Cherry Blossom Peak Bloom Prediction 2026

Methodology, Data Analysis & Final Predictions · Team 5103

Team 5103 — George Mason University

2026-02-01

Agenda

1. Competition Context
2. Data Sources & Enrichment
3. Exploratory Data Analysis
4. Feature Engineering
5. Model A — Local Trend
6. Model B — Pooled Nonlinear Learner
7. Backtesting & Ensemble Blending
8. Prediction Intervals
9. Cross-Language Robustness
10. Final 2026 Predictions

Competition Context

Aspect	Detail
Organizer	George Mason University, Dept. of Statistics
Task	Predict peak bloom DOY for **5 sites** in 2026
Point scoring	Sum of absolute errors across all 5 sites
Interval scoring	Coverage count (tiebreak: sum of squared widths)
Deadline	February 28, 2026 (AoE)

Sites

Kyoto (Japan) · Washington D.C. (USA) · Liestal-Weideli (Switzerland) · Vancouver BC (Canada) · New York City (USA)

Site Characteristics

Site	Latitude	Longitude	Alt (m)	Record span	Species
Kyoto	35.01	135.68	44.0	812 – 2025	P. jamasakura
Washington DC	38.89	-77.04	0.0	1921 – 2025	P. × yedoensis
Liestal	47.48	7.73	350.0	1895 – 2025	P. avium
Vancouver	49.22	-123.16	24.0	2022 – 2025	P. × yedoensis ‘Akebono’
New York City	40.73	-74.00	8.5	2019 – 2025	P. × yedoensis

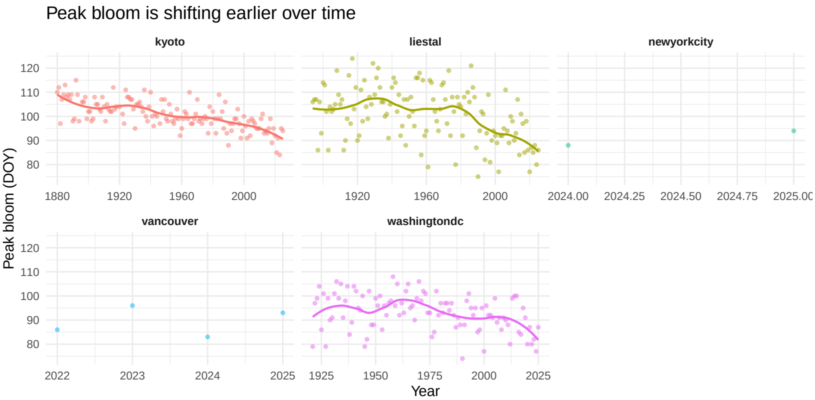
Data Sources Overview

Category	Files	Rows
Competition core	kyoto, washingtondc, liestal, vancouver, nyc	1080
Auxiliary (Japan, MeteoSwiss, S. Korea)	japan, meteoswiss, south_korea	14209
USA-NPN (NYC enrichment)	status-intensity + individual phenometrics	~4 extra NYC bloom-years

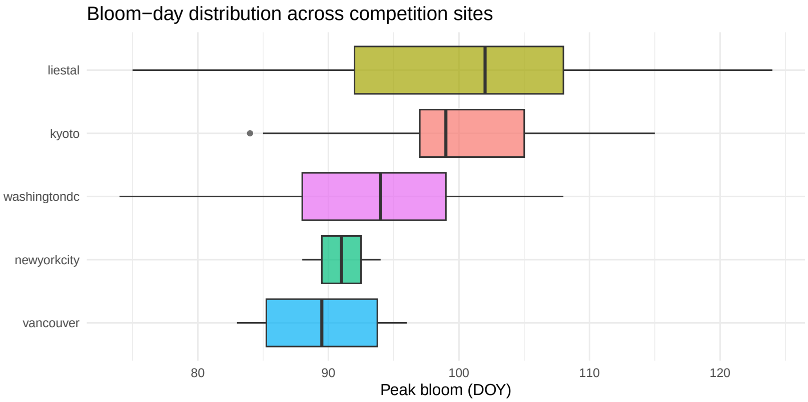
NYC Data Enrichment (USA-NPN)

- **Site 32789** (Washington Square Park), Species 228, Phenophase 501
- Status-intensity: first `Phenophase_Status == 1` per year
- Phenometrics: `min(First_Yes_DOY)` per year
- Merge rule: status takes priority; phenometrics fills gaps
- **Result:** 5 extra NYC bloom years added

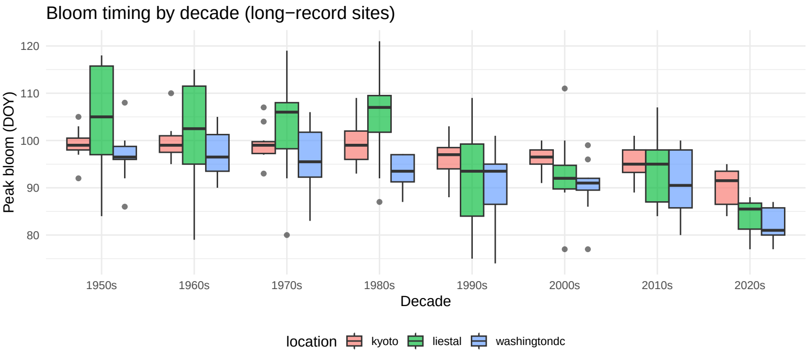
EDA: Long-Term Trend



EDA: Bloom Distribution by Site



EDA: Recent Decade Acceleration



Over the last two decades, median bloom at all three long-record sites is the **earliest on record**.

Feature Engineering

Feature	Formula / Description	Ecological rationale
`year_c`	`year - 1950`	Centers the time axis
`year_c`^{2}	`(year - 1950)`^{2}	Captures trend acceleration
`lat`, `long`	Raw coordinates	Spatial climate gradients
`alt_log1p`	`log(1 + max(alt, 0))`	Diminishing altitude effect
`site_obs`	Count of records per site	Data-reliability proxy
`source`	competition / auxiliary / npn	Data-provenance indicator

Model A: Local Recency-Weighted Trend

Approach:

- Per-site quadratic regression: `bloom_doy ~ year + year²`
- Exponential decay weights: $w_i = e^{(i-n)/6}$ (half-life 6 yr)
- Recent years dominate while long history provides curvature

Fallback rules:

- 4 obs $\rightarrow$ weighted quadratic
- 2–3 obs $\rightarrow$ unweighted linear
- 1 obs $\rightarrow$ site mean

Strengths:

- Captures site-specific momentum
- Adapts to recent bloom acceleration

Weaknesses:

- Cannot leverage cross-site info
- Poor for sparse sites (Vancouver, NYC)

Model B: Pooled Nonlinear Learner

R pipeline — Generalized Additive Model (GAM):

$$\text{bloom\_doy} \sim s(\text{year}, k{=}25) + s(\text{lat}, \text{long}, k{=}40) + s(\text{alt}, k{=}8) + s(\text{site\_obs}$$

- Estimation method: REML · Trained on **all ~14 K+ records** (competition + auxiliary + NPN)

Python pipeline — Gradient Boosting Regressor (GBR):

- Huber loss, 700 estimators, learning rate = 0.02, max depth = 3
- Same feature set; `OneHotEncoder` for source, `MedianImputer` for numerics

Key advantage: Learns transferable spatial-temporal structure — sparse sites borrow strength from thousands of auxiliary records.

Rolling-Origin Backtesting

Procedure: For each year y from 1900 to 2025:

1. Train on all data with `year < y`
2. Predict competition sites observed at year y
3. Record absolute errors for Model A & Model B

→ Ensures **no future leakage**. → Provides honest MAE estimates and residuals for interval calibration.

Ensemble Blending — Quantitative Results

Inverse-MAE weighting:

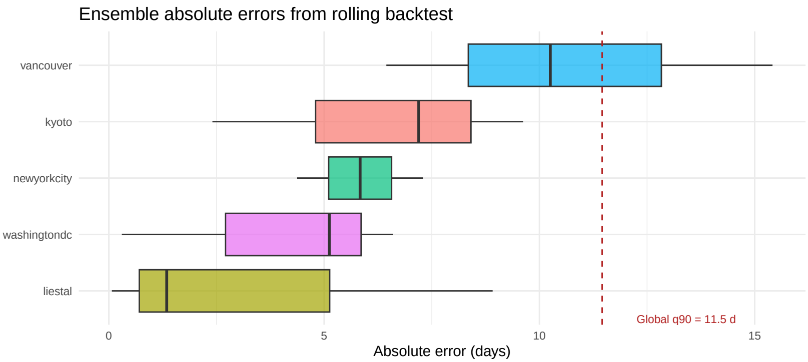
$$w_A = \frac{1/\mathsf{MAE}_A}{1/\mathsf{MAE}_A + 1/\mathsf{MAE}_B}, \quad w_B = 1 - w_A$$

Model	MAE (days)	Weight
Local (Model A)	7.01	50.7%
GAM (Model B)	7.21	49.3%
Ensemble	6.10	—

$$\hat{y} = 0.507 \times \hat{y}_{\mathsf{local}} + 0.493 \times \hat{y}_{\mathsf{GAM}}$$

The ensemble outperforms both individual models on held-out years.

Backtest Residual Analysis



Prediction Intervals

Split-conformal calibration:

- Half-width = 90th percentile of backtest |residuals| per location
- Interval: $[\hat{y} - q_{90},\ \hat{y} + q_{90}]$
- Fallback to global q_{90} for unseen sites
- Clipped to valid range [1, 366]

Location	Half-width	Widt
kyoto	9.6	
liestal	8.9	
newyorkcity	7.3	
vancouver	15.4	
washingtondc	6.6	

Design goal: 90% empirical coverage while minimizing $\sum(\text{width}^2)$ (competition tiebreaker).

Cross-Language Robustness Check

Two **fully independent** pipelines:

Pipeline	Model B	Output
R (primary)	GAM (REML)	`cherry-predictions.csv`
Python	GBR (Huber, 700 trees)	`cherry-predictions-python.`

Location	R pred	Python pred	$\lvert$Gap$\rvert$
kyoto	88	92	4
liestal	88	87	1
newyorkcity	93	92	1
vancouver	92	93	1
washingtondc	82	84	2

Mean point gap = **1.8 days** (4 $\rightarrow$ blended submission used).

Final 2026 Predictions — R Pipeline

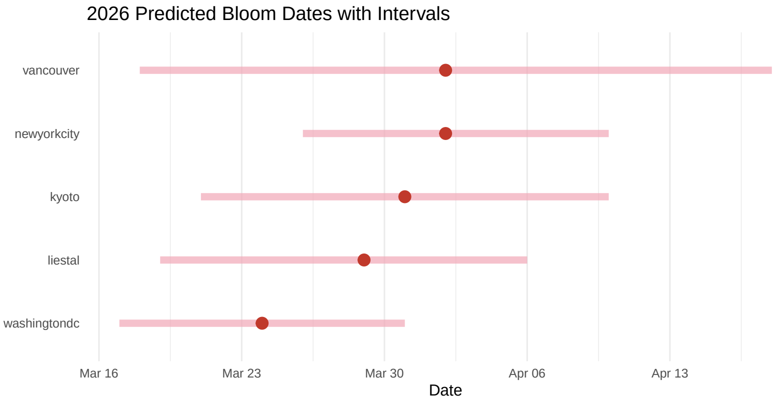
Location	Prediction (DOY)	Lower	Upper	Predicted date	Interval
kyoto	88	78	98	2026-03-29	Mar 19 – Apr 08
liestal	88	78	97	2026-03-29	Mar 19 – Apr 07
newyorkcity	93	85	101	2026-04-03	Mar 26 – Apr 11
vancouver	92	76	108	2026-04-02	Mar 17 – Apr 18
washingtondc	82	75	89	2026-03-23	Mar 16 – Mar 30

Final 2026 Predictions — Blended Submission

Location	DOY	Lower	Upper	Predicted date	Interval	Width
kyoto	90	80	100	2026-03-31	Mar 21 – Apr 10	20
liestal	88	78	96	2026-03-29	Mar 19 – Apr 06	18
newyorkcity	92	85	100	2026-04-02	Mar 26 – Apr 10	15
vancouver	92	77	108	2026-04-02	Mar 18 – Apr 18	31
washingtondc	83	76	90	2026-03-24	Mar 17 – Mar 31	14

Sum of squared interval widths (tiebreaker): 2106

Predictions Visualized



Interpretation

Why the ensemble works:

- Local trend captures site-specific acceleration
- GAM captures cross-site spatial structure
- Inverse-MAE blending is purely data-driven
- Backtest MAE 6.1 days

Climate signal:

- Bloom DOY is decreasing at all 5 sites
- Last two decades are the earliest on record
- Sparse sites (Vancouver, NYC) borrow strength from 14K+ auxiliary records

Interval design: Per-site conformal widths adapt to each location's predictability (wider for Vancouver 31 d, narrower for NYC 15 d).

Limitations & Future Work

Limitation	Potential improvement
No direct temperature covariates	NOAA API winter/spring degree-day features
Short records at Vancouver (4 yr) & NYC (5 yr)	Transfer learning / Bayesian priors
Residual weather shocks unpredictable	Ensemble with current-season temperature forecasts
Point uncertainty only from conformal residuals	Full Bayesian credible intervals or quantile regression
No phenological process model	Chill-unit accumulation (e.g., Utah model)

Reproducibility & Repository Map

```
peak-bloom-prediction_5103-Team/
   solution.qmd                 ← Primary R pipeline (abstrac
   Solution.ipynb               ← Independent Python pipeline
   cherry-predictions.csv        ← R output
   cherry-predictions-python.csv ← Python output
   cherry-predictions-final.csv  ← Blended final submission
   abstract.md                  ← Competition abstract (335 w
   slides.qmd                   ← This presentation source
   data/                        ← All datasets + README
   demo_analysis.qmd            ← Competition-provided demo
```

Reproduce everything:

```
quarto render solution.qmd        # R analysis + predictior
jupyter nbconvert --execute Solution.ipynb --inplace  # Pyt
quarto render slides.qmd          # This deck
```

Thank You

Peak Bloom Prediction 2026

Team 5103 · George Mason University

> Tip
>
> All code, data, and outputs are publicly available and fully reproducible.