Cherry Blossom Peak Bloom Prediction

Methodology and 2026 Results (LM vs LASSO)

Callie Ann Pokorski
Daul-MS Operations Research and Statistical Science
George Mason University

February 21, 2026

Objective and Scope

- Predict 2026 peak bloom day-of-year for Kyoto, Washington DC, Liestal, Vancouver, and New York City.
- Compare two models built on the same pre-bloom climate features:
  - Linear regression (LM)
  - LASSO (CV-selected $\lambda_{\min}$)
- Report uncertainty as $\pm$ days for both models.

Data window

All workflows are restricted to years ≥ 1973.

Data Pipeline and Feature Engineering

- Merge bloom-history and climate-station data by mapped location/station pairs.
- Keep core climate variables: T_{max}, T_{min}, precipitation.
- Fill short internal temperature gaps only (< 3 consecutive days).
- Build yearly pre-bloom aggregates up to observed bloom date:
  - mean adjusted max temperature
  - mean adjusted min temperature
  - total precipitation

Altitude Adjustment Methodology

Rationale

Stations and bloom sites differ in elevation, introducing temperature bias if uncorrected.

$$\Delta T = 6.5\,^{\circ}\mathrm{C/km} \cdot \frac{h_{\text{station}} - h_{\text{bloom}}}{1000}$$

$$T_{\text{max}}^{\text{adj}} = T_{\text{max}} + \Delta T, \qquad T_{\text{min}}^{\text{adj}} = T_{\text{min}} + \Delta T$$

- If station is higher than bloom site, adjusted temperatures increase.
- If station is lower, adjusted temperatures decrease.

Model Design and Split Strategy

Training / Holdout

- Train locations: Kyoto, Washington DC, Liestal
- Holdout locations: Vancouver, New York City
- Holdout split by time:
  - earlier half → validation
  - later half → test

Models

LM and LASSO use:

- mean_tmax_adj_prebloom
- total_prcp_prebloom

Uncertainty Methodology

Linear Model

90% confidence bounds from regression prediction intervals.

LASSO

Bootstrap-based 90% intervals:

1. Resample training rows (with replacement)
2. Refit LASSO at fixed λ_{min}
3. Predict 2026 DOY per location
4. Use empirical 5th/95th quantiles

Validation/Test Model Metrics

Model	Split	MAE (days)	RMSE (days)
Linear	Validation ($n = 24$)	6.62	8.06
Linear	Test ($n = 15$)	7.59	9.45
LASSO	Validation ($n = 24$)	6.62	8.05
LASSO	Test ($n = 15$)	7.59	9.45

Interpretation

LM and LASSO perform nearly identically under the reduced predictor set.

Parameter Estimates (90% CI)

Model	Term	Estimate	90% CI
LM	Intercept	106.59	[100.40, 112.78]
LM	mean_tmax_adj_prebloom	-1.87	[-2.48, -1.26]
LM	total_prcp_prebloom	0.031	[0.017, 0.045]
LASSO	Intercept	106.54	[100.45, 113.22]
LASSO	mean_tmax_adj_prebloom	-1.86	[-2.53, -1.23]
LASSO	total_prcp_prebloom	0.031	[0.015, 0.047]

Source

Exported from `data/model_outputs/model_parameter_estimates_90ci_comparison.csv`.

2026 Results: LM vs LASSO

For the model-date comparison, we report the predicted bloom date (in YYYY-MM-DD) and the $\pm$ uncertainty width (in days) (90% confidence) for each location and model. The difference is calculated as (LASSO date - LM date) in terms of day-of-year (DOY).

Location	LM Date	LM $\pm$	LASSO Date	LASSO $\pm$
Kyoto	2026-03-29	3.39	2026-03-29	3.00
Washington DC	2026-04-02	1.55	2026-04-02	1.57
Liestal	2026-04-04	1.54	2026-04-04	1.58
Vancouver	2026-04-10	1.85	2026-04-10	1.99
New York City	2026-04-09	1.58	2026-04-09	1.78

Difference (LASSO - LM, DOY)

Kyoto: +0.05, DC: +0.02, Liestal: +0.01, Vancouver: -0.03, NYC: -0.02

Takeaways

- Reduced-feature LM and LASSO produce nearly identical validation/test performance.
- Parameter estimates are stable across LM and LASSO with consistent sign and magnitude.
- 2026 predictions are highly aligned across models (all DOY differences within about 0.05 days).
- Final submission can choose a single model or ensemble based on validation/test metrics.

Thank You

Any Questions?