Methodology: Cherry Blossom Bloom-Date Prediction
Linear Model and LASSO

Peak Bloom Prediction Workflow

February 21, 2026

1 Objective

This document summarizes the modeling methodology used to predict annual cherry blossom peak bloom timing (day-of-year) for each location, with a focus on the standard **linear model (LM)** and a secondary **LASSO** model.

2 Data Sources and Initialization

The workflow loads and harmonizes:

- Bloom history data by location and year (peak bloom date and day-of-year).
- Daily climate station data (temperature, precipitation, station metadata).

Key preprocessing decisions include:

- Restrict records to years ≥ 1973.
- Map NOAA stations to bloom locations.
- Keep selected climate variables (temperature and precipitation inputs for the standard model pipeline).
- Fill only short internal temperature gaps (less than 3 consecutive days) by linear interpolation.

3 Feature Engineering for Standard Model

From daily climate records and bloom events, yearly pre-bloom predictors are constructed.

3.1 Altitude/Location Adjustment

For each location, station-to-bloom-site altitude difference is computed using:

- Station altitude from climate station elevation.
- Bloom-site altitude from bloom history metadata.

A lapse-rate temperature correction is applied:

$$\Delta T = \gamma \cdot \frac{h_{\text{station}} - h_{\text{bloom}}}{1000}, \quad \gamma = 6.5\,^{\circ}\text{C/km}. \tag{1}$$

Adjusted daily temperatures are:

$$T_{\max}^{\text{adj}} = T_{\max} + \Delta T, \qquad T_{\min}^{\text{adj}} = T_{\min} + \Delta T. \tag{2}$$

3.2 Pre-bloom Annual Aggregates

For each location-year, using all days from Jan 1 through bloom date:

- mean_tmax_adj_prebloom
- mean_tmin_adj_prebloom
- total_prcp_prebloom
- Bloom-site altitude feature bloom_alt_m

The target is bloom day-of-year, $y = \text{bloom\_doy}$.

4 Train / Validation / Test Design

The standard model uses location-based domain split:

- **Training locations**: Kyoto, Washington DC, Liestal.
- **Holdout locations**: all remaining locations.

For holdout locations, years are ordered chronologically and split into:

- first half → validation,
- second half → test.

This tests geographic transferability while preserving temporal order within each holdout location.

5 Primary Model: Linear Regression (LM)

The baseline model is:

$$\text{bloom\_doy} = \beta_0 + \beta_1\,\text{mean\_tmin\_adj\_prebloom} + \beta_2\,\text{mean\_tmax\_adj\_prebloom} + \beta_3\,\text{total\_prcp\_prebloom} + \beta_4\,\text{bloom\_alt} \tag{3}$$

Evaluation metrics on validation/test:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}. \tag{4}$$

6 Secondary Model: LASSO

A LASSO regression is fit using the same predictors and target, with λ selected by cross-validation on training data only.

Optimization objective:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{5}$$

The selected model uses $\lambda_{\min}$ from CV and is scored on validation/test using the same MAE and RMSE metrics as LM.

7 2026 Forecasting and Uncertainty

7.1 Feature Construction for 2026

For each location, the most recent available feature row is used as the predictor baseline and assigned year = 2026.

7.2 LM 90% Confidence Bounds

For the linear model, 90% confidence intervals are produced directly from the regression prediction interval output for the mean response:

- predicted DOY,
- lower/upper 90% DOY bounds,
- converted calendar-date bounds,
- uncertainty width as $\pm$ days.

7.3 LASSO 90% Confidence Bounds

Because standard closed-form confidence bounds are not directly provided for penalized fits, uncertainty is estimated via bootstrap on training rows:

1. Resample training rows with replacement.
2. Refit LASSO at fixed $\lambda_{\min}$.
3. Predict 2026 DOY for each location.
4. Repeat (e.g., 500 replicates) and take empirical 5th/95th percentiles for a 90% interval.

The reported outputs include lower/upper bounds and $\pm$ day summaries.

8 Model Comparison Outputs

The process stores:

- Validation/test metrics by model (LM vs LASSO).
- Prediction-level residual tables by split and model.
- 2026 prediction tables for LM and LASSO.
- A side-by-side 2026 comparison including DOY difference (LASSO – LM).

9 Outputs for Final Report

This section can be used directly in the final competition report to summarize model performance and 2026 predictions.

9.1 Suggested Performance Summary Table

Report validation and test metrics for both models:

- MAE (days),
- RMSE (days),
- split sample size n.

9.2 2026 Prediction Table Template (LM vs LASSO)

Location	LM Date	LM 90% CI	LM ± Days	LASSO Date	LASSO 90% CI	LASSO ± Days
Kyoto						
Washington DC						
Liestal						
Vancouver						
New York City						

Table 1: Predicted 2026 peak bloom dates with 90% confidence bounds and uncertainty width.

9.3 How to Populate the Table

Use the saved model artifacts as follows:

- LM values from `predictions_2026`: `pred_bloom_date`, `conf_low_date_90`, `conf_high_date_90`, `conf_pm_days_90`.
- LASSO values from `predictions_2026_lasso`: `pred_bloom_date_lasso`, `conf_low_date_90_lasso`, `conf_high_date_90_lasso`, `conf_pm_days_90_lasso`.
- Optional model-difference column from `predictions_2026_comparison`: `doy_diff_lasso_minus_linear`.

10 Reproducibility Notes

- A fixed random seed is used for splitting and bootstrap reproducibility.
- All intermediate and final objects are saved as RData artifacts.
- Modeling scripts are modularized into data prep, feature construction, and model fitting stages.