023) and a variety of R packages for data processing, analysis, and visualization. Specifically, tidyverse(Wickham et al. 2023g), lubridate(Wickham et al. 2023c), arrow(Richardson et al. 2023), tidyr(Wickham et al. 2023f), dplyr(Wickham et al. 2023a), and readr(Wickham et al. 2023d) were used for data manipulation and cleaning. caTools(P. et al. 2023) supported data splitting, while testthat(Wickham et al. 2023e) facilitated testing of scripts. For geospatial analysis and visualization, sf(Pebesma et al. 2023), rnaturalearth(South et al. 2023a), and rnaturalearthdata(South et al. 2023b) were employed, along with ggplot2(Wickham et al. 2023b), patchwork(Pedersen et al. 2023), and ggthemes(Arnold et al. 2023) for creating and enhancing visualizations. The here(Müller et al. 2023) package streamlined file path management. For dynamic report generation, knitr(Xie et al. 2023) and kableExtra(Zhu et al. 2023) were used, allowing for well-structured and formatted outputs. Together, these packages provided a toolkit for data preparation, analysis, and visualization throughout the study.

2.1 Overview

The data utilized in this study comprises three primary datasets: historical cherry blossom bloom records, modern bloom records, and temperature data. The historical dataset encompasses records of cherry blossom bloom dates spanning several decades in the Kyoto region, offering a long-term perspective on bloom trends in this specific area of Japan. The modern dataset contains recent bloom records from various regions across Japan, capturing current climatic conditions and bloom dynamics. Lastly, the temperature dataset includes temperature readings for various locations across Japan, which are essential for assessing temperature-related effects on bloom duration.

To prepare these datasets for analysis, several data cleaning steps were undertaken. For the historical dataset, the predicted and actual temperature estimates were consolidated into a single “temperature” variable, with the actual temperature given the highest priority. The modern dataset underwent similar steps, where modern bloom records and temperature data were merged based on corresponding time and regional information. A new variable, “mean temperature per month,” was created to represent the average temperature in each region during the bloom month. Additionally, incomplete entries were removed, and irrelevant attributes were filtered out from the two cleaned datasets.For variables with NA values, we remove the entire row of data to ensure the reliability of the analysis.

Two cleaned datasets were created to facilitate the study of relationships between various factors. The analysis dataset was further refined by selecting relevant variables and splitting it into training and testing subsets, with 70% used for training and 30% for testing, to support the development of Model 1 and Model 2. Notably, the “full bloom” data was excluded from this study, as it was irrelevant to our research focus on the complete florescence period, specifically the bloom duration.

By integrating both historical and modern data, our analysis captures both long-term trends and recent changes in bloom timing, providing intuitions into the effects of climate variability on cherry blossom phenology. Two cleaned datasets are used in explorating variable relationships in Section 2 and four analysis datasets are used in model training and validation presented in Section 3 and Section 4.

2.2 Measurement

The transformation of real-world phenomena into dataset entries begins with understanding the cherry blossom bloom dynamics. According to a recent BBC article (BBC Travel 2024), climate change has significantly impacted cherry blossom blooming in Japan, with rising temperatures leading to earlier blooming times. Warmer winters and earlier springs are causing the iconic cherry blossoms to bloom weeks earlier than usual, which has disrupted traditional cultural events that celebrate this natural phenomenon (BBC Travel 2024). This link between rising temperatures and earlier bloom periods illustrates the sensitivity of cherry blossoms to temperature fluctuations, underscoring the role of climate change as a important factor influencing bloom dynamics.

From these historical and modern records, we derived several key environmental factors that may influence the florescence period, including temperature, latitude, and the timing, especially in late winter and early spring. These factors were selected based on their potential impact on bloom dynamics, as discussed in related studies (BBC Travel 2024). Temperature plays an important role, as evidenced by documented earlier blooming trends in response to rising temperatures attributed to climate change (BBC Travel 2024). By considering both historical temperature reconstruction and modern observational data, we aimed to develop an understanding of how these variables impact the bloom duration.

The cleaned dataset enabled us to analyze the interactions between bloom dates and climate factors in a structured manner. Each entry in the dataset represents the transformation of raw observations into analyzable data points—quantifying environmental influences such as average monthly temperature and geographic location. This structured data allowed us to systematically assess the relationships between variables and draw intuitions about the key factors affecting cherry blossom phenology, ultimately enhancing our predictive models of bloom duration.

2.3 Outcome Variables

2.3.1 Florescence

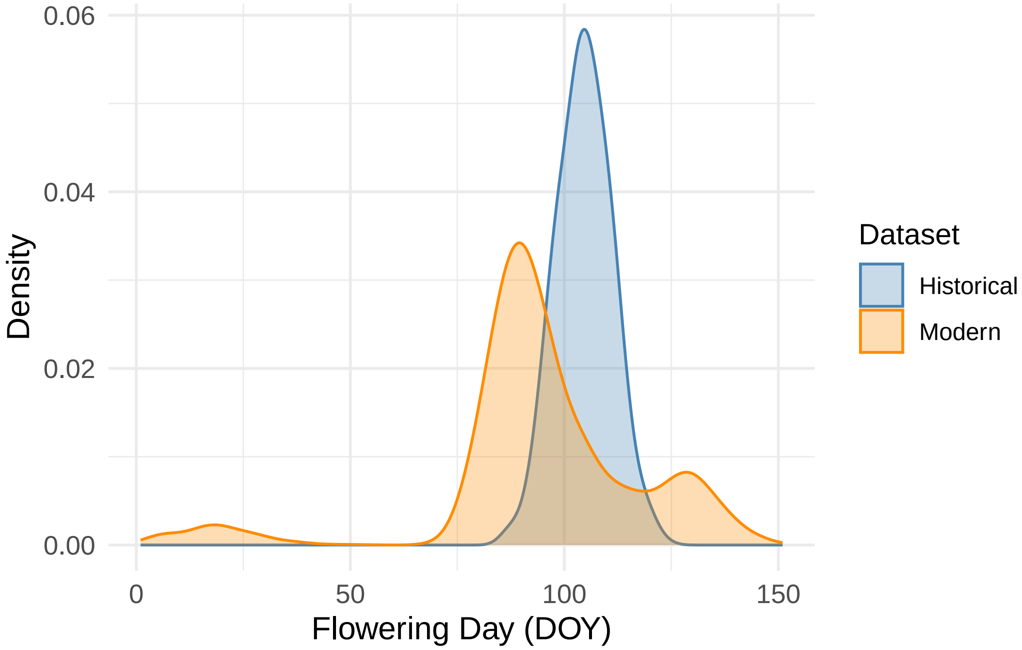

Figure 1: Sakura Florescence: Historical vs Modern Data

Figure 1 illustrates the density distribution of the day of year (DOY) for full bloom, comparing historical and modern datasets. The variable 'florescence' represents the duration of cherry blossom blooming, defined as the number of days during which cherry blossoms are in bloom. In the historical dataset, which focuses on the Kyoto region, florescence data shows a narrower distribution, indicating relatively stable bloom timings over centuries. In contrast, the modern dataset includes bloom data from multiple regions across Japan, resulting in a wider distribution that reflects greater variability due to regional climatic differences.

As the outcome variable in our modeling approach, florescence is essential for understanding bloom dynamics. It allows us to assess the impact of temperature and other environmental factors on the timing and length of cherry blossom blooms, providing intuitions into the effects of climate variability. The stability observed in the historical data can be largely attributed to the consistent climatic conditions in the Kyoto region over centuries. Conversely, the modern data discloses pronounced regional variations in bloom duration, influenced by the diverse environmental conditions across Japan. These variations highlight how regional climate differences have increasingly affected bloom timing in recent years.

2.4 Predictor Variables

2.4.1 Temperature

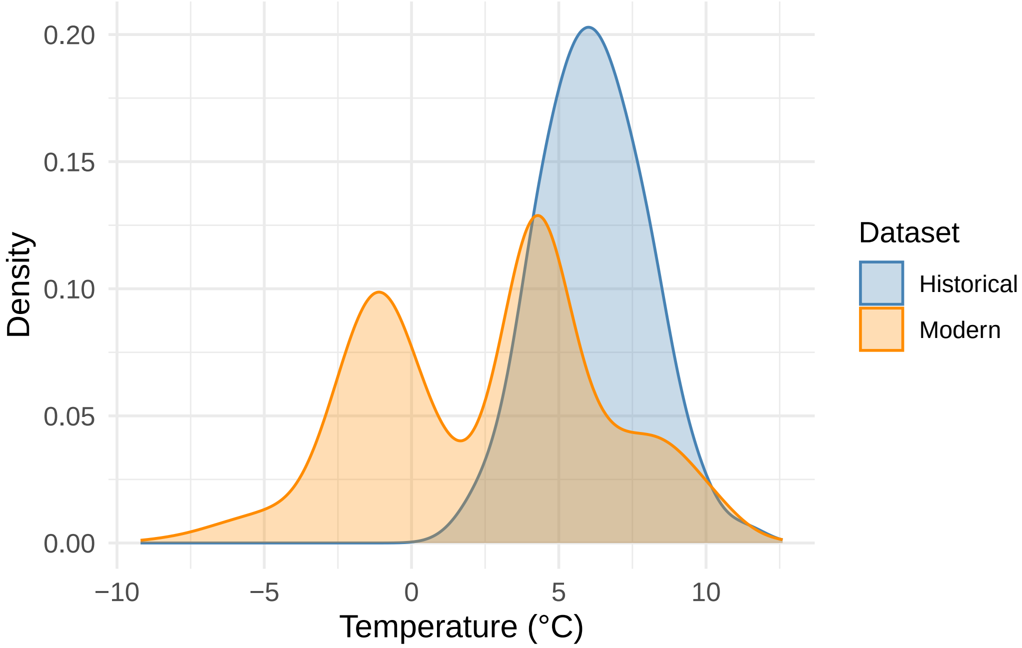

Figure 2: Temperature: Historical vs Modern Data

Temperature plays a essential role in determining the timing and duration of cherry blossom blooms. Figure 2 shows the density distribution of temperatures in historical and modern datasets, highlighting significant differences between the two periods. The historical dataset, primarily focused on the Kyoto region, shows a relatively narrow temperature range centered around moderate values, while the modern dataset presents a broader range with lower temperatures being more prevalent. This difference in temperature distributions likely reflects the increased regional variability in Japan.

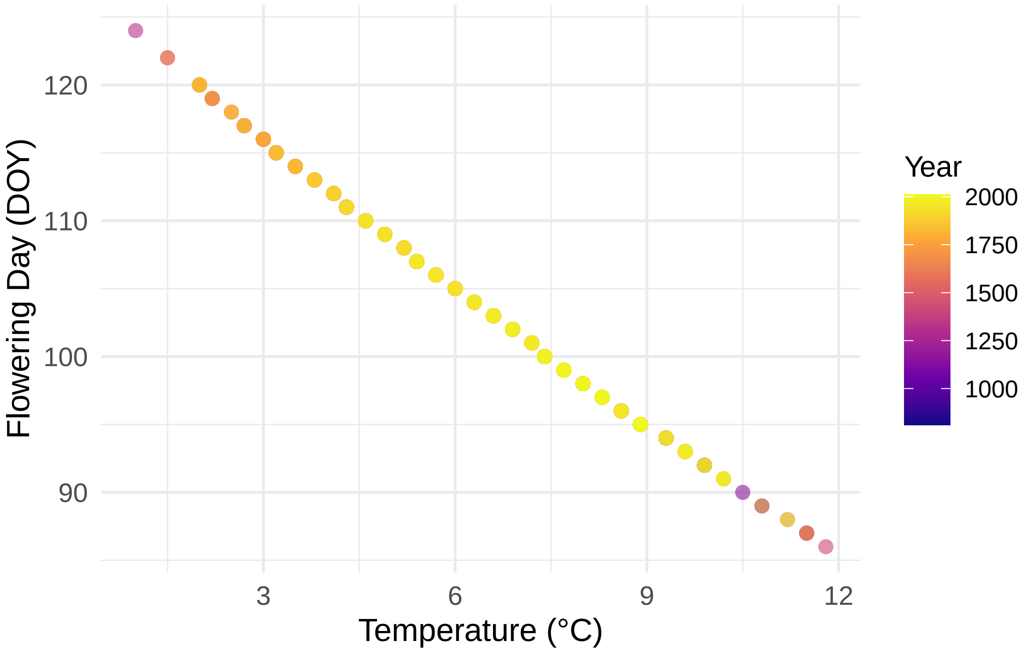

Figure 3: Temperature VS Sakura Florescence

The strong linear relationship between temperature and the day of year (DOY) for florescence period is evident in Figure 3. As temperatures increase, the DOY for blooming decreases, indicating that higher temperatures lead to earlier blooming. This relationship is consistent across both historical and modern datasets, underscoring the important impact of temperature on bloom timing. The temporal progression in Figure 3, represented by the color gradient, illustrates how this relationship has persisted over time, even as overall climatic conditions might have changed.

In our modeling approach, temperature serves as a key predictor for florescence, allowing us to quantify how fluctuations in temperature directly affect the timing and duration of cherry blooms. By examining the historical and modern temperature distributions, we can better understand the impacts of climate variability and identify trends that are essential for predicting future bloom behavior.

2.4.2 Latitude

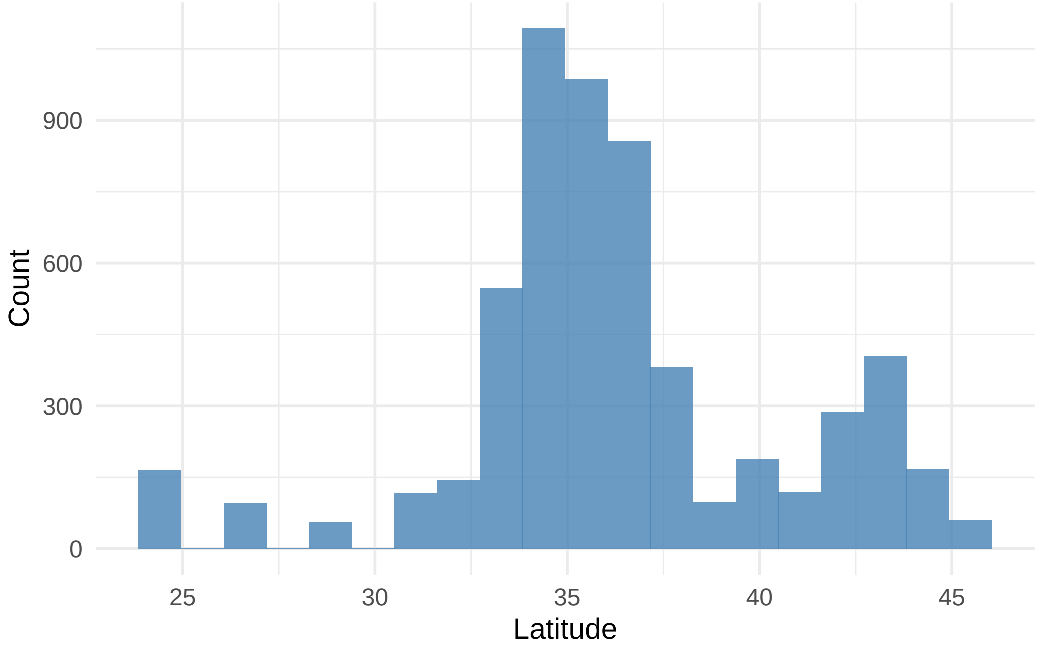

Figure 4: Latitude Distribution

Figure 4 shows the distribution of latitudes for the locations included in the dataset. The majority of observations are concentrated between latitudes 34° and 38° N, which represents the regions most commonly associated with cherry blossom observations in Japan. There is a smaller number of observations at lower and higher latitudes, reflecting the geographic range of sakura coverage in the country.

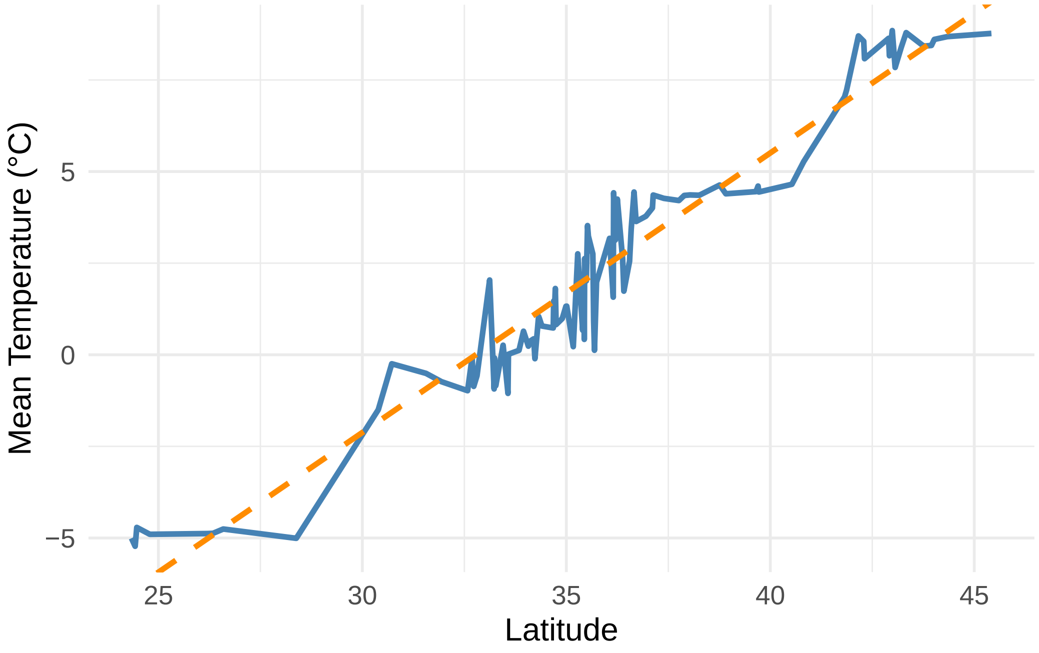

Figure 5: Latitude VS Temperature

Figure 5 illustrates the relationship between latitude and mean temperature, showing a clear positive trend. As latitude increases, so does the average temperature. The dashed orange line represents the linear trend, indicating a consistent increase in temperature as we move towards the northern regions of Japan. This relationship is essential for understanding how geographic position influences local climatic conditions, which in turn affects bloom duration.

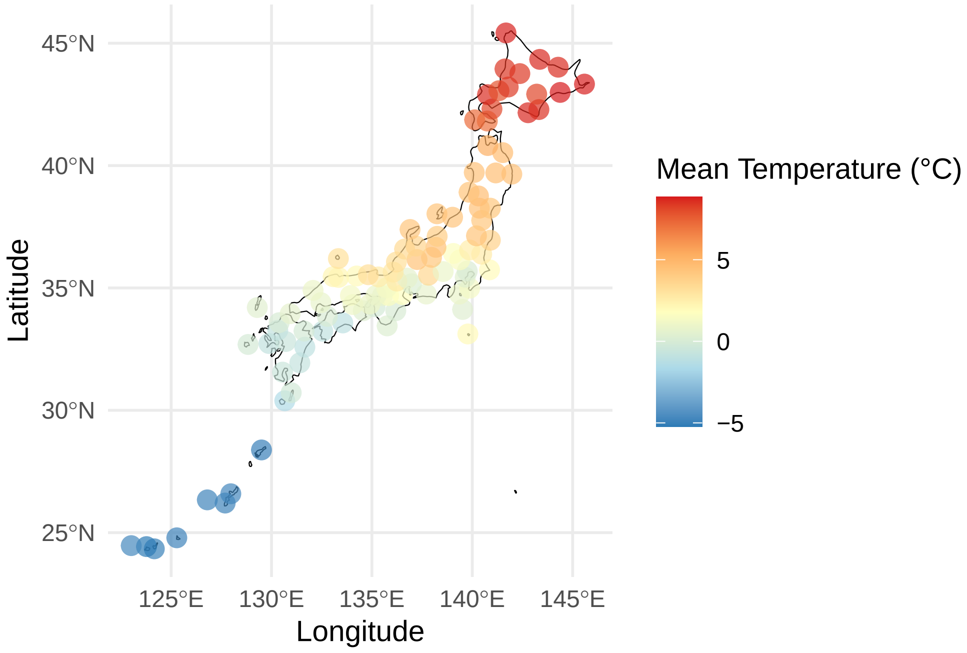

Figure 6: Average Temperatures Across Japan

Figure 6 presents a geographic distribution of mean temperatures across Japan, with warmer colors indicating higher temperatures. This visualization highlights the spatial temperature variability across the country, where northern regions tend to have higher average temperatures compared to southern regions. The color gradient helps to illustrate how temperature conditions vary as a function of both latitude and longitude.

In summary, latitude has a significant influence on temperature, which directly impacts the duration of cherry blooms. The relationship between latitude and temperature is well-captured by Figure 4, Figure 5, and Figure 6, which collectively demonstrate that regions at higher latitudes tend to experience higher average temperatures, contributing to earlier bloom periods. These intuitions are essential for understanding the geographic factors that affect sakura florescence and for improving the accuracy of predictive models.

2.4.3 Time of Blossom

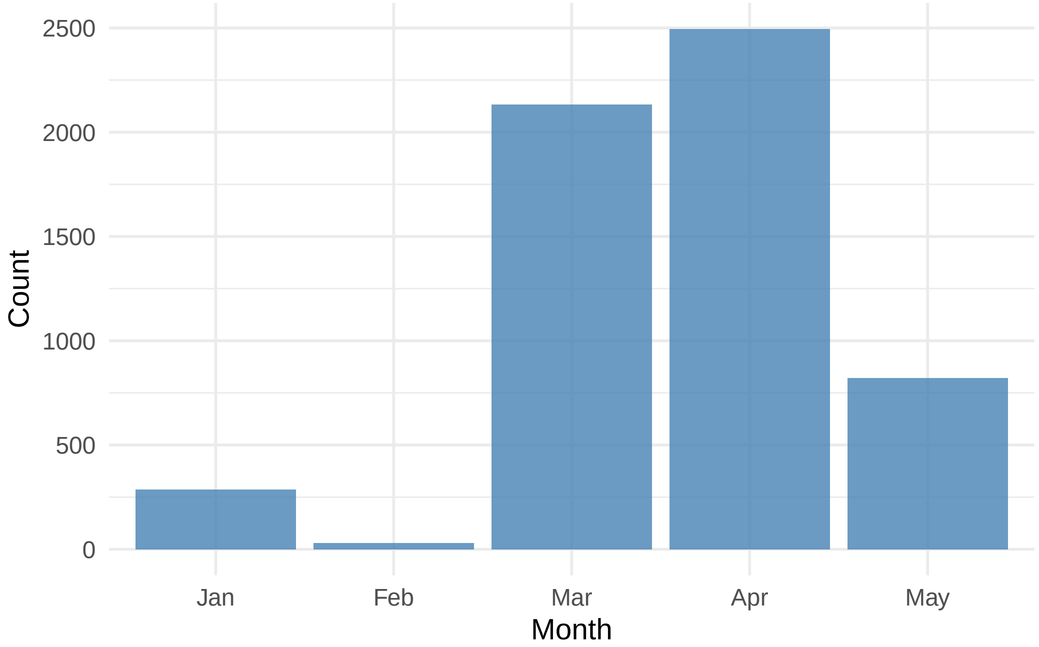

Figure 7: Monthly Distribution of Flower Date

Figure 7 shows the monthly distribution of flower dates, with the majority of cherry blossoms blooming in March and April. This aligns with the typical cherry blossom season in Japan, which occurs in early spring. There are fewer occurrences of blooms in January, February, and May, which indicates that bloom timings outside of this window are uncommon, reflecting the strong seasonal pattern of sakura flowering.

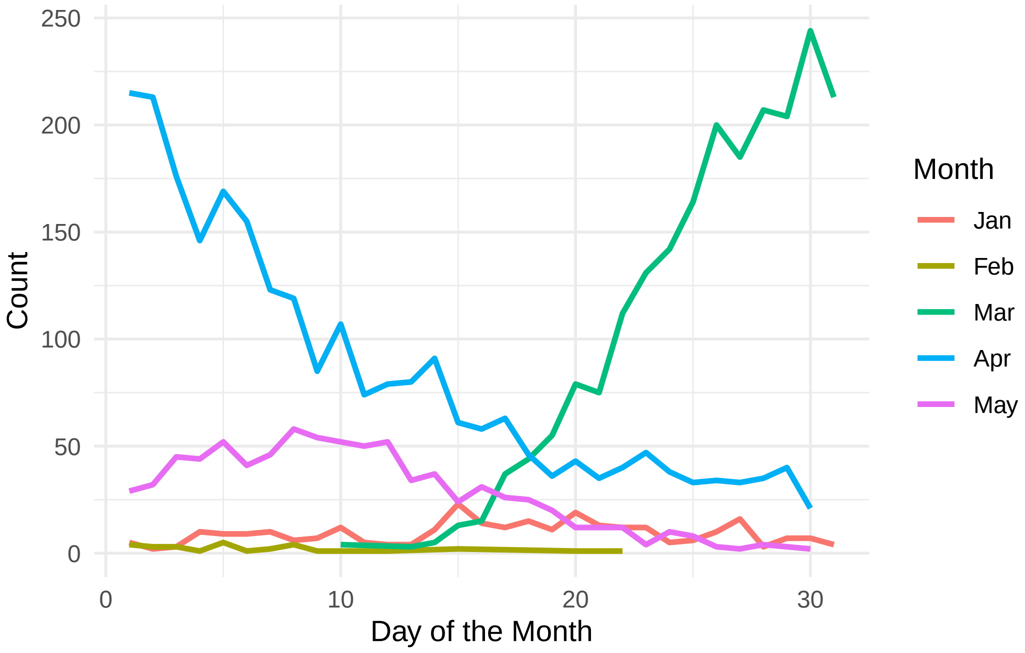

Figure 8: Flower Date Distribution by Month and Day

Figure 8 presents a more detailed view of the flowering dates by month and day. The peak bloom periods are concentrated towards mid March and mid April, with the highest counts occurring in late March. This detailed distribution highlights the specific bloom dates, emphasizing how the bulk of flowering happens within a narrow time frame in early spring, largely driven by favorable temperature conditions.

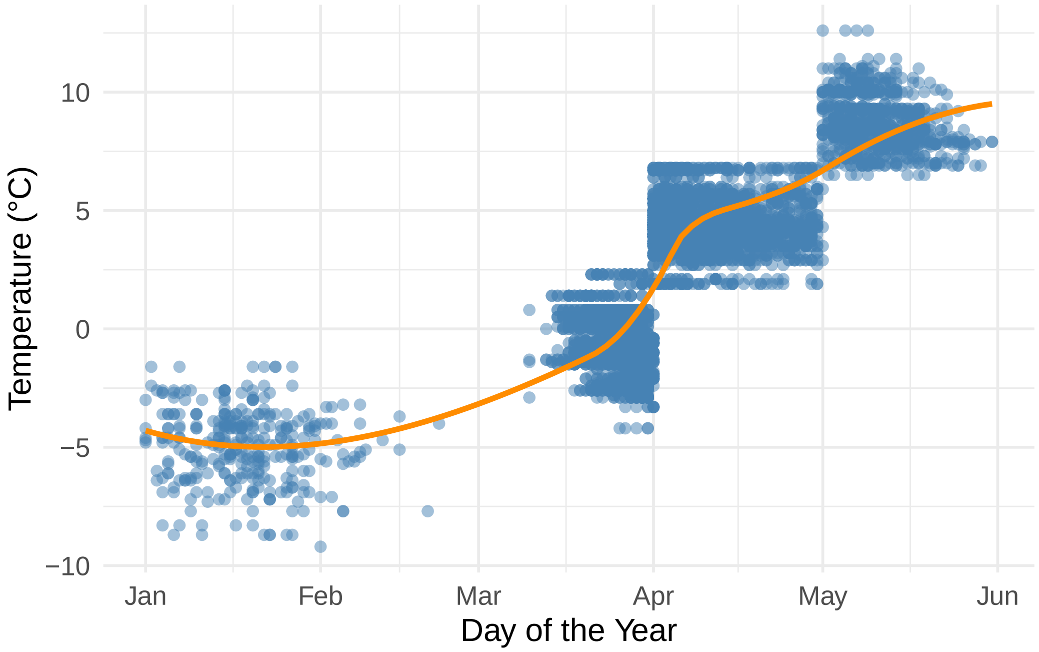

Figure 9: Temperature Distribution by Date

Figure 9 shows the relationship between day of the year (DOY) for blooming and temperature. The trend line suggests a potential linear relationship, where later dates are associated with higher temperatures. This trend indicates that as the year progresses, temperatures rise, which influences the timing of cherry blossom blooms. The scatter plot further emphasizes the variation in temperatures experienced during different bloom periods, illustrating how the blooming date affects the observed temperature.

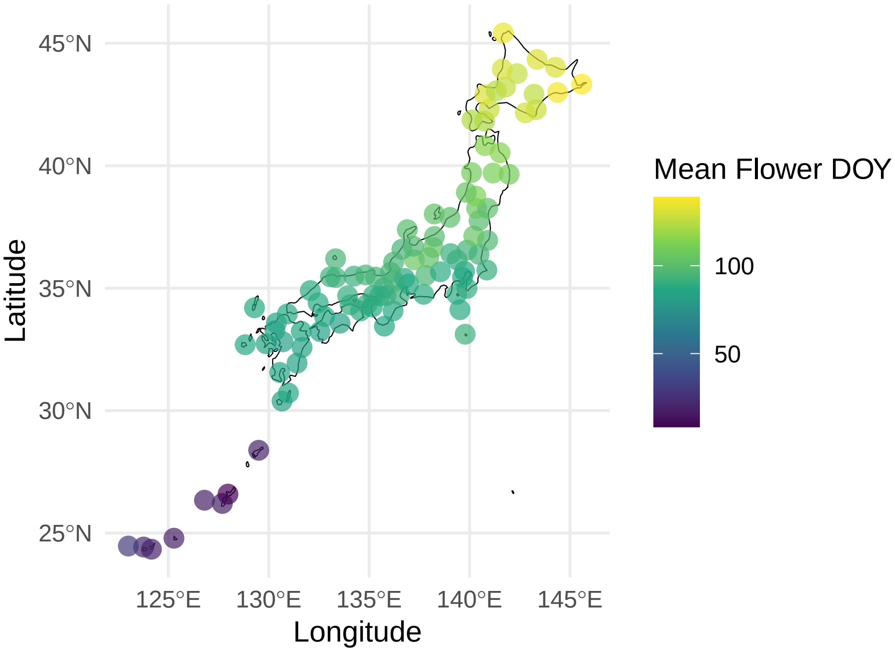

Figure 10: Average Flowering Dates Across Japan

Figure 10 provides a geographic overview of average flowering dates across Japan. The color gradient represents the mean flowering DOY, with darker points indicating later flowering dates. This map illustrates how flowering dates vary geographically, with earlier blooms occurring in warmer, southern regions and later blooms in cooler, northern regions. This geographic variation in bloom dates underscores the influence of temperature gradients across Japan.

In summary, the timing of cherry blossom blooms significantly affects temperature patterns. The majority of cherry blossoms bloom in March and April, which in turn correlates with observed temperature trends. Geographic variation also plays a significant role, with southern regions experiencing earlier blooms compared to northern regions. These intuitions are essential for understanding the temporal dynamics of sakura florescence and for enhancing the accuracy of predictive models that aim to understand temperature dynamics under varying bloom conditions.

2.5 Excluded Variables

2.5.1 Year

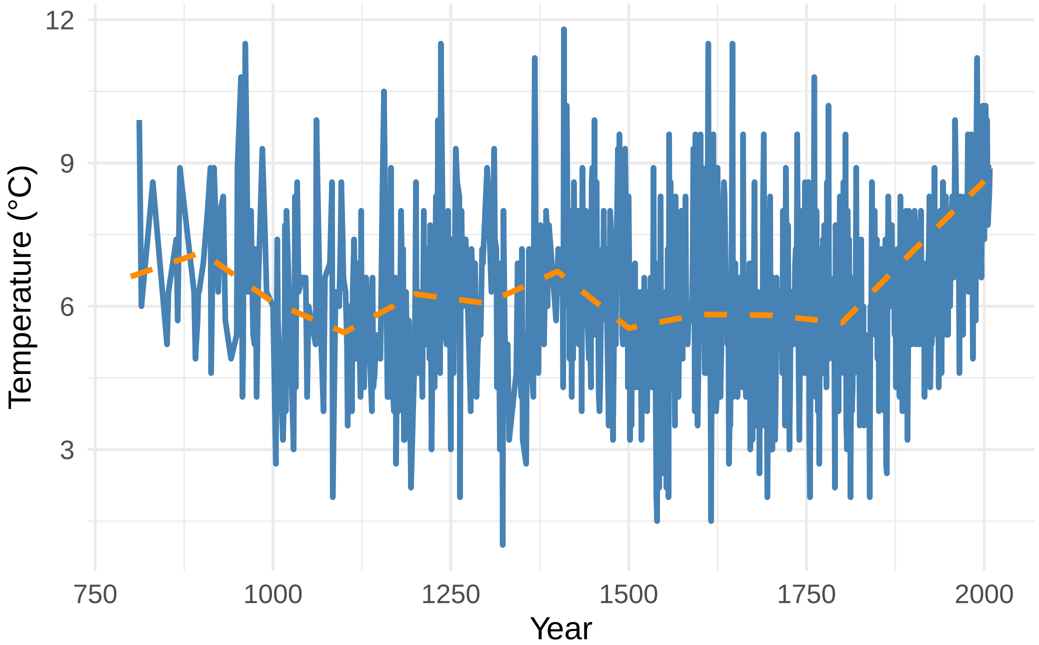

Figure 11: Temperature Trend Over Years

Figure 11 depicts the temperature trend over several centuries, highlighting the distribution of temperatures across different years. The orange dashed line represents the average temperature of each century, providing a baseline for comparison. The plot shows considerable variation in temperatures from year to year, with no consistent linear trend. This lack of a clear relationship between temperature and year suggests that, over the long term, temperature fluctuations have been influenced by a complex interplay of factors beyond a simple time progression.

The data indicates that temperature levels have not exhibited a significant overall change over time. This suggests that, despite certain fluctuations, there is no consistent trend of temperature increase or decrease across the recorded period. Such stability is important for understanding the broader context of climate effects, as it implies that temperature-related changes in bloom timing are influenced more by regional variability rather than long-term shifts. This is why we chose to exclude this variable from further analysis.

2.5.2 Longitude

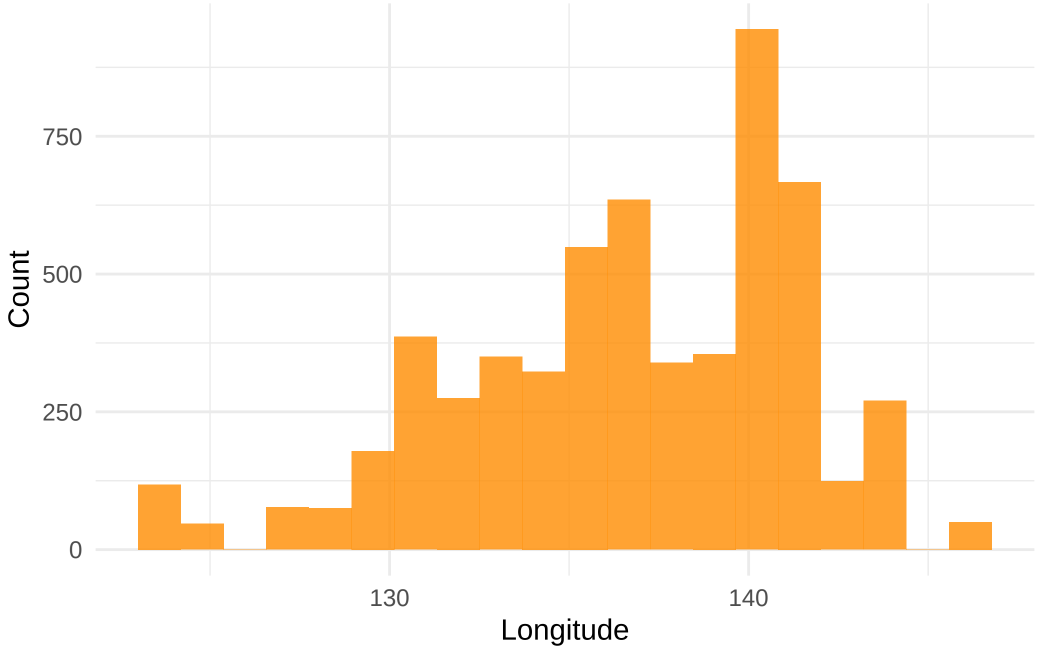

Figure 12: Longitude Distribution

Figure 12 shows the distribution of longitudes for the locations included in the dataset. The majority of the observations are concentrated between 135°E and 141°E, representing regions across central and eastern Japan. This distribution highlights the geographic focus of our study, which includes areas where cherry blossoms are commonly observed and where temperature data is readily available.

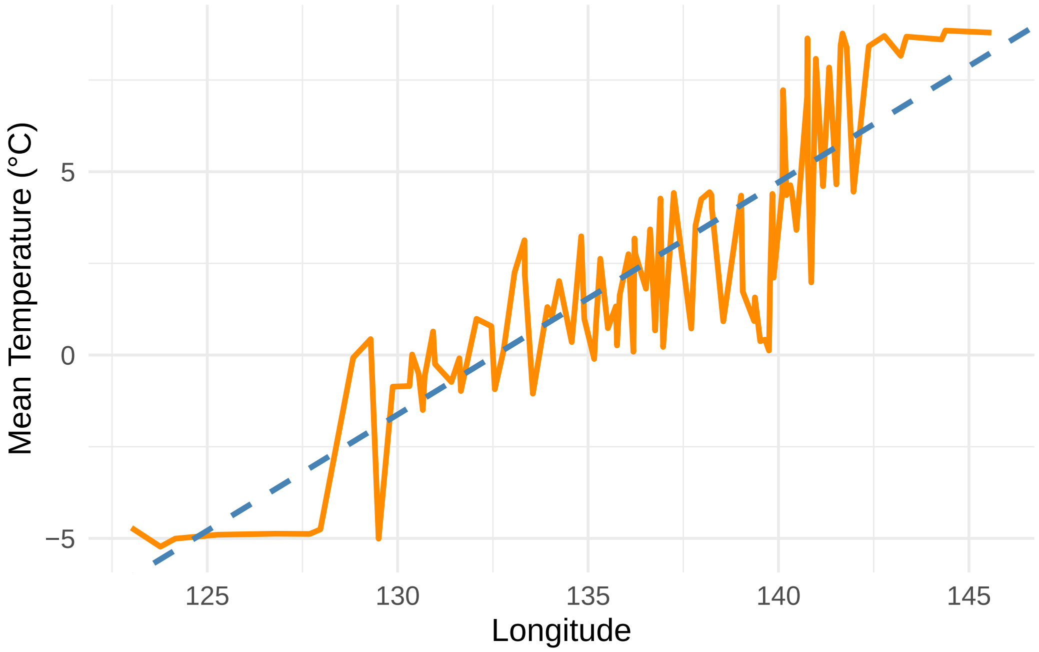

Figure 13: Temperature VS Longitude

Figure 13 illustrates the relationship between longitude and mean temperature. The dashed blue line represents the linear trend, indicating a positive correlation between longitude and temperature. However, compared to latitude, this relationship is weaker and exhibits more variability. The positive trend suggests that regions further east tend to have slightly higher average temperatures, though the variability indicates that other geographic and climatic factors are also at play.

In summary, while longitude does show some relationship with temperature, it is not as strong as the correlation observed with latitude. This weaker relationship and greater variability led us to exclude longitude as a predictor in Model 2. The rationale for this decision will be discussed as an alternative model in Section 3, where we explain the selection of the most relevant variables for predicting temperature and bloom dynamics.

3 Model

We chose to use linear models for both models. Linear models provide a clear structure that makes it easier to interpret the relationships between variables, and they are computationally efficient, allowing for quick predictions. They are also suitable in this study for capturing the basic relationships between temperature and bloom timing, as well as the influence of geographic and temporal factors on temperature, all of which have been shown in previous research to exhibit linear trends. While more complex nonlinear models might explain additional residuals, they tend to sacrifice interpretability and practicality, which could be detrimental for research that aims to balance theoretical understanding with real-world application.

Regarding variable selection, we chose variables based on their established significance in the literature and the availability of data. Temperature, latitude, month, and day were selected as key predictors, while we excluded variables like year, which showed minimal contribution to the predictions. Specifically, the alternative model demonstrates that latitude has a significant and consistent effect on temperature, whereas longitude had a weaker and less significant impact. By excluding variables with low relevance to the outcome, we simplified the model, improving its efficiency and reliability, while avoiding unnecessary complexity that could hinder the interpretation of results.

3.1 Alternative Model

$$\text{Alternative Model: } \text{mean\_temp\_month} = \beta_1 \cdot \text{latitude} + \beta_2 \cdot \text{longitude} + \beta_0 \tag{1}$$

where:

- latitude $\in [20, 50]$, representing geographical latitude in degrees (°N).
- longitude $\in [120, 150]$, representing geographical longitude in degrees (°E).
- β_1 and β_2 are coefficients for latitude and longitude, respectively.
- β_0 is the intercept term.
- Both latitude and longitude are numerical variables.

As shown in equation~1,the objective of this alternative model is to evaluate whether both geographic coordinates, latitude and longitude, contribute significantly to predicting mean monthly temperature. Including both factors allows us to examine their combined effect on temperature estimation, particularly in regions where geographic positioning might lead to climatic variation.

Table 1: Alternative Model Summary

	Variable	Estimate	P-Value
(Intercept)	(Intercept)	-25.1642530	<2e-16
latitude	latitude	0.7781008	<2e-16
longitude	longitude	-0.0033193	0.8063
1	R-squared	0.6721877	
2	Adjusted R-squared	0.6720738	

Table 1 provides the summary of the alternative model, which includes estimates and p-values for both latitude and longitude. The results indicate that latitude has a strong and statistically significant relationship with temperature, whereas longitude does not ($p - value = 0.806$ and $estimates = -0.003$). This finding suggests that latitude is the predominant predictor, which aligns with the analysis in Section 2 that showed a stronger correlation between latitude and temperature compared to longitude.

Given these findings, incorporating longitude into the model did not substantially enhance the temperature estimation, and thus it was excluded from the final temperature estimation model (Model 2). The decision to exclude longitude was based on the statistical insignificance and the simplicity of maintaining an interpretable model without sacrificing accuracy.

3.2 Model 1

Model 1,as shown in equation~2, is intended to predict the duration of cherry blossom blooms, represented by flower_doy (day of year), based solely on the temperature variable:

$$\text{Model 1: flower\_doy} = \beta_1 \cdot \text{temp} + \beta_0 \tag{2}$$

where:

- temp $\in [1, 12]$, representing the temperature range in degrees Celsius (°C).
- β_1 is the coefficient of the variable temp, and β_0 is the intercept term.
- The variable temp is a numerical variable.

The primary objective of Model 1 is to quantify the direct effect of temperature on the flowering date of cherry blossoms. This model is straightforward, utilizing only one predictor variable—temperature—which simplifies interpretation and provides an essential baseline understanding of how temperature alone influences bloom dynamics.

Model 1's simplicity makes it particularly useful in understanding temperature's direct role in influencing bloom timing. A linear model was selected to quantify the direct impact of temperature and assess the potential shifts in bloom periods due to changes in temperature. This model also serves as a foundation against which more complex models can be benchmarked, offering intuitions into the fundamental climatic drivers of cherry blossom florescence.

Table 7 summarizes the results of Model 1, including residuals, coefficients, and overall model performance metrics. A detailed breakdown of model diagnostics and performance metrics is available in Section 6 for further reference.

3.3 Model 2

Model 2,as shown in equation~3, is designed to estimate the mean monthly temperature using three predictors: day of the month (day), latitude, and month. The model is represented as follows:

$$\text{Model 2: mean\_temp\_month} = \beta_1 \cdot \text{day} + \beta_2 \cdot \text{latitude} + \beta_3 \cdot \text{month} + \beta_0 \tag{3}$$

where:

- latitude $\in [20, 50]$, representing geographical latitude in degrees (°N).
- day $\in [1, 31]$, representing the day of the month.
- month is a categorical variable representing months (January to May).
- β_1, β_2, β_3 are coefficients of the linear model, and β_0 is the intercept term.
- All variables are numerical variables, except for month, which is a categorical variable.

Model 2 provides a framework for estimating temperature as a function of both spatial and temporal variables. The inclusion of latitude captures the north-south temperature gradient, which is essential given Japan's geographic diversity. The day variable allows for finer temporal resolution within each month, and month as a categorical variable captures the broad seasonal temperature variations.

This model aims to address the complexity of temperature dynamics by accounting for both geographical and temporal variability. By integrating these factors, Model 2 helps to estimate localized temperatures that are subsequently used to predict bloom timing. The decision to incorporate both numerical and categorical predictors ensures that the model reflects the degrees of temperature changes, accounting for both seasonal effects and geographic variations.

Table 8 provides a summary of Model 2, including residuals, coefficients, and overall model performance metrics. The detailed analysis and diagnostics of Model 2 are provided in Section 6, where we evaluate its accuracy and predictive capability in estimating temperature based on the selected predictors.

3.4 Model Validation

There are potential limitations to our model, including the exclusion of other climatic factors such as precipitation, humidity, and soil moisture, potentially influencing bloom timing. Additionally, the models do not account for microclimatic effects or local geographical features such as urban heat islands, which might introduce variability not captured by latitude and temperature alone.

Therefore, to validate the models to see whether the results are largely affected by unknow variables, we split the cleaned data into training and testing sets using a 70-30 ratio.Then, we assess the performance of the models using various diagnostic tools and statistical metrics.Specifically, we include the actual vs. fitted values plot to evaluate how well the predicted values align with the observed data. A strong alignment in this plot indicates high predictive accuracy and suggests that the model captures the underlying trends in the data effectively.

We also use the Q-Q residual plot to assess whether the residuals follow a normal distribution. This helps validate one of the key assumptions of the models, ensuring that the variability and spread of the residuals are appropriate for statistical inference and prediction.

Additionally, we rely on metrics such as RMSE, R-squared, adjusted R-squared, F-statistics, and p-values to quantitatively measure the accuracy and reliability of each model. These metrics provide understanding of the overall fit, predictive power, and statistical significance of the predictors.

Together, these visual and numerical tools offer an evaluation of model performance and help ensure that the models are robust and suitable for the intended analyses.

3.4.1 Model 1 Validation

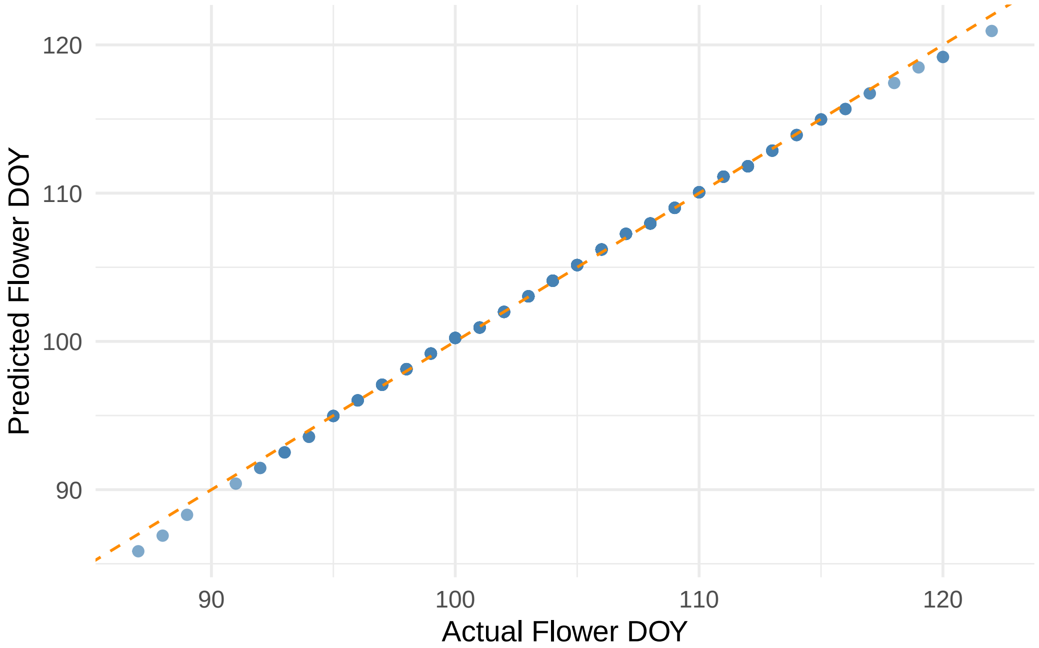

Figure 14: Model 1: Actual vs Predicted Sakura Florescence

Figure 14 depicts the actual versus predicted flowering day of the year (DOY) for cherry blossoms using Model 1. The points are closely aligned along the dashed orange line, indicating that the model predictions are highly accurate.

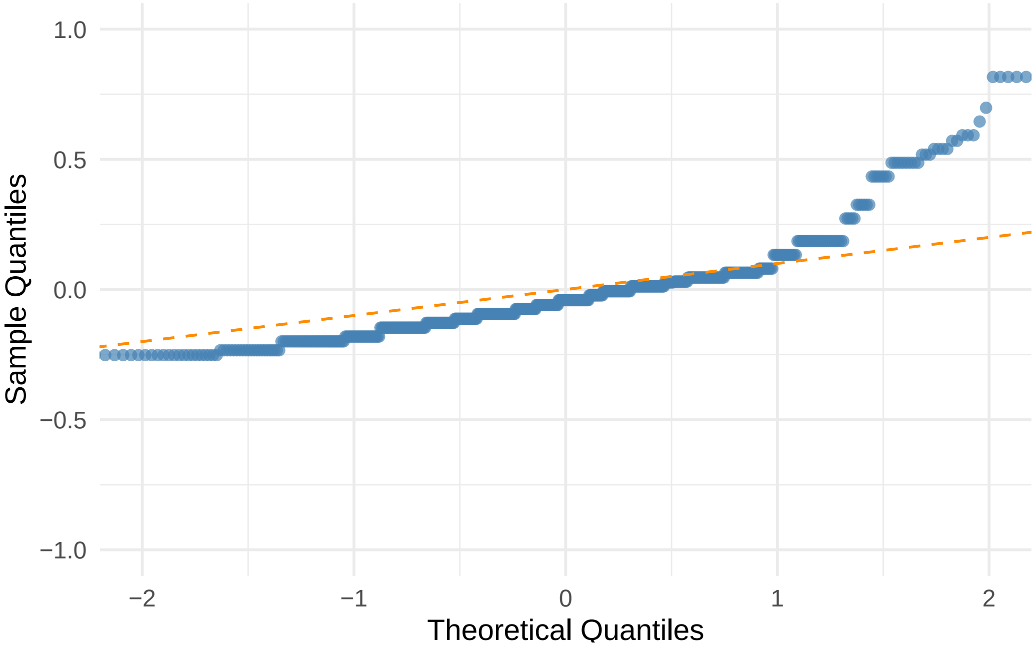

Figure 15: Model 1: QQ Plot of Residuals

Figure 15 depicts the Q-Q residual plot of model 1. The orange line represents a residual slope of 0.1, indicating that the spread of the residuals is much narrower compared to the theoretical normal distribution which means the residuals have a smaller variance than a standard normal distribution.However, this doesn't interfere the accuracy of our model.

Table 2: Model 1: Regression Diagnostics

Metric	Value
Residual Standard Error	0.2337 on 572 degrees of freedom
Multiple R-squared	0.9987
Adjusted R-squared	0.9987
F-statistic	429436.7017 on 1 and 572 DF
P-value	< 2.2e-16

The strong linear relationship observed suggests that temperature is an effective predictor of bloom timing, validating our choice of using temperature as the sole predictor in Model 1.The RMSE, R-squared, adjusted R-squared, F-statistics and overall P-value of model 1 are listed above in Table 2. Model 1 exhibited a low RMSE, high adjusted R-suqared and low overall p-value, indicating strong performance in predicting bloom dates based on temperature. Full summary table is in Section 6.

3.4.2 Model 2 Validation

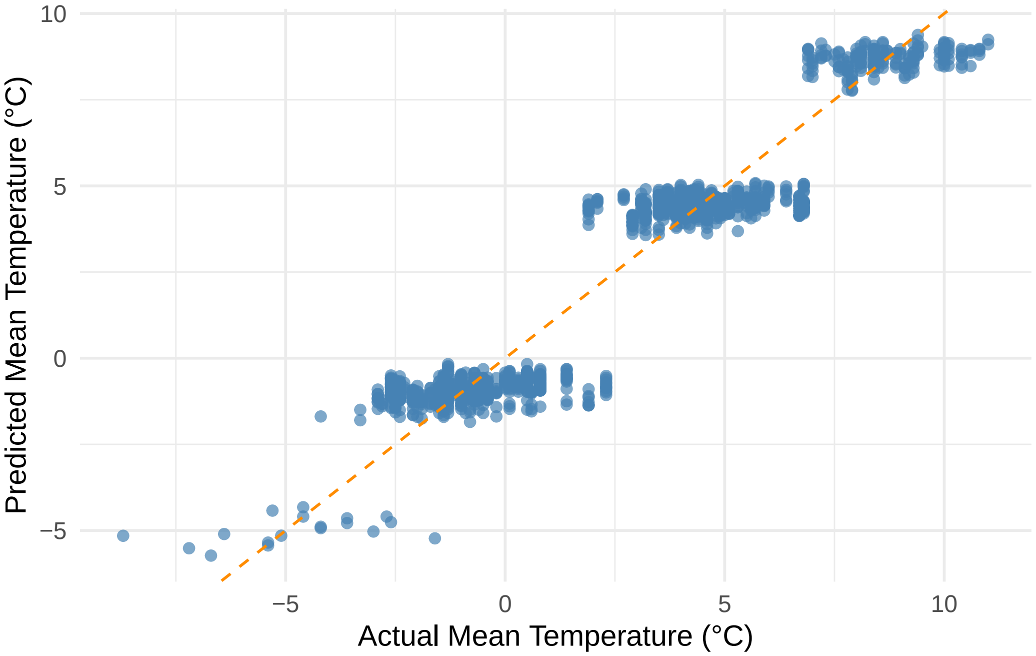

Figure 16: Model 2: Actual vs Predicted Temperature

Figure 16 illustrates the actual versus predicted mean temperature values for Model 2.While there is more variability compared to Model 1 the overall trend aligns well with the dashed orange line, indicating reasonable predictive performance. This can be attributed to the increased complexity of the model and the influence of multiple interacting factors, such as day, latitude, and month, which contribute to temperature variations.

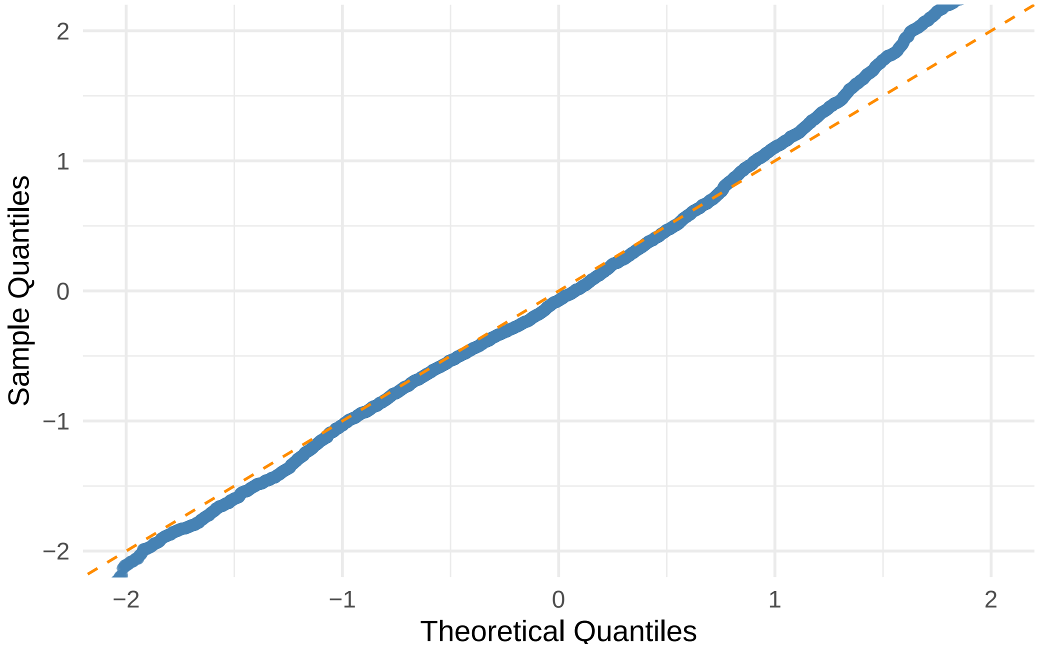

Figure 17: Model 2: QQ Plot of Residuals

Figure 17 depicts the Q-Q residual plot of model 2. The orange line represents a residual slope of 1, indicating that that the variability of the residuals matches that of a standard normal distribution, which is a desirable property.The model adequately captures the underlying relationships in the data and the residual variance reflects the inherent variability in the data without being artificially compressed or inflated.

Table 3: Model 2: Regression Diagnostics

Metric	Value
Residual Standard Error	1.1156 on 4256 degrees of freedom
Multiple R-squared	0.9255
Adjusted R-squared	0.9254
F-statistic	8813.9897 on 6 and 4256 DF
P-value	< 2.2e-16

Nonetheless, the model performs well in capturing the overall pattern of temperature change.The RMSE, R-squared, adjusted R-squared, F-statistics and overall P-value of model 2 are listed above in Table 3. Model 2 also demonstrated reasonable accuracy in estimating mean temperatures,with a low RMSE, high adjusted R-suqared and low overall p-value. Full summary table is in Section 6.

3.4.3 Validation Conclusion

In conclusion, both models show strong predictive capabilities. The results demonstrate the utility of the dual-model approach in enhancing our understanding of the factors driving cherry blossom bloom duration, particularly in the context of climate variability and change.The simplicity of Model 1 and the more subtle complexity of Model 2 offer complementary intuitions into the determinants of cherry blossom bloom duration. Model 1 provides a clear understanding of temperature's direct influence, while Model 2 enhances this by incorporating additional geographic and temporal variables, thereby improving the accuracy of bloom forecasts in the context of changing climatic conditions.

4 Result

4.1 Model Result

The results of the modeling process are summarized in two tables and two figures, highlighting the performance and predictive capabilities of Model 1 and Model 2. In this section, we present and analyze the key findings, utilizing both statistical summaries and visual aids to effectively communicate the outcomes.

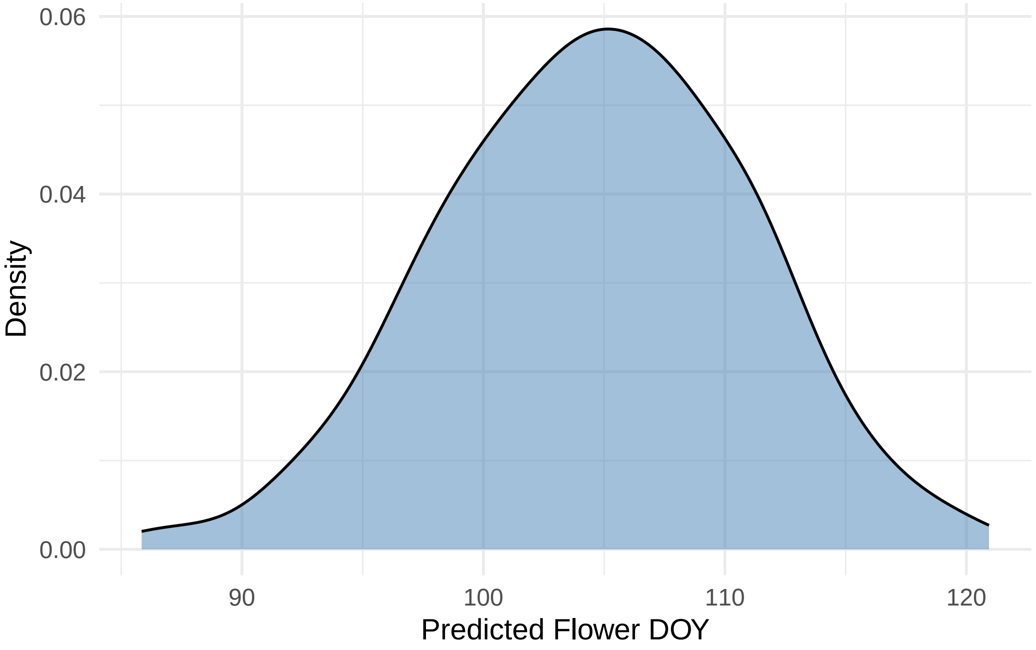

Figure 18: Distribution of Predicted Bloom Duration

Figure 18 illustrates the distribution of the predicted flowering day of the year (DOY) for cherry blossoms using Model 1. The distribution displays a bell-shaped curve, indicating the overall spread and concentration of predicted bloom dates. The model shows consistency in predicting bloom timing, with most values clustering around a central point. This result suggests that the temperature variable, as used in Model 1, is effective in capturing the general trends in cherry blossom bloom timing.

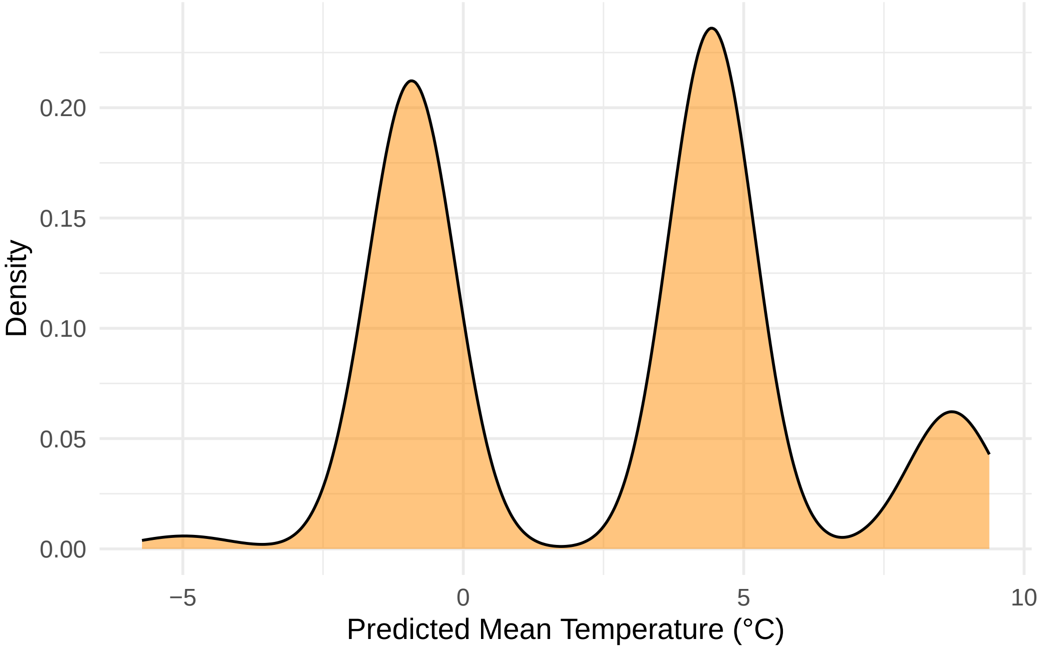

Figure 19: Distribution of Predicted Bloom Temperature

Figure 19 represents the distribution of the predicted mean temperature values, generated by Model 2. The density plot discloses multiple peaks, which may indicate variations in temperature predictions across different regions and times of the year. The observed variability aligns with the geographic and temporal diversity included in the dataset, as Model 2 incorporates latitude, day, and month as predictors. These factors contribute to the complex pattern observed, demonstrating the influence of both geographic location and seasonal changes on temperature dynamics.

Table 4: Coefficients for Model 1

	Estimate	Std. Error	t Value	P-Value
(Intercept)	126.202188	0.0343959	3669.1015	< 2.2e-16
temp	-3.509279	0.0053551	-655.3142	< 2.2e-16

Table 4 presents the coefficients for Model 1, which predicts the bloom day of the year based on temperature. The negative coefficient for temperature (-3.509) indicates that as temperature increases, the flowering day occurs earlier in the year. This aligns with established phenological observations, suggesting that warmer temperatures accelerate the blooming process. The strong statistical significance of the temperature variable (p-value $< 2 \times 10^{-16}$) confirms its importance in the model.

Table 5: Coefficients for Model 2

	Estimate	Std. Error	t Value	P-Value
(Intercept)	-9.0004393	0.2894642	-31.09344	< 2.2e-16
day	-0.0492231	0.0032911	-14.95648	< 2.2e-16
latitude	0.1890333	0.0120889	15.63688	< 2.2e-16
month2	-0.6179283	0.2198314	-2.81092	0.004963
month3	2.8693830	0.1143644	25.08982	< 2.2e-16
month4	6.9661077	0.1649841	42.22291	< 2.2e-16
month5	10.0473632	0.2374526	42.31312	< 2.2e-16

Table 5 provides the coefficients for Model 2, which estimates the mean monthly temperature based on latitude, day, and month. The small p-values for all variables indicate strong statistical significance, confirming their importance in predicting mean temperature (nearly all p-values are less than 2×10^{-16}, with the largest one around 0.005). The coefficient for latitude (0.189) suggests that temperature increases with latitude, albeit at a modest rate. The day variable has a small negative coefficient (-0.049), indicating a slight decrease in temperature as the month progresses. The month coefficients, treated as categorical, reflect expected seasonal changes in temperature, with notable increases from March to May (from around -0.62 to 10.05). The significance of all predictors underscores their relevance in estimating temperature variations, suggesting that both geographic and temporal factors play important roles in determining temperature dynamics.

In conclusion, both models provide substantial intuitions into the factors influencing cherry blossom bloom duration. Model 1 effectively captures the direct impact of temperature on bloom timing, while Model 2 incorporates geographic and temporal variables to estimate temperature. The dual-model approach offers an understanding of the climatic and geographic determinants of cherry blossom phenology, contributing to our broader understanding of the effects of climate variability on natural events.

4.2 Practical Applications

To provide a practical example of how our models perform, we selected a specific data point from the raw dataset representing the city of Esashi in 1969. The data for this location included all relevant attributes needed for the models, such as latitude, day, month, and mean temperature. Using these values, we predicted the mean temperature and flowering day of the year (DOY), then compared the predicted values to the actual historical data.

Table 6: Prediction Compared to Actual Data

Metric	Actual	Predicted
Mean Temperature (°C)	Esashi	NA
Flower DOY	40.82	NA
City	140.77	NA
Latitude	3.9	4.254538
Longitude	119	111.271828

Table 6 presents the information of selected data point as well as the actual versus predicted values for mean temperature and flowering DOY. The results show that the predicted mean temperature is $4.25°C$, compared to the actual value of $3.9°C$. For flowering DOY, the model predicted a value around 111 days, while the actual recorded value was 119 days. This example demonstrates that our models are highly accurate in predicting both temperature and bloom timing, effectively capturing the relationship between environmental factors and cherry blossom florescence.This prediction example highlights the practical application of the models developed in this study and their potential to provide meaningful intuitions into the timing and conditions of cherry blossom blooms in Japan.

By combining Model 1, which estimates bloom timing based on temperature, with Model 2, which predicts temperature using latitude, day, and month, this framework effectively captures the interplay between environmental and geographic factors in bloom prediction. These predictions can aid in organizing cultural events and managing tourism centered around cherry blossoms, while also serving as a foundation for ecological research and conservation planning. Furthermore, the models can be applied to study how climate change may influence seasonal events, offering information to researchers and policymakers as they address potential shifts in timing and their broader implications.

5 Discussion

5.1 Achievements of This Study

This study presented a dual-model framework to forecast the blooming duration of cherry blossoms (sakura) in Japan. The first model established a predictive relationship between temperature and bloom duration, while the second model estimated temperature based on geographic and temporal factors, such as latitude, day, and month. By leveraging historical and modern datasets, the study provided a robust method for understanding how environmental variables influence bloom dynamics. The findings offer practical applications for managing cultural events, ecological conservation, and planning in the context of climate variability.

5.2 Environmental and Phenological Dynamics

The findings from this study provide substantial intuitions into the factors that influence cherry blossom bloom timing in Japan. One key takeaway is the strong relationship between temperature and bloom duration, confirming the important role of climatic conditions in determining the timing of sakura florescence. This information is particularly relevant in the context of climate change, as rising temperatures could lead to earlier and potentially shorter bloom periods, impacting both cultural and ecological aspects of cherry blossom events. The implications extend to tourism, agriculture, and ecological conservation, as shifting bloom timings could affect local economies and biodiversity.

Another important finding is the relative importance of geographic factors, such as latitude, in determining local temperatures. Our results suggest that latitude plays a more significant role than longitude in temperature estimation, leading us to exclude longitude from the final model. This decision reflects a balance between model complexity and predictive accuracy, ensuring that the model remains interpretable while retaining its predictive power. It also underscores the importance of simplifying models without sacrificing key predictive capabilities, which can be especially useful in practical applications such as ecological forecasting.

5.3 Limitation

Despite the promising results of our models, several limitations must be acknowledged. First, the first model relies heavily on the availability and accuracy of historical phenological data. However, the historical data predominantly originates from the Kyoto region, which may limit its generalizability to other areas in Japan. While the results show no significant decrease in prediction accuracy when applied to broader datasets, regional variations in climate and geography may introduce potential biases when the model is used in locations with different environmental conditions.

Temperature was identified as a strong predictor of bloom timing, but it is not the sole factor influencing cherry blossom florescence. Other variables, such as precipitation, soil moisture, light exposure, and regional influences like microclimates or urban heat islands, may also play important roles but were not included due to data limitations. These omissions reduce the models' universality and may contribute to the deviations observed in Model 2's predictions.

Additionally, the reliance on linear assumptions may oversimplify the complex interactions among these factors, as seen in Figure 16. Future model improvements should incorporate additional environmental variables, account for regional influences, and explore non-linear or machine learning approaches to enhance accuracy and robustness.

5.4 Future Research Directions

Future research should explore additional environmental variables that may affect bloom timing, such as soil moisture, precipitation, and light availability. Incorporating these factors could improve model accuracy and provide an understanding of the factors driving sakura florescence. Furthermore, future studies could consider non-linear models or machine learning approaches to capture complex interactions between variables, potentially enhancing predictive capabilities. Machine learning models, for instance, could uncover hidden patterns and relationships in the data that are not easily captured by traditional linear approaches, offering deeper intuitions into the dynamics of cherry blossom blooming.

It would be beneficial to extend the scope of the analysis beyond Japan to examine cherry blossom bloom patterns in other regions, such as South Korea or China. By comparing bloom dynamics across different climates and geographies, we could gain a deeper understanding of the global factors affecting cherry blossom phenology and assess the broader implications of climate change on these iconic events. Such comparative studies could disclose how different environmental and cultural contexts modulate the response of cherry blossoms to climatic changes, thereby contributing to a more global understanding of phenological shifts.

6 Appendix

6.1 Survey Methodology Overview

In this section, we provide an overview of the survey methodologies relevant to this study, focusing on their design, sampling strategies, and implications for data quality. Surveys serve as a useful tool for capturing public sentiment and translating it into actionable intuitions, especially in environmental research where individual perceptions and regional trends play an essential role.

Our primary data collection involved historical and modern records of cherry blossom blooming and temperature data. To complement these observations, an ideal survey could be implemented to gauge local experiences of sakura blooming and capture potential environmental influences not represented in raw data. This survey would be designed using a combination of stratified random sampling and purposive sampling, ensuring representation across various geographic regions and demographics within Japan.

The survey would be stratified by latitude and elevation bands, recognizing that blooming times differ significantly across these dimensions. This approach would ensure our sample includes participants from urban, rural, coastal, and inland areas, providing a richer context for the impact of climatic changes on cherry blossom timing. Sampling would also consider socio-demographic factors such as age and occupation, as these factors may influence awareness and historical knowledge of bloom events. Data collection methods would include both online surveys and in-person interviews to maximize reach and accessibility.

Moreover, measurement tools within the survey would follow well-established practices. Questions would address individuals' observations on bloom timing, any perceived shifts over years, and associated environmental conditions. The survey would utilize a Likert scale to capture perceptions of climatic changes, and open-ended questions to gather qualitative intuitions into local conditions and environmental anecdotes that may not be captured by weather stations.

6.2 Idealized Survey

To supplement the models presented in this paper, we propose an idealized survey that could provide more in-depth intuitions into the local experiences of climate effects on cherry blossom blooming. If we had an extended budget of $100,000, this survey would be expanded to cover additional aspects not readily accessible through observational or temperature datasets alone.

The idealized survey would be conducted over a full blooming season, collecting data from diverse regions across Japan. The goal would be to capture spatial and temporal variability in bloom timing and link it to local microclimatic conditions, participant observations, and other environmental factors such as soil moisture and local flora interactions. We would employ a mixed-methods approach, combining both quantitative (closed questions) and qualitative (interviews, focus groups) data.

Sampling would be implemented through a mixture of probability and non-probability approaches to achieve both representativeness and depth. Stratified random sampling would be used to ensure geographic diversity, while snowball sampling would help gather data from specific cultural or expert groups (e.g., horticulturalists, elders who possess traditional knowledge about historical blooming patterns). Each participant would provide data on observed bloom dates, perceived changes in bloom patterns over time, and any unusual climate events that they believe might have influenced sakura blooming.

Survey questions would be developed with guidance from both climate scientists and cultural experts to ensure that they adequately capture the climatic, geographic, and cultural contexts of sakura blooming. Likert scales would be used to gauge perceptions of change in bloom patterns and the impact of factors such as temperature and rainfall. In addition, open-ended questions would allow respondents to describe any unusual events or trends they have noticed over time.

The survey results would be analyzed using statistical techniques, including logistic regression to identify factors associated with changes in bloom timing, and thematic analysis for qualitative responses. The intuitions gained would be used to refine our predictive models by incorporating more subtle, location-specific environmental factors that are not currently represented in our dataset.

The implementation of this idealized survey could significantly enhance the robustness of our models, offering deeper intuitions into the role of microclimates and local environmental influences on cherry blossom phenology. Moreover, it would provide an opportunity to include public perception in our analysis, offering a human dimension to the understanding of climatic impacts on a culturally significant event.

6.3 Survey Design

To provide an understanding of local environmental factors impacting cherry blossom blooming, we propose a survey targeted at residents across Japan. This survey aims to gather data regarding personal observations of bloom dates, temperature changes, and any noted environmental anomalies. Below is a suggested questionnaire to gather intuitions from participants.

1. **What is your age group?**
   - Under 18
   - 18-34
   - 35-49
   - 50-64
   - 65 and above
2. **What is your gender?**
   - Male
   - Female
   - Prefer not to say
3. **In which region of Japan do you currently reside? (Select one)**
   - Hokkaido
   - Tohoku
   - Kanto
   - Other: _________ (PleaseTypeHere)
4. **How long have you been observing cherry blossom bloom in your area?**
   - Less than 1 year
   - 1-5 years
   - More than 5 years
5. **Have you noticed any changes in the blooming dates over the past years?**
   - Yes, blooming is earlier
   - Yes, blooming is later
   - No significant change
   - Not sure
6. **Which environmental factors do you think have the most impact on bloom timing? (Select all that apply)**

- Temperature changes
- Rainfall levels
- Urban development
- Other (please specify)

7. **On average, when do cherry blossoms begin to bloom in your area?**
   - Before March
   - March (early)
   - March (mid)
   - March (late)
   - April (early)
   - April (mid)
   - April (late)
   - After April

8. **How would you describe the temperature during the blooming period in the past few years?**
   - Much warmer
   - Slightly warmer
   - About the same
   - Slightly cooler
   - Much cooler

9. **Have there been any unusual climate events (e.g., heavy snowfall, prolonged warm spells) in recent years that might have affected blooming?**
   - Yes (please specify)
   - No
   - Not sure

10. **Do you think climate change is influencing the blooming pattern of cherry blossoms in your region?**
    - Strongly agree
    - Agree
    - Neutral
    - Disagree
    - Strongly disagree

11. **Are there other plants in your region that show similar blooming patterns to cherry blossoms?**
    - Yes (please specify)
    - No
    - Not sure

6.4 Incorporating Survey Findings

Integrating the survey findings into the discussion of cherry blossom bloom dynamics enhances the understanding of how local perceptions and environmental conditions influence cherry blossom blooming. The survey would capture personal observations of bloom timing, temperature changes, and unusual climate events, helping to refine predictions and account for local variations not represented by environmental data alone. By combining both quantitative and qualitative data, the survey would provide a clearer view of how climatic changes, such as temperature shifts and extreme weather events, are affecting sakura blooms in different regions. These findings would support more accurate forecasts and improve conservation and event planning efforts.

6.5 Model Details

This section provides full summary table for model 1 and model 2.

Table 7: Model 1 Summary Tables

(a)

Residuals for Model 1		
	Statistic	Value
0%	Min	-0.2520814
25%	1Q	-0.1465140
50%	Median	-0.0409466
75%	3Q	0.0460629
100%	Max	1.3070912

(b)

Coefficients for Model 1				
	Estimate	Std. Error	t Value	P-Value
(Intercept)	126.202188	0.0343959	3669.1015	< 2.2e-16
temp	-3.509279	0.0053551	-655.3142	< 2.2e-16

(c)

Model Summary for Model 1	
Metric	Value
Residual Standard Error	0.2337 on 572 degrees of freedom
Multiple R-squared	0.9987
Adjusted R-squared	0.9987
F-statistic	429436.7017 on 1 and 572 DF
P-value	< 2.2e-16

Table 8: Model 2 Summary Tables

(a)

Residuals for Model 2		
	Statistic	Value
0%	Min	-4.2185455
25%	1Q	-0.6993186
50%	Median	-0.0729067
75%	3Q	0.6641482
100%	Max	4.1113800

(b)

Coefficients for Model 2				
	Estimate	Std. Error	t Value	P-Value
(Intercept)	-9.0004393	0.2894642	-31.09344	< 2.2e-16
day	-0.0492231	0.0032911	-14.95648	< 2.2e-16
latitude	0.1890333	0.0120889	15.63688	< 2.2e-16
month2	-0.6179283	0.2198314	-2.81092	0.004963
month3	2.8693830	0.1143644	25.08982	< 2.2e-16
month4	6.9661077	0.1649841	42.22291	< 2.2e-16
month5	10.0473632	0.2374526	42.31312	< 2.2e-16

(c)

Model Summary for Model 2	
Metric	Value
Residual Standard Error	1.1156 on 4256 degrees of freedom
Multiple R-squared	0.9255
Adjusted R-squared	0.9254
F-statistic	8813.9897 on 6 and 4256 DF
P-value	< 2.2e-16

References

Arnold, Jeffrey B. et al. 2023. *Ggthemes: Extra Themes, Scales and Geoms for Ggplot2*. https://CRAN.R-project.org/package=ggthemes.

BBC Travel. 2024. "Climate Change Thwarts Cherry Blossom Travel." https://www.bbc.com/travel/article/20240223-climate-change-thwarts-cherry-blossom-travel.

Müller, Kirill et al. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Ocko, Ilissa. 2024. "Cherry Blossoms: A Microcosm of the Global Climate Crisis." https://blogs.edf.org/climate411/2024/03/21/cherry-blossoms-a-microcosm-of-the-global-climate-crisis/.

P., Tina et al. 2023. *caTools: Tools: Moving Average, ROC, Etc*. https://CRAN.R-project.org/package=caTools.

Pebesma, Edzer et al. 2023. *Sf: Simple Features for r*. https://CRAN.R-project.org/package=sf.

Pedersen, Thomas Lin et al. 2023. *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/.

Richardson, Neal et al. 2023. *Arrow: Interface to Apache Arrow*. https://CRAN.R-project.org/package=arrow.

Rohan Alexander. 2021. "Starter Folder for GitHub Papers." https://github.com/RohanAlexander/starter_folder.

South, Andy et al. 2023a. *Rnaturalearth: World Map Data from Natural Earth*. https://CRAN.R-project.org/package=rnaturalearth.

——— et al. 2023b. *Rnaturalearthdata: Large-Scale World Map Data from Natural Earth*. https://CRAN.R-project.org/package=rnaturalearthdata.

tacookson. n.d. "Sakura Flowering Dataset." https://github.com/tacookson/data/tree/master/sakura-flowering.

Wickham, Hadley et al. 2023a. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

——— et al. 2023b. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

——— et al. 2023c. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

——— et al. 2023d. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

——— et al. 2023e. *Testthat: Testing for r*. https://CRAN.R-project.org/package=testthat.

——— et al. 2023f. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

——— et al. 2023g. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wikipedia contributors. 2024. "Cherry Blossom — Wikipedia, the Free Encyclopedia." https:

//en.wikipedia.org/wiki/Cherry_blossom.

Xie, Yihui et al. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.

Zhu, Hao et al. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.