STAT 207 Take-Home Quiz 1

Department of Statistics, UC Santa Cruz

Released: April 23
Due: April 27

Turning In Your Exam: Submit your complete written report (as a single PDF) and any supplementary code on Canvas no later than the deadline above. **Late submissions will have 10% points deducted for each late day.**

---

1 General Instructions

You are to analyze the provided dataset to address the objectives above. You should rigorously justify your modeling choices and use appropriate methods for quantifying uncertainty.

Report format. Your final analysis should be presented as a written report of **no more than 8 pages** using the template provided on Canvas, including all relevant tables and figures. You are required to submit your codes as the Appendix of the report. You may also include additional convergence diagnostics or sensitivity analyses in the report, but any materials other than the source codes need to be explained and clearly labeled.

Required sections. Your report must include, at minimum:

- **Abstract**: a brief summary of your findings.
- **Introduction**: motivation and background for the problem that are relevant to your model choice later.
- **Methods**: description and justification of all statistical methods used.
- **Results**: presentation and interpretation of results, organized around the analytical tasks
- **Discussion**: synthesis of findings and limitations

Collaboration policy. You may consult any written or electronic references. You may **not** communicate by any means with any other person about this exam or your analysis.

AI use policy and disclosure. The use of AI-assisted tools (e.g., large language models such as ChatGPT, Claude, or Copilot) is **permitted** for this exam. You will **not** be penalized for using such tools. However, your grade is based solely on the *quality* of the submitted work: the clarity and rigor of your statistical reasoning, the correctness and depth of your analysis, and the coherence of your written report, regardless of how it was produced.

You **must** include a brief **AI Use Disclosure** section at the end of your report (after the Discussion, before the References) structured as follows:

1. **Tools used.** List each AI tool you used (name and version if known), or state "None."
2. **Role in analysis.** Describe specifically how each tool was used in the *statistical analysis* (e.g., "used to help debug Stan code," "used to suggest prior distributions," "not used for analysis").
3. **Role in writing.** Describe specifically how each tool was used in *writing or editing* the report (e.g., "used to proofread for grammar," "used to draft the Introduction," "not used for writing").

4. **Critical evaluation.** Briefly note any cases where you found AI-generated suggestions to be incorrect, misleading, or unhelpful, and how you resolved them. If no such cases arose, state that explicitly.

This disclosure is required regardless of whether you used AI tools. Omitting it will result in point deduction.

2 Background

The annual flowering of cherry trees (*Prunus* spp.) is one of the most iconic harbingers of spring across the temperate world. In Japan, the blossoming of *Prunus jamasakura* has been celebrated as *hanami* (flower viewing) for over a thousand years, and historical records of peak bloom dates in Kyoto, the longest continuous phenological record on Earth, extend back to 812 CE. Comparable long-term records exist in Europe and North America, providing a rare opportunity to study how a sensitive ecological indicator has responded to climate over centuries and decades.

Peak bloom timing is determined primarily by the accumulation of warmth in late winter and early spring. As a result, it functions as a natural thermometer: earlier blooms signal warmer springs. Studies using both the Kyoto record and instrumental climate data have documented a striking shift toward earlier bloom timing in recent decades compared to the mid-20th century. However, the *magnitude* of this shift is not uniform: local geography, land use, and the degree of continental versus maritime influence all modulate the sensitivity of bloom timing to temperature.

In this exam you will analyze a publicly available dataset compiled from six phenological monitoring sites spanning a range of latitudes and climatic settings:

- **Kyoto, Japan** (35.0°N) — records from 812 CE to 2021; compiled by Aono & Kazui [1].
- **Liestal, Switzerland** (47.5°N) — records from 1894 to 2021; part of the Swiss Phenological Network.
- **Washington DC, USA** (38.9°N) — records from 1921 to 2021; compiled from National Park Service data.
- **Vancouver, Canada** (49.2°N) — records from 2022 to 2025 (4 years); recently added to the competition dataset.
- **Seoul, South Korea** (37.6°N) — records from 1980 to 2020; compiled from Korea Meteorological Administration station data.
- **Yeosu, South Korea** (34.7°N) — records from 1980 to 2020; compiled from Korea Meteorological Administration station data.

The primary outcome of interest is **peak bloom day of year (DOY)**: the calendar day on which cherry trees at each site reached peak flowering, expressed as an integer from 1 (January 1st) to 365 (December 31st). A smaller DOY indicates an earlier, typically warmer spring.

3 Data Description

The dataset `cherry_bloom.csv` contains annual peak bloom observations for six cherry blossom monitoring sites. A reproducible `R` script to download and prepare this file is provided alongside the exam (`download_data.R`). The script retrieves data directly from the public repository of the George Mason University Cherry Blossom Peak Bloom Prediction Competition (https://github.com/GMU-CherryBlossomCompetition/peak-bloom-prediction), which compiled the per-site records from their respective primary sources [1, 2].

Variable	Description
`location`	Site name: Kyoto, Liestal, Washington DC, Vancouver, Seoul, or Yeosu.
`latitude`	Approximate latitude of the site (degrees North).
`longitude`	Approximate longitude of the site (degrees).
`year`	Calendar year of the observation.
`period`	Climatological period of the observation: `"Early"` for years 1950–1987, or `"Recent"` for years 1988–2025. This variable is provided for convenience and defines the two comparison groups used in this exam.
`bloom_date`	Calendar date of peak bloom (YYYY-MM-DD format).
`bloom_doy`	Day of year of peak bloom (integer, 1–365). **This is the primary outcome variable.** Smaller values indicate earlier bloom.

4 Analytical Tasks

The overarching goal of this analysis is to characterize whether cherry blossom peak bloom timing has shifted between the Early (1950–1987) and Recent (1988–2025) climatological periods, and to assess whether the magnitude of this shift varies across the six monitoring sites. Your report should address the following specific tasks.

(a) Conduct a thorough exploratory data analysis of `cherry_bloom.csv`. At a minimum, your EDA should address: the distribution of `bloom_doy` within each site; differences in bloom timing across sites; a comparison of `bloom_doy` distributions between the Early and Recent periods, both overall and within each site; and any apparent patterns in variability or data coverage (e.g., which sites have limited observations in one or both periods).

(b) Based on your EDA, propose a Bayesian hierarchical model to estimate the mean peak bloom DOY for each site within each period, and to compare those means between periods. You will fit one model per period (Early and Recent). Write out your model fully in mathematical notation, specifying:

- the likelihood (data-level model) for `bloom_doy`;
- the random-effects structure (which parameters vary by site and how they are related across sites);
- the prior distributions for all parameters;
- any hyperpriors on variance components.

Clearly justify each choice. For prior distributions, discuss whether your choices are informative or weakly informative, and explain how you would check for prior sensitivity.

(c) Fit your proposed hierarchical model (separately for each period) using software of your choice. In addition, fit the **no-pooling** special case of your model: the same model structure applied independently to each site within each period, with no sharing of information across sites. Note that Vancouver has no observations in the Early period; for the no-pooling fit you should discuss whether a meaningful site-specific estimate can be obtained for Vancouver in the Recent period and handle it accordingly. For all fits, assess convergence using appropriate MCMC diagnostics and report whether satisfactory convergence was achieved.

(d) Provide a brief interpretation of the posterior distributions for the site-specific mean bloom DOY in each period from your hierarchical model. Do the posterior means shift in the expected direction (earlier bloom in the Recent period)? Is this shift consistent in sign and magnitude across sites? Which sites show the most or least evidence of a shift?

(e) Produce a figure comparing the site-specific mean bloom DOY estimates under the no-pooling and hierarchical (partial-pooling) models for at least one of the two periods, along with 90% posterior credible intervals. Which site(s) show the most pronounced shrinkage toward the population mean, and why? What does this imply about those sites scientifically?

(f) Discuss the bias-variance tradeoff illustrated by the comparison between the no-pooling and partial-pooling fits. Under what conditions would you expect partial pooling to offer the largest benefit? Relate your answer to specific features of this dataset (e.g., sites with limited observations in one of the two periods).

(g) Define the quantity "average shift in peak bloom timing between the Early and Recent periods, pooled across sites" as a precise estimand expressed in terms of your model parameters. Then estimate this quantity from the posteriors of your hierarchical models: report the posterior mean and a 95% credible interval, and interpret the result in plain scientific language. Is there strong evidence that bloom timing has shifted? By how many days on average?

(h) Based on the site-specific shifts in mean bloom timing between the two periods from your hierarchical model, assess whether there is evidence that higher-latitude sites are experiencing larger or smaller shifts in bloom timing compared to lower-latitude sites. You are not required to specify or fit a new model for this question; instead, use the posterior distributions already obtained and any appropriate summaries or visualizations. Discuss what you can and cannot conclude given the number of sites and their geographic diversity.

References

[1] Aono, Y., & Kazui, K. (2008). Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century. *International Journal of Climatology*, 28(7), 905–914.

[2] Primack, R. B., Higuchi, H., & Miller-Rushing, A. J. (2009). The impact of climate change on cherry trees and other species in Japan. *Biological Conservation*, 142(9), 1943–1949.