Using Generalised additive mixed effects models (GAMMs) to model flowering phenology across Western North America

[1]Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, Illinois 60022, USA

Reed Clark Benkendorf[1]*

Abstract

Premise: Increasing the quantity of native seed available for habitat restoration is hindered by impediments to wild land seed collection. These challenges can be minimized by having wild land seed collection teams scout for populations near peak flowering - allowing more accurate assessments of populations census sizes and geographic extents. However, spatially explicit estimates of major phenological events across the geographic and ecological ranges of many taxa are rare.

Methods: Generalised additive mixed effects models were used to provide estimates of flowering phenology for 271 taxa. Herbaria data were manually scored, and records were then spatially clustered, before using the weibull distribution to generate pseudoabsences for flowering events before modelling.

Results: Models can be useful for providing a general order by which species can be scouted for by crews, and identifying work priorities throughout a season. Further they can provide ecologically relevant information at a variety of temporal and spatial resolutions.

Conclusions: Phenology can readily be modelled for a wide range of plant species using publically available data, and predicted into space for various applied applications. We provide an outline of the process here, but recommend using density estimate quantiles for generating pseudo-absences which will greatly improve the speed of the process.

Introduction

Changes in a species phenology, the timing of life history events, is one of the most common and pronounced responses to climate change (Parmesan and Yohe, 2003). Accordingly, considerable effort has been directed towards exploring the causal links between climate change and phenology (Tang et al., 2016). Given the importance of phenology to biodiversity, and the ready identification of causal agents, many meteorological

*Author for Correspondence: rbenkendorf@chicagobotanic.org

explanatory variables have been produced, as well as remotely sensed vegetation attributes linked to phenology, e.g. vegetation stand wide leaf out, and leaf senescence dates (Dronova and Taddeo, 2022). Generally studies which tend to treat species as vegetation complexes or communities have found that early season phenophases, any noticeable stage in a life cycle, have tended to advance, while late season phenophases have often become delayed (Parmesan and Yohe, 2003). However, for the individual plant species analyzed to date, responses to warming have been idiosyncratic preventing the generalization of results across, clades, most functional groups, resulting in a need for continued species specific modelling of phenology for many applications (CaraDonna et al., 2014; Augspurger and Zaya, 2020). These observational studies and manipulative experiments have generally been limited to a few dozen species in only one to a couple populations, or when using herbarium sheets from many populations across a spatial domain - than only a few species (Katal et al., 2022), although some exceptions exist s(Park et al., 2023).

Rather than the documentation of trends over time, e.g. flowering initiation advancing by 2.3 days per decade, associated with phenology research in a climate change context the capability to predict the timing of phenophases in an individual year based on realized weather are required for several applications, most notably agriculture and related disciplines. Species specific models have been generated for crop varieties for over half a century (Hodges, 1990), and increasingly incorporate data sources which seldom exist for wild species, e.g. genes, near real-time remote sensing data of pure stands of individuals, and large amounts of training data capable of training artificial intelligence (Nagai et al., 2020; Gao and Zhang, 2021; Deva et al., 2024). Additionally, these agricultural systems minimize several environmental factors e.g. the severity of drought, and are operating on lineages breed for consistent windows of phenophases; hence these recent innovations in crop science are difficult to transfer to wildland settings.

The number of papers attempting to predict flowering events in an individual year, across geographically large portions of a species range are fewer than either of the above use-cases (Hodgson et al., 2011). Attempting to model these events are complicated because not only are the responses of individual species idiosyncratic to climate change, the response of populations varies across species ranges, due not only to differing levels of climate change, but to existing broad environmental climate (Park et al., 2019, 2023) Generalized Additive Mixed Models (GAMM's) are often used to document a phenophase because a single model can have their splines fit to both initiation, peak, and cessation of an phase, using a single or multiple independent variables - a limitation of several other methods of estimation (Polansky and Robbins, 2013). The use of independent variable(s) alongside GAMM's ability to incorporate an error-correlation structure which accommodates spatial autocorrelation allows them to model the phenological parameters of a species across it's geographic and concomitant environmental range. However, the data sets which cover the wide range of species which

may be desired to model are few, with herbaria and citizen science initiatives being the two largest sources.

Disciplines which straddle ecology and agriculture, such as wildland seed harvest, require useful models of major phenological events (flowering, and fruit dispersal) to optimize the detection of populations, estimates of census sizes, and the eventual collection of native seed. Hererin we use GAMM's to model phenophases, inferred from herbarium specimens and using environmental predictors identified as casual cues of phenology, in space. Our necessity to more accurately understand the phenology of species arose from our goal of native seed collection for both native plant germplasm development, and *ex-situ* conservation. The identification of putative populations with enough individuals to warrant germplasm development is a time consuming process, because most plant species can only be identified when they have reproductive organs, and a populations ability to support these collections varies wildly with the years weather, pathogen load, and various stochastic processes. The collection of seeds, which is generally occurring for both many species and many populations each year, is challenged by both the need for crews to collect from other species and the natural dispersal of seeds - put simply - *timing is everything.*

Methods

Data Sources & Cleaning

Species records were derived from the Symbiota herbarium portal for all years from 1981-2021, these years reflected the climate means used as independent variables (Michonneau and Collins (2024)). All records were downloaded, and the records in the 2.5% Day of Year (DOY) quantile were manually reviewed. These early records were reviewed because novice collectors, especially with graminoids, may actually collect material without reproductive organs yet reaching anthesis ('in bud'). The later records were reviewed because collectors may have collections of individuals entirely post-anthesis - a situation very common with certain clades where species are commonly distinguished by morphological characteristics of their fruits (e.g. the Fabaceae or Leguminosae). In both scenarios analysts proceeded towards the mean of the distribution until they encountered 5 consecutive sheets with the desired phenophase.

Layer	Description	Source	Abbrev
1.	Mean Temperature of Warmest Quarter (BIO10)	Chelsa	bio10
2.	Precipitation of Driest Month (BIO14)	Chelsa	bio14
3.	Mean Monthly vapour pressure deficit (vpd)	Chelsa	vpd_mean

Layer	Description	Source	Abbrev
4.	Heat accumulation of Degree-days above 0C (gdd0)	Chelsa	gdd5
5.	First growing degree day above 0C (gdgfgd0)	Chelsa	gdgfgd5
6.	Number of Degree-days above 0C (ngd0)	Chelsa	ngd5
7.	Heat accumulation of Degree-days above 5C (gdd5)	Chelsa	gdd5
8.	First growing degree day above 5C (gdgfgd5)	Chelsa	gdgfgd5
9.	Number of Degree-days above 5C (ngd5)	Chelsa	ngd5
10.	Heat accumulation of Degree-days above 10C (gdd10)	Chelsa	gdd10
11.	First growing degree day above 10C (gdgfgd10)	Chelsa	gdgfgd10
12.	Number of Degree-days above 10C (ngd10)	Chelsa	ngd10
13.	Net Primary Productivity (npp)	Chelsa	npp
14.	Soil Bulk Density	SoilGrids	sbd
15.	Compound Topographic Index	terra	cti

Independent variables reflected climate, and landform and soil parameters which modulate soil moisture. The climate variables from CHELSA, were 1981-2010 annual means, for growing degree days (GDD) heat sums (at 0°C, 5°C, 10°C), first (gdgfgd) and last (gddlgd) growing degree days, vapor pressure deficit (vpd), Bio10 (mean daily mean air temperatures of the warmest quarter), and Bio14 (precipitation amount of the driest month) (Karger et al., 2017). Soil bulk density, which reflects the amount of air/water space in soil, was developed by SoilGrids (Hengl et al., 2017). Compound Topographic Index (cti), which describes the potential of an area to accumulate soil moisture via a combination of its landform position, slope, aspect, and it's up-slope catchment area, was downloaded from geomorpho90m and resampled from 90m to the 250m resolution of the previous data sets (Amatulli et al., 2020).

Generalised additive models require data on when a species was **not** flowering in order to constrain the splines for the onset and cessation of flowering. Pseudo-floral absences were created using known sites, and their observed phenology. All of the CHELSA climate variables were decomposed using principal components analysis, and the first axis (explaining 98.1% of the variation; 750m x 750m cells) was used as a feature space in a Ward-like hierarchical clustering algorithm which seeks to maximize homogeneity of both the

feature and constraint space - here geography (Chavent et al., 2021). A suitable number of clusters from the independent variable were automatically selected using kgs (White and Gramacy, 2022), these clusters were then reanalyzed in light of the constraint space using automatic selection of an alpha parameter which blends the feature and constraint space and re-clustered using hclustgeo (Chavent et al., 2021). Each cluster had weibull estimates of flowering initiation and cessation modelled, and any DOY within 28 days preceding onset or following cessation were drawn for each group (Belitz et al., 2020). These values were arranged by ascending DOY and joined to the members of the group via a decreasing warm to cool gradient along the 1^{st} PCA axis. To avoid having pseudoabsences which coincided too closely with flowering presences points in the assigned clusters which had a nearest geographic neighbor in a different cluster had their randomly generated pseudo-absences overwritten. Thin plate spline regression using the 1^{st} PCA axis as an independent variable, and interpolation was then used to repopulate the floral pseudo-absence dates with a value generally intermediate between the two sets of pseudoabsences estimates - but always less than the flowering date observed at the site (Nychka et al., 2021; Hijmans, 2024).

Modelling

All independent variables were extracted to the dependent variables, and if a value for an independent variable was missing - which was not uncommon for Soil Bulk Density, where the modelers excluded the fringes of several vernal playas - it was imputed as the mean of the variable for the species. All independent variables then underwent feature selection using the Recursive Feature Elimination (rfe) with 10 Cross-Validations (CV) folds, 5 replicates, and from 1-10 variables using caret (Kuhn and Max, 2008). The remaining variable(s) were used as covariates with DOY always included in the models. A GAMM was fit using presence/absence of flowering as a response, as well as GAMM's with error structure of gaussian, spherical, ratio, linear, and exponential variograms, with REML; if any spatial model failed to converge the aspatial model was selected as the top model. (Bartoń, 2023; Pinheiro et al., 2024). If multiple models converged than all models underwent model selection with the top model determined via change in AUC scores ($_{\Delta}$AUC) (Bartoń, 2023).

Surfaces

Predicting models onto raster surfaces is oftentimes the most time consuming part of spatial modelling. Multiple pro-active steps were undertaken to reduce this time commitment. The number of time slices which had rasters generated for them were reduced by initially predicting the fit model onto an aspatial prediction matrix with 15 increments along the observed range of each independent variable for the species range. These predictions were then used to determine start and end dates (DOY) for which flowering was likely

to occur, while initially omitting space as an explicit variable. The first DOY with a $> 55\%$ probability of flowering was used as the start date, and the lat DOY with a >60% probability of flowering, were used as temporal constraints for spatial modelling. The higher tolerance for flowering cessation was used because the distribution of flowering generally follows a right skewed distribution. Models were predicted onto rasters at biweekly (14 day) intervals from the start to end DOY, in areas which Species Distribution Models predicted as having a high probability ($> 60\%$) of suitable habitat (Benkendorf et al. in prep). Rasters which had fewer than 5% of their total cells classified as having a >50% probability of flowering were subsequently discarded.

Interpretation

The spatial data were decomposed into tabular data for ease of interpretation by crews. The first and last date in each cell with a probability of flowering greater than 50% was identified and assigned as the respective start and end dates for that taxon in that cell. The peak flower date was simply determined as the time point between the start and end dates with the highest probability of being in flower.

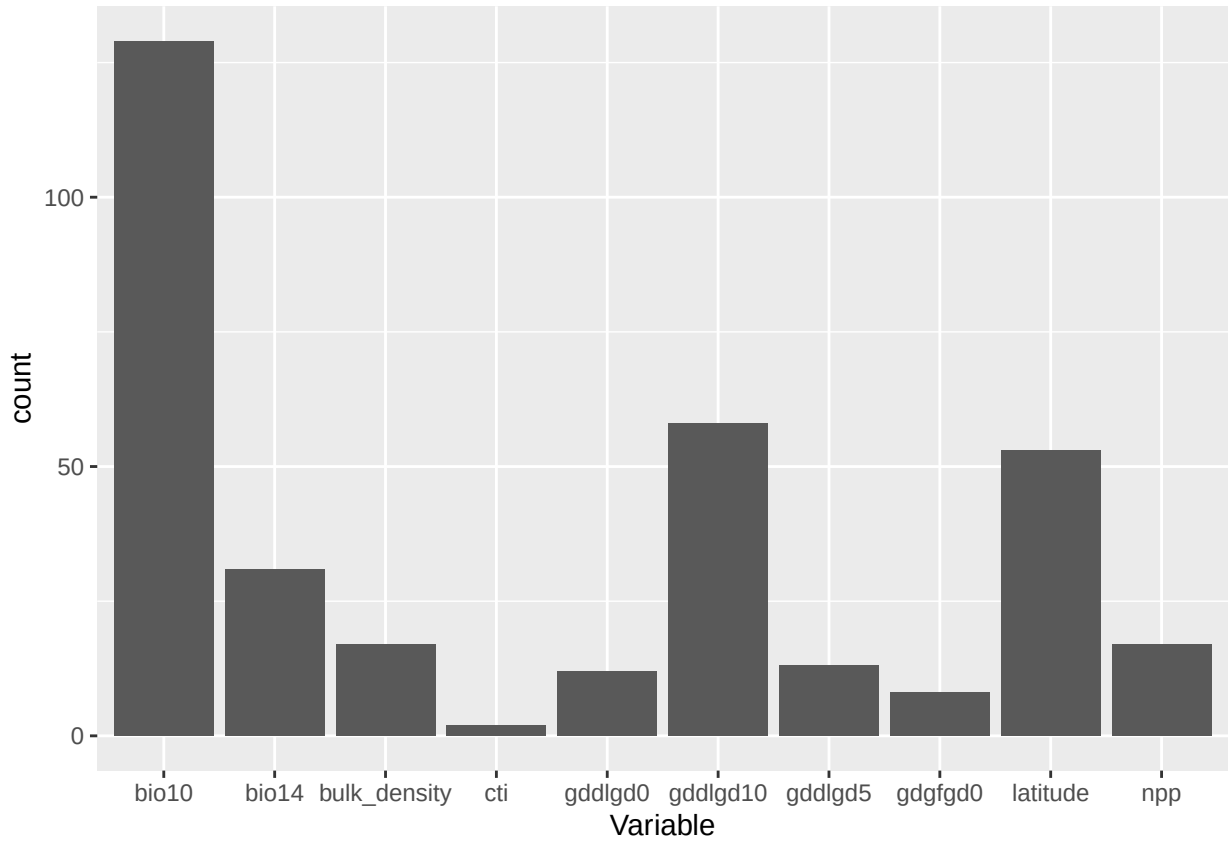
Results

Herbarium Records

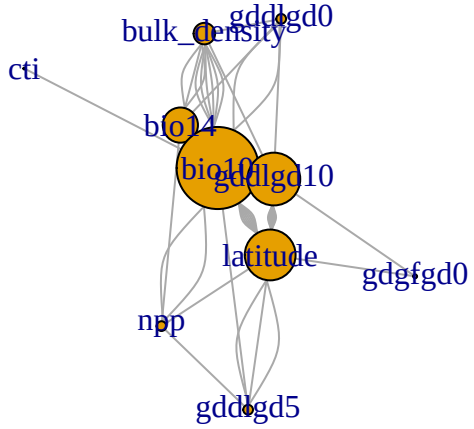
A total of 273 species had a total of 5462 herbarium sheets reviewed. Of these sheets 2618 (47.9%) were not in anthesis. 109 of these species had less than 50% of their records in anthesis. There was strong evidence that lifeforms varied in the proportion of mean number of sheets which were flowering (kruskall wallis p = $1.3383711 \times 10^{-10}$), with more forbs flowering than each of the three other life forms, and some evidence that more graminoids than trees are in anthesis in collections.

Modelling

The number of clusters recovered by hclustgeo ranged from 1 to 5 with a median of 4. The difference in the estimated day of year which flowering started between each of these clusters within a species was mean = 16.46, median = 12.17, and the difference in the estimated day of year flowering ended was mean = 20.62, median = 13.53. The mean difference between the earliest and latest initiation estimate were mean = 37.7 median = 34.68, and for cessation events was mean = 47.21, median = 39.77.




```
##      gddlgd0        bio14     latitude        bio10 bulk_density          npp
##   0.02516537   0.47627628   6.03238703  14.78189666   0.02702703   0.09210526
##     gddlgd10      gddlgd5          cti      gdgfgd0
##   3.56514236   0.00000000   0.00000000   0.00000000
```



In total models for 269 taxa converged.

Which variables ended up making it to the top models? Is this different than species mean LATITUDE???

Ground Verification

What percent of scouting records phen for flowering were within bounds? When was peak % flowering observed?

Discussion

Amatulli, G., D. McInerney, T. Sethi, P. Strobl, and S. Domisch. 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. *Scientific Data* 7: 162.

Augspurger, C. K., and D. N. Zaya. 2020. Concordance of long-term shifts with climate warming varies among phenological events and herbaceous species. *Ecological Monographs* 90: e01421.

Bartoń, K. 2023. MuMIn: Multi-model inference.

Belitz, M. W., E. A. Larsen, L. Ries, and R. P. Guralnick. 2020. The accuracy of phenology estimators for use with sparesely sampled presence-only observations. *Methods in Ecology and Evolution*.

CaraDonna, P. J., A. M. Iler, and D. W. Inouye. 2014. Shifts in flowering phenology reshape a subalpine plant community. *Proceedings of the National Academy of Sciences* 111: 4916–4921.

Chavent, M., V. Kuentz, A. Labenne, and J. Saracco. 2021. ClustGeo: Hierarchical clustering with spatial constraints.

Deva, C., L. Dixon, M. Urban, J. Ramirez-Villegas, I. Droutsas, and A. Challinor. 2024. A new framework for predicting and understanding flowering time for crop breeding. *Plants, People, Planet* 6: 197–209.

Dronova, I., and S. Taddeo. 2022. Remote sensing of phenology: Towards the comprehensive indicators of plant community dynamics from species to regional scales. *Journal of Ecology* 110: 1460–1484.

Gao, F., and X. Zhang. 2021. Mapping crop phenology in near real-time using satellite remote sensing: Challenges and opportunities. *Journal of Remote Sensing*.

Hengl, T., J. Mendes de Jesus, G. B. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, et al. 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one* 12: e0169748.

Hijmans, R. J. 2024. Terra: Spatial data analysis.

Hodges, T. 1990. Predicting crop phenology. Crc Press.

Hodgson, J. A., C. D. Thomas, T. H. Oliver, B. J. Anderson, T. Brereton, and E. Crone. 2011. Predicting insect phenology across space and time. *Global Change Biology* 17: 1289–1300.

Karger, D. N., O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, et al. 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific data* 4: 1–20.

Katal, N., M. Rzanny, P. Mäder, and J. Wäldchen. 2022. Deep learning in plant phenological research: A

systematic literature review. *Frontiers in Plant Science* 13: 805738.

Kuhn, and Max. 2008. Building predictive models in r using the caret package. *Journal of Statistical Software* 28: 1–26.

Michonneau, F., and M. Collins. 2024. Ridigbio: Interface to the iDigBio data API.

Nagai, S., H. Morimoto, and T. M. Saitoh. 2020. A simpler way to predict flowering and full bloom dates of cherry blossoms by self-organizing maps. *Ecological Informatics* 56: 101040.

Nychka, D., R. Furrer, J. Paige, and S. Sain. 2021. Fields: Tools for spatial data.

Park, D. S., I. Breckheimer, A. C. Williams, E. Law, A. M. Ellison, and C. C. Davis. 2019. Herbarium specimens reveal substantial and unexpected variation in phenological sensitivity across the eastern united states. *Philosophical Transactions of the Royal Society B* 374: 20170394.

Park, D. S., Y. Xie, A. M. Ellison, G. M. Lyra, and C. C. Davis. 2023. Complex climate-mediated effects of urbanization on plant reproductive phenology and frost risk. *New Phytologist* 239: 2153–2165.

Parmesan, C., and G. Yohe. 2003. A globally coherent fingerprint of climate change impacts across natural systems. *nature* 421: 37–42.

Pinheiro, J., D. Bates, and R Core Team. 2024. Nlme: Linear and nonlinear mixed effects models.

Polansky, L., and M. M. Robbins. 2013. Generalized additive mixed models for disentangling long-term trends, local anomalies, and seasonality in fruit tree phenology. *Ecology and Evolution* 3: 3141–3151.

Tang, J., C. Körner, H. Muraoka, S. Piao, M. Shen, S. J. Thackeray, and X. Yang. 2016. Emerging opportunities and challenges in phenology: A review. *Ecosphere* 7: e01436.

White, D., and R. B. Gramacy. 2022. Maptree: Mapping, pruning, and graphing tree models.