ECOGRAPHY

Research article

Herbarium data accurately predict the timing and duration of population-level flowering displays

Isaac W. Park[1], Tadeo Ramirez-Parada[2], Sydne Record[3], Charles Davis[4], Aaron M. Ellison[5,6] and Susan J. Mazer[2]

[1]Department of Biology, Georgia Southern University, Statesboro, GA, USA
[2]Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA, USA
[3]Department of Wildlife, Fisheries, and Conservation Biology, University of Maine, Bangor, ME, USA
[4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA
[5]Harvard University Herbaria, Harvard University, Cambridge, MA, USA
[6]Sound Solutions to Sustainable Science, Boston, MA, USA

Correspondence: Isaac W. Park (ipark@georgiasouthern.edu)

Ecography
2024: e06961
doi: 10.1111/ecog.06961

Subject Editor: Alice C. Hughes
Editor-in-Chief: Miguel Araújo
Accepted 24 February 2024

Forecasting the impacts of changing climate on the phenology of plant populations is essential for anticipating and managing potential ecological disruptions to biotic communities. Herbarium specimens enable assessments of plant phenology across broad spatiotemporal scales. However, specimens are collected opportunistically, and it is unclear whether their collection dates – used as proxies of phenological stages – are closest to the onset, peak, or termination of a phenophase, or whether sampled individuals represent early, average, or late occurrences in their populations. Despite this, no studies have assessed whether these uncertainties limit the utility of herbarium specimens for estimating the onset and termination of a phenophase. Using simulated data mimicking such uncertainties, we evaluated the accuracy with which the onset and termination of population-level phenological displays (in this case, of flowering) can be predicted from natural-history collections data (controlling for biases in collector behavior), and how the duration, variability, and responsiveness to climate of the flowering period of a species and temporal collection biases influence model accuracy. Estimates of population-level onset and termination were highly accurate for a wide range of simulated species' attributes, but accuracy declined among species with longer individual-level flowering duration and when there were temporal biases in sample collection, as is common among the earliest and latest-flowering species. The amount of data required to model population-level phenological displays is not impractical to obtain; model accuracy declined by less than 1 day as sample sizes rose from 300 to 1000 specimens. Our analyses of simulated data indicate that, absent pervasive biases in collection and if the climate conditions that affect phenological timing are correctly identified, specimen data can predict the onset, termination, and duration of a population's flowering period with similar accuracy to estimates of median flowering time that are commonplace in the literature.

Keywords: bioclimatology, herbarium specimen, phenology

NORDIC SOCIETY OIKOS
www.ecography.org

Introduction

Climate change has caused widespread shifts in the reproductive periods of populations across species, which may result in profound consequences across levels of ecological organization. To date, the majority of phenological studies has focused on magnitudes of phenological responses in flowering onset or (in the case of specimen-based studies), mean flowering time to climate conditions. However, many of the ecological effects of phenological changes are caused by changes in the duration of a plant species' synchrony with pests or pollinators, or the duration over which a species is exposed to adverse conditions during vulnerable phenophases such as flowering or fruit production (Inouye 2008, Park et al. 2020). Dates of flowering onset or mean flowering dates are not necessarily useful in evaluating these processes, as changes in climate affect may also affect flowering duration (CaraDonna et al. 2014). In such cases, phenological shifts in flowering duration may alter the synchrony among interacting taxa, affecting plant-pollinator interactions (Bodley et al. 2016), interspecific competition for pollinators (Harris 1977, Waser 1978, Anderson and Schelfhout 1980, Rathcke 1988, Forrest et al. 2010), and susceptibility to herbivory (Asch and Visser 2007, Singer and Parmesan 2010) in ways that are not apparent when considering only shifts in their mean timing. Therefore, more fully determining the ecological consequences of phenological shifts attributable to climate change requires that we develop the ability to forecast changes in the duration of each phenophase (e.g. flowering) within local populations by modeling changes in the dates of population-level onsets and terminations for that phenophase. However, field-based phenological records documenting the onset and termination of phenophases across multiple species are limited in geographic or taxonomic scope (Sherry et al. 2011, Crimmins et al. 2013, Bock et al. 2014), and frequently focus on repeated observation of specific individuals rather than local populations. While modern observation networks such as the USA-NPN and iNaturalist have greatly broadened the spatial and taxonomic breadth of records capable of evaluating the timing of flowering onset, duration and termination, (Rosemartin et al. 2014, Pearse et al. 2019, Li et al. 2021), these data are limited in their temporal depth, and still exhibit significant taxonomic and spatial gaps. To date, this has limited our ability to assess climate-driven shifts in phenological synchrony across regions and taxa, highlighting the need for taxonomically and spatially extensive data sources that offer the capacity to estimate the duration of targeted phenophases.

Herbarium records and other specimen-based data represent the most taxonomically, geographically, and temporally extensive source of phenological information for wild and naturalized species (Davis et al. 2015, Willis et al. 2017). Moreover, herbarium specimens have been widely used to estimate phenological responses to climate in temperate regions (Davis et al. 2015, Rawal et al. 2015, Jones and Daehler 2018, Park and Mazer 2018, Park et al. 2019, Taylor 2019, Ramirez-Parada et al. 2022), have captured patterns of phenological variation that are similar to those observed in the field (Miller-Rushing et al. 2006, Ramirez-Parada et al. 2022), and with sufficient statistical correction, can infer similar dates of flowering onset within well-sampled locations to those produced through in-situ observation (Pearse et al. 2017).

Despite their growing use, the utility of herbarium specimens for estimating phenological onset and termination dates may be affected by several limitations. Crucially, herbarium records are sampled opportunistically, providing single snapshots of the phenological status (e.g., flowering) of an individual at a given place and time. As such, it is rarely possible to discern whether a specimen was collected immediately after the onset of a given phenophase, at its peak, or shortly before its termination. Similarly, it is typically not possible to determine whether the sampled individual represents a collection from an early- or late-flowering individual within its local population. Due to these sources of uncertainty, the phenological dates of individual specimens may not reflect the date on which any specific individual- or population-level phenological event occurred. This limitation has the potential to restrict their utility for measuring the precise timing of a given phenophase at the individual level or for estimating the responses to climate of the extremes of a population's temporal phenological distribution (i.e., the onset and termination of a phenophase at the population level). Other researchers have developed methods to infer the flowering duration of individual plants using specimen data (Rossington Love et al. 2019). However, these methods… However, we have also identified certain phenological modalities, such as species that flower close to the start and end of the growing season, where inferences from collections should be examined cautiously.

Whereas methods do exist for inferring onset or termination timing of flowering from herbarium specimens, they require visual assessment of the number of buds, flowers, and fruits (Pearson 2019, Rossington Love et al. 2019), assume a constant flowering duration across the species' geographic range, and are only applicable to species that exhibit multiple flowers and for which specimen imagery is available. These methods are highly time consuming and therefore cannot be easily applied to most digital herbarium holdings.

Additionally, while validation studies have shown that herbarium-based estimates of the temperature sensitivity of mean flowering dates typically match those derived from field observations (Robbirt et al. 2011, Ramirez-Parada et al. 2022), estimates of the first (and last) occurrence of a phenophase are more subject to the effects of outliers and to variation in sampling intensity, population size, and other confounding effects than estimates of mean flowering (Tryjanowski et al. 2005, Miller-Rushing and Primack 2008). These qualitative limitations of specimen data may intrinsically limit the accuracy with which population-level flowering onset and termination can be predicted even when plant phenology responds strongly to well documented aspects of climate. Similarly, it is possible that the number of specimens required to overcome these limitations and produce accurate phenological estimates from these data are prohibitive.

Finally, herbarium records may be subject to several forms of bias when used to estimate the timing of phenological events. First, some species may be preferentially collected during the early or late portion of their local population-level flowering displays. This is most likely to occur with the earliest and latest-flowering species, which flower partially outside of the typical growing season. If specimen collection efforts in general are highest when most species are in flower at a given location, then collection effort is likely to be relatively high in the later portions of early-flowering species' flowering period (i.e. when most other species are flowering) and in early portions of late-flowering species' flowering period. Similarly, species are likely to be less frequently collected during portions of their flowering period that frequently overlap with inclement weather or storm events, as poor weather is associated with reduced collector activity (Daru et al. 2017).

Alternatively, collectors may preferentially collect specimens from individuals within certain portions of their individual flowering period. Evaluations of collections across multiple species have found that collectors often preferentially collect specimens from individuals that are close to their peak flowering date, when the largest numbers of flowers are present (Primack et al. 2004, Davis et al. 2015, Panchen et al. 2019). Conversely, species with fragile flowers or that are subjected to high rates of herbivory may be preferentially collected shortly after the onset of flowering, when petals and other delicate structures are most likely to be intact. Additionally, collectors who prefer specimens bearing both flowers and fruits may collect specimens shortly before flowering termination, when both structures are likely to be present. Despite such biases (Daru et al. 2017, Panchen et al. 2019), recent work by Ramirez-Parada et al. (2022) found that herbarium- and field-based estimates of flowering sensitivity to temperature closely agreed in magnitude and direction despite substantial differences in the timing, location, and associated climate conditions captured by both types of data. However, as this work examined mean flowering time, the implications of these forms of bias for predictions of the timing and duration of the local flowering period for each species remain unknown.

Nevertheless, forecasting changes to the entire distribution of phenological events within a population – rather than simple changes in mean timing – is essential to understanding the effects of climate change on seasonal floral resource availability as well as on a host of ecological processes from pollinator activity to floral vulnerability to frost damage. Determining whether predictions of population-level flowering onset and termination are less accurate than predictions of median flowering, or require larger sample sizes is therefore necessary to leverage the unparalleled taxonomic and spatiotemporal scope of natural-history collections with confidence. Despite this, no studies to date have sought to validate herbarium-based estimates of phenological onsets and terminations, likely because such assessments require a suitable reference dataset of population-level phenological timings against which the accuracy of phenological predictions derived from specimen data can be tested. Unfortunately, extensive field datasets of population-level phenological events across several locations throughout the range of a species are exceedingly rare, limiting our ability to validate such herbarium-based estimates.

In this study, we used simulated phenological data to assess the accuracy of climate-driven models of population-level flowering onset and termination derived from opportunistically sampled data (henceforth, 'phenoclimate' models). These data incorporated uncertainty or bias in the timing of specimen collection relative to the start and end of the flowering period of the sampled individual, and in the relative timing of flowering of the individual relative to its source population. Using these data, we assessed the accuracy of estimated population-level flowering onsets and terminations of simulated plant taxa. We also assessed the degree to which flowering duration of individual plants, intrapopulation variation in flowering time among individual plants, and phenological responsiveness (of mean flowering dates) to differing climate conditions impacted the accuracy of phenological models. We then determined the relationship between data availability and model performance, from which we inferred the number of specimens required to produce reliable phenoclimate models of population-level flowering onset and termination. Finally, we evaluated the effects of 1) biases towards collection of early or late individuals within local populations and 2) biases towards collection of individuals proximate to their flowering onset or termination dates on model accuracy.

Material and methods

Creating a reference dataset: generating sample locations representing known population-level phenological distributions and individual phenological parameters

We simulated phenological data for 1200 hypothetical 'species' in the coterminous USA that varied in the attributes of their individual- and population-level flowering phenology. For each of these simulated species, we selected 1000 locations within the continental United States, each representing a local population observed during a single year from which a simulated specimen was later obtained (Fig. 1a). The coordinates for each location, year, and associated mean annual temperature in the year of collection were randomly selected without replacement from 4-km^2 PRISM pixels (PRISM Climate Group 2011) between the years 1901–2020, and were restricted to locations with 1991–2020 temperature normals of 1–20°C and mean annual precipitation normals for the same period of 60–3800 mm.

Each species generated this way was assigned a series of attributes defining its individual- and population-level flowering phenology. The peak flowering date of an individual was assumed to coincide with its mean flowering date. We then defined a linear equation describing the relationship between the mean date of peak flowering among individuals within a population and local temperature conditions. Each species

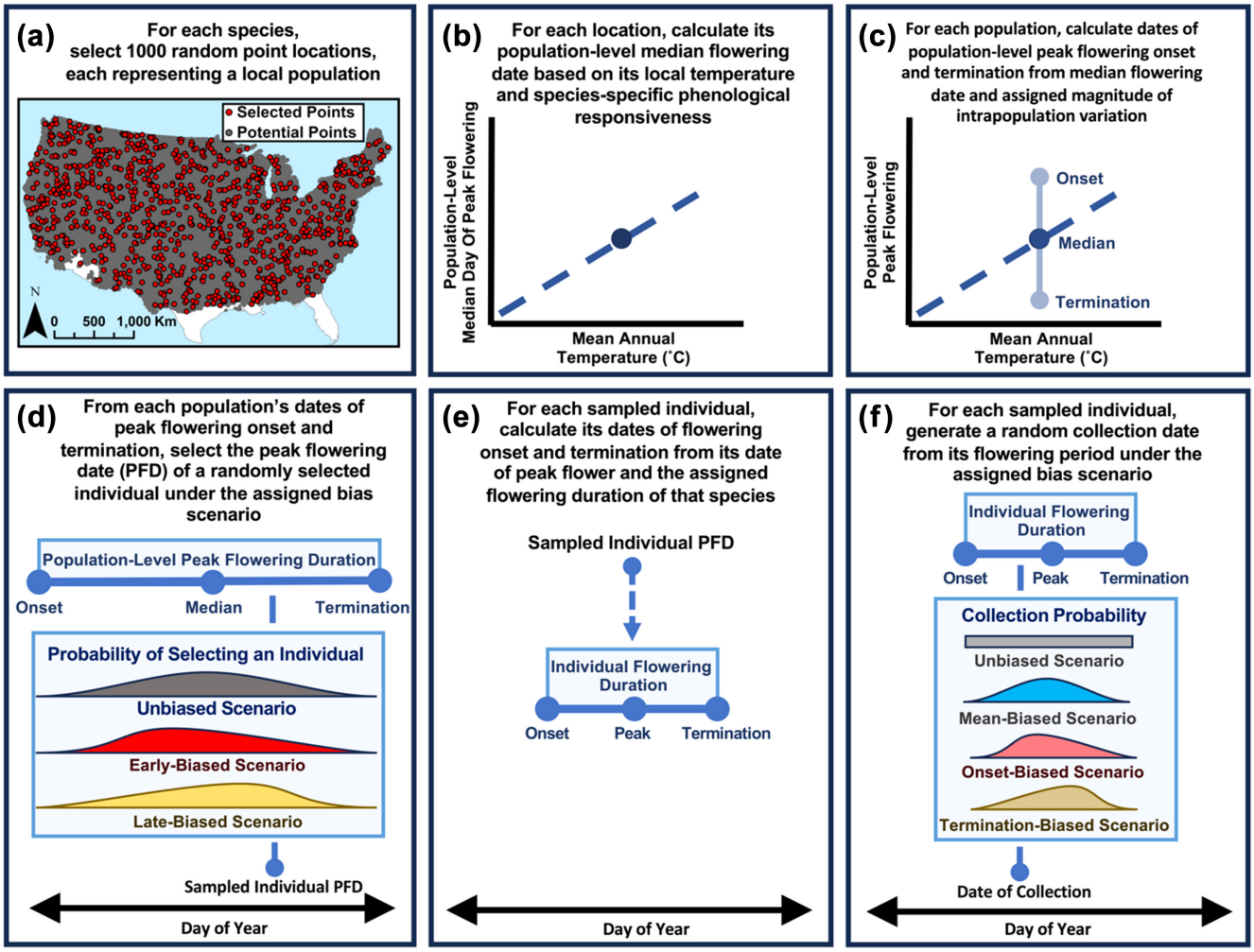

Figure 1. Process by which simulated specimens were generated, beginning by (a) randomly selecting 1000 points for each species, each representing a local population. (b) For each of these points, the median date of peak flowering for a given species was calculated based on the local temperature at that location and the assigned phenological responsiveness of that species. (c) From each population's calculated median date of peak flowering, the dates of peak flowering onset and termination were then calculated using the species' assigned magnitude of intrapopulation variation. (d) From each population's dates of peak flowering onset and termination, we then calculated the duration of peak flowering and randomly selected a date from within that period, which was defined as the date of peak flowering of a single individual selected from that population. While these dates were selected randomly, the probability of selecting each date from within the peak flowering period of a population depended on the type of bias scenario under examination. Individuals could be selected with no bias (shown in grey), with a bias towards collection from the early portion of the population's peak flowering period (shown in red), or with a bias towards collection from the late portion of the population's peak flowering period (shown in yellow). (e) From each individual plant's selected peak flowering date (PFD) we then calculated the dates of that individual's flowering onset and duration (i.e., its individual flowering period) using the individual flowering duration assigned to that species. (f) We then randomly selected a date from within each individual's flowering period to represent the date on which a specimen of that individual was collected. While these dates were selected randomly, the probability of selecting each date from within the peak flowering period of an individual depended on the type of bias scenario under examination. Specimen collection dates were selected either with no bias (shown in grey), with a bias towards collection proximate to the individual's date of peak flowering (shown in blue); with a bias towards collection shortly after flowering onset (shown in red); or with preference towards collection shortly before flowering termination (shown in yellow).

was assigned a median population flowering DOY of 50 at 0°C (i.e. the intercept) as well as a phenological responsiveness (i.e. slope) of median flowering DOY to mean annual temperature: advancing by 1, 4 or 8 days per increase in °C. Next, we assigned each species a low or high magnitude of intrapopulation variation in phenological timing (i.e. in peak flowering DOYs) among individuals (based on normal distributions with standard deviations (σ) of either 10 or 30 days), representing the magnitude of variation in the flowering times of early- to late-flowering individuals within each local population. Then, each species was assigned a short, moderate, or long duration of the flowering period by each individual within each population (15, 30 or 60 days, representing the duration of time each individual plant was in flower. Fifty

species were simulated for each of these 18 combinations of phenological responsiveness, flowering duration, and intrapopulation variation in phenological timing (Table 1).

To accommodate the possibility that the magnitude of variation in phenological timing within a population could depend on local climate conditions, we also simulated 50 species with temperature-sensitive intrapopulation phenological variation (σ) ranging from 10 to 30 days. For these species, σ of the DOY among individuals in a given population increased by 1 day for every 1°C increase in the mean annual temperature of its location (this class of σ is labelled as 'variable' in Table 1). For these simulated species, individual flowering duration was fixed at 30 days. Additionally, to accommodate the possibility that individual flowering durations could exhibit linear relationships with local climate conditions, we also simulated 50 species that exhibited individual-level variation in flowering duration resulting from changes in temperature (increasing by 1 day per °C increase in mean annual temperature, and ranging from 10 to 30 days). For these species, the degree of intrapopulation variation in peak flowering dates was held constant at σ = 30 days (i.e. high intrapopulation variation).

Calculation of population-level onset, median and termination dates of flowering

For each population of each species described above, we calculated a distribution of individual-level peak flowering dates – assumed to be normally distributed (Clark and Thompson 2011) – based on the flowering attributes of the species and the temperature conditions corresponding to its site and year of observation. First, we calculated the median flowering DOY at the location and year from which each specimen was collected based on its pre-defined intercept and phenological responsiveness to mean annual temperature (i.e. 1, 4 and 8 days per °C, Fig. 1b). Then, we obtained the standard deviation of each local population (i.e. its degree of intrapopulation variation in flowering dates) based on the flowering attributes of the simulated species as outlined above. Next, we arbitrarily defined population-level flowering onset DOYs for each population and year as the 10th percentile of a normally distributed population whose mean and standard deviation we obtained in the previous steps (i.e. the DOYs by which the first 10% of individuals in a local population at a given location and year would have reached their median flowering dates). Similarly, the population-level flowering termination dates were calculated as the 90th percentile of a normally distributed population with the same characteristics as described above (i.e. the DOYs by which all but 10% of individuals in a local population at a given location and year would have reached their peak (or mean) flowering dates).

Through this process, we obtained a sample of 1000 annual population-level distributions of flowering dates for each of 1200 hypothetical species. For each of these populations, the quantiles of their flowering distribution – representing the n_{th} individual reaching peak flowering within a population – were known a priori, representing a benchmark against which to compare estimates derived from simulated specimen data (Fig. 1c).

Table 1. The combinations of parameters used to simulate phenological data. For each combination, 50 simulated 'species' were constructed using identical parameters, but with different randomized sample locations and individual collection dates.

Phenological responsiveness	Flowering duration (days)	Intrapopulation variation (sigma)
1 day/°C	15	10 days
		30 days
	30	10 days
		30 days
		Variable
	60	10 days
		30 days
	variable	30 days
4 days/°C	15	10 days
		30 days
	30	10 days
		30 days
		variable
	60	10 days
		30 days
	variable	30 days
8 days/°C	15	10 days
		30 days
	30	10 days
		30 days
		variable
	60	10 days
		30 days
	variable	30 days

Simulating randomly selected (unbiased) phenological snapshots from pre-defined populations

For each species, we then generated simulated specimens by: 1) randomly selecting an individual within each population, and 2) selecting a random DOY within its individual-level flowering period that emulated the phenological snapshot provided by real herbarium specimens. Specifically, using the distribution of peak flowering dates of each population, we selected an individual at random (Fig. 1d). From its peak flowering date, we then obtained onset and termination dates by subtracting (for flowering onset) or adding (for flowering termination) half the individual's flowering duration for that species to the sampled date of peak flowering (Fig. 1e). To simulate a phenological snapshot for that individual, we then randomly selected a DOY between the onset and termination of that individual's flowering period (Fig. 1f). As a result, the simulated datum represented a simulated herbarium specimen generated accounting for uncertainty in both the timing of the individual relative to its source population, and in the timing of the collection relative to the onset and termination of that individual's flowering period. This procedure was repeated across all locations for each simulated species, generating 1000 data points (i.e. simulated specimens or phenological snapshots) per species.

Simulating biases in collection effort across population-level flowering periods

To simulate biases towards collection of specimens during the early or late portion of their local population-level flowering displays, we selected an individual at random within each population and year using both left- and right-skewed normal probability distributions. These distributions were constructed by modulating the parameter α in the python package 'scipy.stats.skewnorm' ver. 1.10.1 (Azzalini and Capitanio 1998), such that if the underlying plant population was treated as exhibiting a normal distribution (α = 0), samples were collected from that population with a left-skewed (α = −1.0) or right-skewed (α = −1.0) probability distribution (Fig. 1d, Supporting information). Once an individual was selected from these skewed distributions, the timing of sample collection from within the individual flowering durations of these 'specimens' was generated using similar methods as unbiased specimens. We then determined the accuracy of the model predictions generated from datasets exhibiting biased and unbiased sampling of local populations by comparing predicted population-level flowering onset and termination dates with the actual (i.e. known, simulated) flowering dates that were produced using a normal distribution. To minimize computation time, population-level biases were examined only for the subset of species for which phenological responsiveness to mean annual temperature equaled 4 days/°C (representing moderate responsiveness to climate stimuli), intrapopulation variation was high (σ = 30), and individual flowering duration was moderate (30 days).

Simulating biases in the timing of collection within flowering periods of individuals

In addition to biases towards collection of early or late individuals within a local population, botanists may also preferentially collect individuals from the early or late portion of their individual flowering period (i.e. individual collection bias). In some cases, collectors may preferentially collect individuals that are proximate to their peak flowering date because this is when the most flowers are displayed. In other cases, collectors may preferentially collect specimens that have only recently begun to flower, when floral structures may exhibit less damage from inclement weather or herbivores, or proximate to flowering termination in cases where the collector prefers specimens that include both flowers and fruits. Accordingly, for each population of each species, we simulated DOYs within each individual's flowering period both at random (i.e., without bias) and with three different types of bias (Fig. 1f, Supporting information). Unbiased collections were simulated by selecting a random date chosen uniformly within the flowering period of each sampled individual (Fig. 1f, Supporting information). To represent a bias toward collection of individuals close to their peak (median) flowering DOY, we sampled collection dates from a truncated normal distribution centered on an individual's mean flowering date and with σ = 25% of the flowering duration for that species and location (henceforth referred to as mean-biased collection data, Fig. 1f, Supporting information). To represent a bias toward collection dates shortly after flowering onset (henceforth, onset-biased collection data), we sampled collection dates from a truncated normal distribution centered on a date 25% earlier than the mean flowering onset date of that individual (σ = 25%; Fig. 1f, Supporting information). Finally, to represent a bias toward collection on dates shortly before flowering termination (henceforth termination-biased collection data), we sampled collection dates from a truncated normal distribution centered on a date 25% later than the mean flowering onset date of that individual (σ = 25%; Fig. 1f, Supporting information). As with examinations of population-level bias, collection biases within the flowering periods of individuals were examined only for the subset of species for which phenological responsiveness to mean annual temperature equaled 4 days/°C, intrapopulation variation was high (σ = 30), individual flowering duration was moderate (30 days), and no population-level bias was present.

Estimating population-level flowering onsets and terminations from simulated herbarium data

We generated phenoclimate models for each species from each set of simulated specimen collection dates using quantile regression (Koenker et al. 2018) in RStudio (www.r-project.org). In all cases, each model regressed observed DOYs of the phenological snapshots of all sampled individuals of a

Table 2. Summary of linear models designed to detect significant effects of collection bias, sample count, intrapopulation phenological variation (sigma), flowering duration, and phenological responsiveness on mean absolute error (MAE) of predicted dates of onset, peak (median), and termination of the flowering period for a given population.

Parameter	df	Response variable					
		Flowering onset		Peak flowering		Flowering termination	
Predictor variables		F-Score	p	F-Score	p	F-Score	p
Collection bias	3	9565.4	**<0.01**	29590.7	**<0.01**	9400.2	**<0.01**
Sample count	9	201.7	**<0.01**	267.5	**<0.01**	193.3	**<0.01**
Sigma	2	229.0	**<0.01**	341.4	**<0.01**	223.2	**<0.01**
Flowering duration	3	21128.3	**<0.01**	14390.5	**<0.01**	20624.9	**<0.01**
Phenological responsiveness	2	0.3	0.79	2.1	0.13	0.2	0.79
Error	59980						
Total	59999						

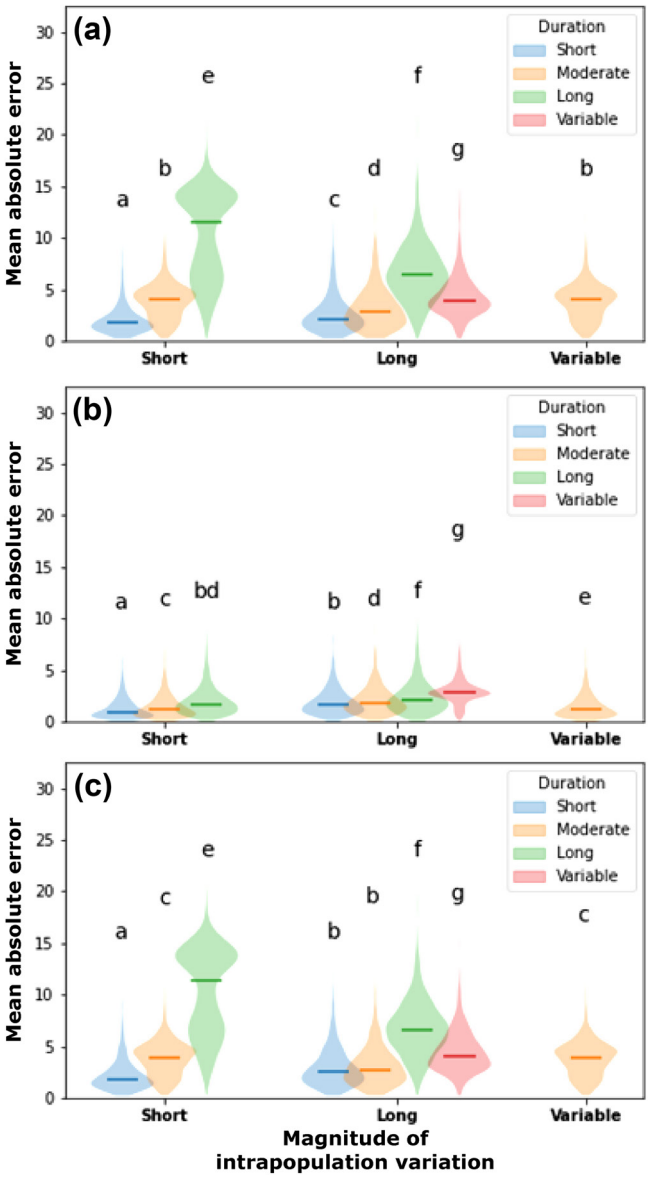

Figure 2. Distribution of mean absolute error (MAE) among phenoclimate models of (a) flowering onset DOY, (b) median flowering DOY and (c) flowering termination DOY trained using simulated species exhibiting low ($\sigma = 10$ days), high ($\sigma = 30$ days), or variable intrapopulation variation in flowering DOY, as well as short (15 days), moderate (30 days), long (60 days) or variable individual flowering duration. Within each panel, groups of models associated with different letters exhibit statistically different mean MAEs among groups of taxa. Where statistically significant differences in MAE were detected, statistical significance was high ($p < 0.001$) in all cases.

given species against mean annual temperature. From these 1450 models (representing each of the species-specific models for all 1200 species plus the additional 150 models exhibiting population-level collection biases and the 100 models exhibiting individual-level collection biases), we predicted the 10th, 50th and 90th percentiles of flowering DOYs for each species from mean annual temperatures corresponding to the years and locations of their source populations. We then calculated the mean absolute error (MAE) of the linear regression of the known timing of the onset (or termination) of the peak flowering period for each reference population on the predicted DOYs produced by each phenoclimate model based on the simulated herbarium data. For each metric of population-level phenology (i.e. flowering onset, peak (i.e. median DOY), and termination), we then used Tukey HSD tests to compare the mean accuracies (estimated as MAE) of these predicted DOYs versus the actual population-level metrics among models constructed from species that differed in their phenological sensitivities to climate, flowering durations, degrees of intrapopulation variation in phenological timing, and collection biases.

Similarly, we tested whether the mean MAE of estimated peak flowering onset and termination dates among groups of species that exhibited the same flowering duration, phenological responsiveness, and intrapopulation phenological variation differed significantly from the mean MAE of estimated median flowering dates for each group of simulated species that exhibited the same flowering duration, phenological responsiveness, and intrapopulation phenological variation. We used Tukey HSD tests to compare the accuracy of estimated onset, median, and termination dates of the peak flowering period among all species produced from each of the simulated datasets.

Finally, we re-fit all 1200 models (including all 24 combinations of species parameters but excluding models constructed to test the effects of collection biases) with randomly selected subsets of data (100–1000 specimens per species) to determine how sample size affected model performance and predictive accuracy. To evaluate whether more data would be needed when variation in phenology among populations is not perfectly explained by the climate variables included in the model, we ran additional simulations in which population-level mean DOYs (and associated onset and termination DOYs of the flowering period) of each species at each sampled location and year included random variation not associated with local climate: adding either $\pm$ 5 days (i.e. a low-noise scenario) or $\pm$ 15 days (i.e. a high-noise scenario) to the DOYs of the onset, median, and termination of flowering DOYs. For each location and year, the random offsets of the DOYs of flowering onset, median flowering DOY, and flowering termination were identical, such that random variation was incorporated only into the timing of flowering, and not its duration.

Results

Effects of species characteristics on model accuracy

We obtained five general results from our comparisons of model accuracy when using simulated data sets characterized by different combinations of phenological sensitivity to temperature, phenological parameters (e.g. duration and standard deviation of flowering times), and sample size. First, the magnitude of species' responsiveness to climate had no significant effect on model accuracy when predicting DOYs of flowering

Table 3. Median MAE (in days) of estimated onset, peak (median), and termination of flowering period by phenoclimate models constructed using species simulated using each combination of parameters. Bold numbers indicate significant differences between the mean MAEs of estimated onset and/or termination dates of the flowering period and the MAE of peak flowering dates within each group of taxa. Where statistically significant differences were detected, differences were highly significant ($p < 0.001$) in all cases.

Flowering duration (days)	Intrapopulation variation (sigma)	MAE in onset of flowering period	MAE in peak flowering	MAE in termination of flowering period
15 days	10 days	**1.5**	0.6	**1.5**
	30 days	**2.2**	1.0	**2.5**
30 days	10 days	**4.5**	0.8	**4.6**
	30 days	**2.8**	1.8	**2.8**
	variable	**1.8**	1.3	**1.8**
60 days	10 days	**14.0**	1.4	**13.9**
	30 days	**6.4**	2.1	**6.7**
Variable	30 days	**3.7**	1.9	**3.5**

onset, median, or termination using unbiased collections ($F \leq 0.3$, $p \geq 0.13$, Table 2). The magnitude of intrapopulation phenological variation, individual flowering duration, sample size, and collection bias did exhibit significant effects on the accuracy of predicted flowering onset, median, and termination DOYs ($p < 0.01$ in all cases, Table 2). However, mean MAE of predicted median (or peak) flowering DOY remained both low and consistent across all categories of taxa (ranging from 0.6 days at minimum, to 1.9 days at maximum, Fig. 2, Table 3). Predictions of flowering onset and termination DOYs exhibited higher MAE than estimates of median flowering DOYs across all categories of species ($p < 0.001$ in all cases, Table 3), but also remained under five days unless individual flowering duration was long (60 days; Fig. 2, Table 3). The mean MAE of predicted median flowering remained under 2.1 days among species exhibiting long individual flowering durations. However, estimation errors for onset and termination DOYs were quite high, with MAEs reaching as high as 14 days when individual flowering durations were long and intrapopulation variation was low (Fig. 2, Table 3).

Effects of sample size on model accuracy

Although sample size exhibited significant effects on model accuracy, the magnitude of the changes in MAE was < 2 days, and remained consistent across predictions of flowering onset, median, and termination DOYs (Fig. 3). In all cases, MAE declined with larger sample sizes, but MAE exhibited < 2 days improvement as sample size increased from 100 to 1000 specimens in all cases (Fig. 3, Supporting information). Increases in model performance as sample size increased above 300 were minimal, and never exceeded a one-day reduction in MAE (Supporting information).

Increased magnitudes of unexplained (i.e. stochastic) variation in phenological timing among populations within a species also were associated with increased MAE as expected (Fig. 3), but exhibited similar relationships to sample size as noiseless models. This implies that unexplained phenological variation inherently degrades the accuracy of phenoclimate models, but the effects of unexplained variation in phenological timing cannot be remedied by greater quantities of sample data.

Effects of population-level collection biases on model accuracy

Phenoclimate models derived from sample data that exhibited population-level biases in collection timing (i.e. biases towards collection of early or late individuals from within each population) exhibited substantially higher MAE than models produced using unbiased collections (Fig. 4). The greatest increases in MAE were observed among predictions of flowering termination derived from collections biased towards early-flowering individuals, and among predictions of flowering onset derived from collections biased towards late-flowering individuals. However, predictions of median (peak) flowering DOY derived from early- or late-biased collections also exhibited significant reductions in accuracy, with mean MAE among models derived from biased collections sometimes exceeding two weeks (16.4 days, Fig. 4). Moreover, phenoclimate models appear to be highly sensitive to < 2 days of population-level temporal biases in collections, with MAE of all predictions exceeding five days even when skew was low ($\alpha = \pm 0.25$, Supporting information), more than doubling the observed MAE of phenoclimate models developed from unbiased collections.

Effects of individual collection biases on model accuracy

Phenoclimate models derived from sample data that exhibited biases towards collection of specimens proximate to the beginning or end of their individual flowering periods exhibited higher MAE than models constructed from specimens collected with no inherent bias, with mean MAE among species exceeding eight days in all cases (Fig. 5). However, models constructed from specimens collected with a bias towards collection of specimens proximate to their peak flowering DOY consistently exhibited lower MAE than models derived from unbiased data (Fig. 5). However, the effects of this form of bias are intrinsically linked to individual flowering duration (with longer durations associated with higher MAE); models produced from species exhibiting bias towards collection of specimens shortly after onset or shortly before termination exhibited greater accuracy and lower MAE among species with 15- or 30-day flowering durations (Supporting information) than among species with 60-day durations.

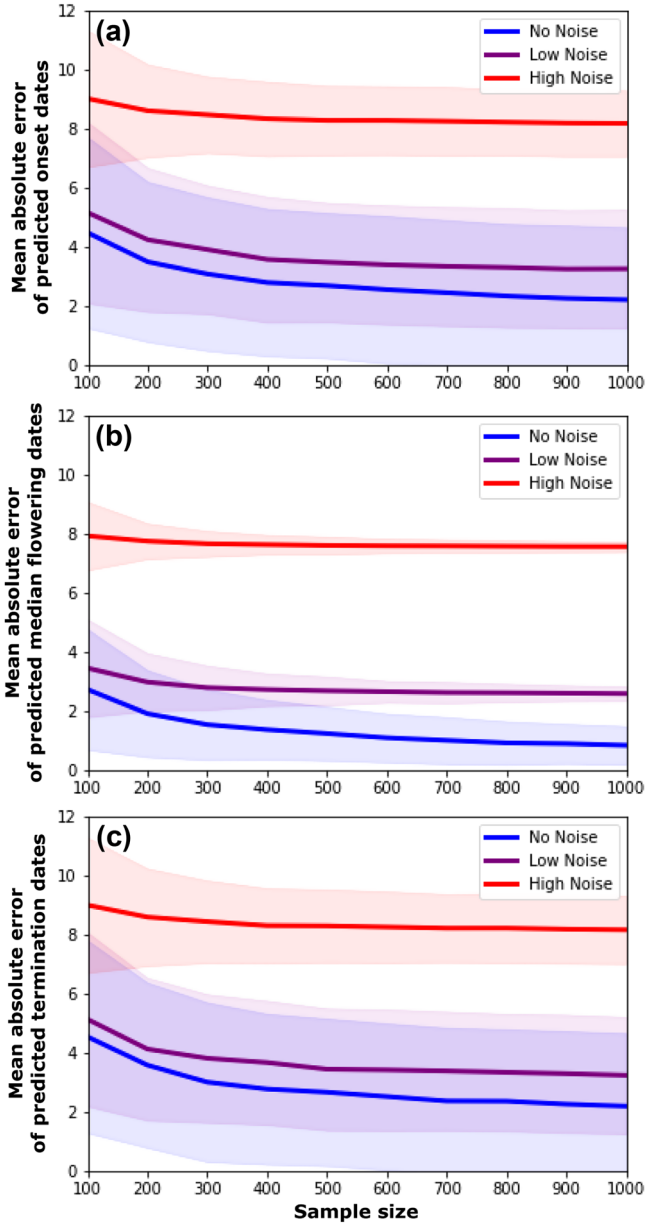

Figure 3. Median MAE of modeled population-level onset, peak, and termination flowering dates. Blue lines correspond to phenoclimate models in which population-level phenological timing varied only with local climate (i.e. a no noise scenario). Purple lines correspond to phenoclimate models in which population-level phenological timing exhibits unexplained variation of ± 5 days (i.e. a low noise scenario), while red lines correspond to phenoclimate models in which population-level phenological timing exhibit unexplained variation of ±15 days (i.e. a high noise scenario).

Discussion

Our results demonstrate that the intrinsic limitations of phenological data derived from herbarium collections – assuming other forms of bias are not pervasive – do not preclude the development of accurate phenoclimate models capable of predicting the timing of population-level flowering onset or termination, and are only slightly less accurate than predictions of median flowering date. Further, the accuracies of these models are not likely to be closely tied to the magnitude of phenological variation among individuals of a species, and can be produced with similar quantities of data as more traditional models of mean flowering phenology (Park and Mazer 2018). However, this study does identify several limitations to the prediction of population-level flowering onset and termination DOYs from herbarium data that may impact the reliability of such predictions.

First, our simulations demonstrate that the accuracy of specimen-derived phenoclimate models can be highly sensitive to biases in collection timing within populations (Fig. 4). The frequency with which such biases occur is not well documented, although they are more likely to be problematic among species that flower at the beginning or end of the local growing season in temperate climates, as collection activity is frequently lower during winter and other unfavorable conditions (Daru et al. 2017). Additionally, collection activity may be reduced during the early portion of the flowering display among the earliest-flowering species, as relatively few species in a given location or region are likely to be vegetatively or reproductively active during those periods. Alternatively, collection activity may be higher than normal in the beginning of the spring. Similarly, collection activity may be low during the later portions of the flowering periods of some late-flowering species that flower after most species have ceased flowering or gone dormant. Thus, predicted timings should be viewed with greater caution when modeling the timing of flowering onset or termination among early spring or late fall-flowering species.

Second, model predictions will likely be less accurate among long-flowering species, as longer individual flowering durations were consistently associated with lower model accuracy across simulated taxa. This pattern has previously been observed in attempts to evaluate accuracy of specimen-based phenoclimate models (Pearson 2019). Long flowering durations also amplify the deleterious effects of biases towards collection of specimens from specific portions of individual bloom displays, as longer individual flowering periods necessarily increase the magnitude of temporal bias that can be introduced by collector preference towards recently opened or nearly completed flowers. Fortuitously, herbarium specimens most frequently have been documented to exhibit biases towards collection proximate to peak flowering DOY (Primack et al. 2004, Davis et al. 2015, Panchen et al. 2019), which notably produced more accurate phenoclimate models of both flowering onset and termination than unbiased collections, particularly among long-flowering species. Thus, for species that exhibit charismatic or notable peaks in their individual flowering displays, collector biases may actually improve rather than hinder phenoclimate modeling conducted using these methods.

Third, and finally, our study assumes that the climate stimuli to which species exhibit phenological responses can be sufficiently captured by available climate data to drive such models; thus, the magnitudes of error presented here should

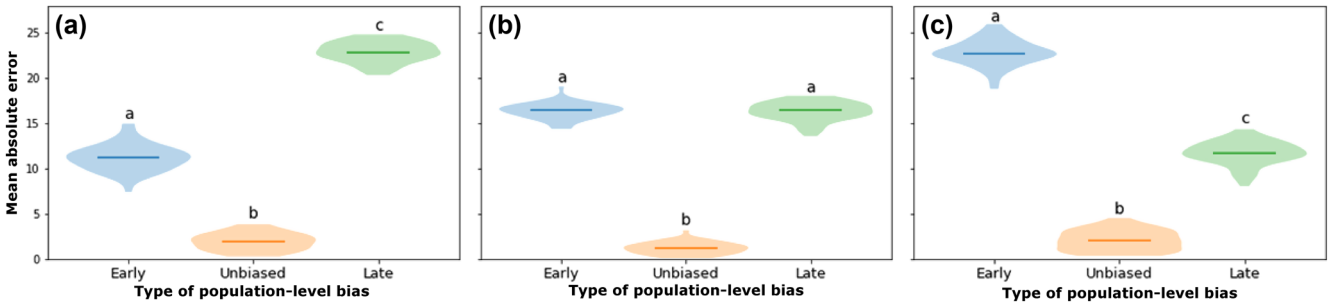

Figure 4. Distribution of MAE among phenoclimate models of (a) population-level flowering onset DOY, (b) population-level peak flowering DOY and (c) population-level flowering termination DOY trained using simulated species collected with a bias towards early individuals, collected without bias, or with bias towards late-flowering individuals within each local population. All species included in these models exhibited a phenological responsiveness of 4 days/°C, a high degree ($\sigma = 30$ days) of intrapopulation variation, and moderate (30 day) individual flowering durations. Within each panel, groups of models associated with different letters exhibit statistically different mean MAEs among groups of taxa. Where statistically significant differences in MAE were detected, statistical significance was high ($p < 0.001$) in all cases.

also not be taken to represent expected model accuracy when predicting phenological timings of real plant taxa, as the simulations presented here corresponded to an ideal situation in which all among-population variation in phenological timing could be explained by a single climate variable. Under real-world conditions, we may expect that some component of phenological variation will be explained by aspects of local conditions that cannot be easily captured using available climate data. Thus, our study demonstrates that specimen-based phenological snapshots enable estimation of population-level onsets and terminations despite noise and biases in the timing of collection, but the accuracy of herbarium-based predictions in actual plant populations will likely depend on i) the degree to which available climate data capture the drivers of its phenological variation over space and time, and ii) whether the most relevant climate factors have been identified and incorporated into phenoclimatic models. Consequently, phenological predictions of species that exhibit highly stochastic phenological timing, occupy sites with high degrees of microhabitat variation, or are highly sensitive to variation in aspects of the local environment that are not easily captured using broad-scale gridded data are likely to be less accurate regardless of what aspects of a given phenophase are being assessed. Similarly, species that exhibit spatial biases towards collection solely in specific habitats or regions (Erickson and Smith 2021) or that exhibit broad seasonal biases in collection effort are likely to be less accurate. However, as many studies have indicated that strong linear phenological responses can be captured from monthly, seasonal, or annual temperature at moderate spatial resolutions (Miller-Rushing et al. 2006, Gerst et al. 2017, Park and Mazer 2018), this is unlikely to represent a major obstacle in modelling the phenology of most plant species in temperate environments.

Future directions

These results indicate that, with some caveats, there is no fundamental barrier that prevents the prediction of population-level flowering timing and duration from specimen-based phenoclimate models. Further, our results show that few additional data are needed than have been used by phenoclimate models predicting simple mean (or median) phenological dates (Park and Mazer 2018, Ramirez-Parada et al.

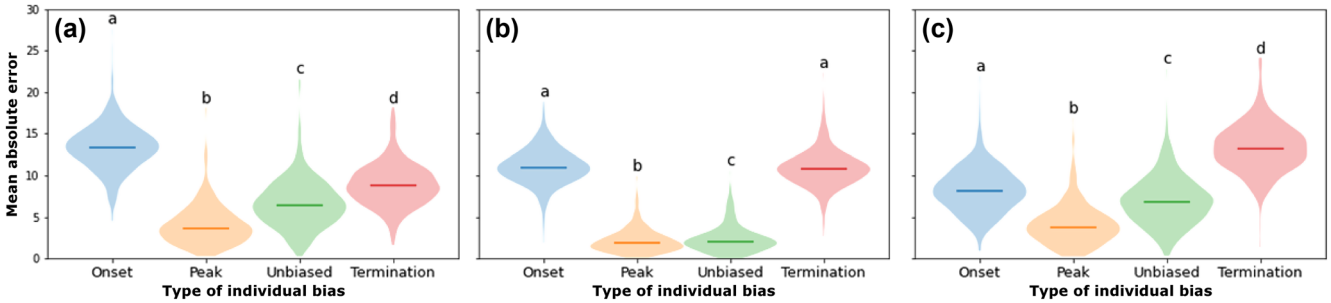

Figure 5. Distribution of MAE among phenoclimate models of (a) population-level flowering onset DOY, (b) population-level peak flowering DOY and (c) population-level flowering termination DOY trained using simulated species bias towards individuals collected shortly after their flowering onset, proximate to their peak flowering DOY, without bias, or with bias towards collection shortly before the end of that individual's flowering period. All species included in these models exhibited a phenological responsiveness of 4 days/°C, a high degree ($\sigma = 30$ days) of intrapopulation variation, and long (60 day) individual flowering durations. Within each panel, groups of models associated with different letters exhibit statistically different mean MAEs between groups of taxa. Where statistically significant differences in MAE were detected, statistical significance was high ($p < 0.001$) in all cases.

2022). However, we have also identified certain phenological modalities, such as species that flower close to the start and end of the growing season, where inferences from collections should be examined cautiously.

Although our simulations were conducted on plant flowering phenology, the underlying results may apply to the development of phenoclimate models of any taxon whose phenology can be assessed from herbaria or other natural history collections data. While the accuracy of those models was not explicitly tested, similar methods have already been used to evaluate the activity period of bee species across the northeastern US (Dorian et al. 2022). Thus, widespread assessment should be possible of the effects of climate change on many other taxa and on synchrony among co-occurring plant species, and plants and their pollinators, pests, and frugivores.

Funding – This work was supported by NSF DEB-1556768 (to SJM, IWP), NSF DEB-2105932 (to SJM, IWP, CCD, AME), and DEB-2242804 (to SR). TRP is grateful to UCSB for fellowship support in the year this manuscript was completed.

Author contributions

Isaac W. Park: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Funding acquisition (equal); Investigation (equal); Methodology (lead); Visualization (lead); Writing – original draft (lead). **Tadeo Ramirez-Parada:** Conceptualization (equal); Formal analysis (supporting); Investigation (equal); Methodology (equal); Visualization (equal); Writing – review and editing (equal); **Sydne Record:** Funding acquisition (equal); Methodology (supporting); Visualization (supporting); Writing – review and editing (equal). **Charles Davis:** Funding acquisition (equal); Methodology (supporting); Visualization (supporting); Writing – review and editing (equal). **Aaron M. Ellison:** Funding acquisition (equal); Methodology (supporting); Visualization (supporting); Writing – review and editing (equal). **Susan J. Mazer:** Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Investigation (equal); Methodology (equal); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at https://publons.com/publon/10.1111/ecog.06961.

Data availability statement

Data and code associated with this paper are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.dbrv15f79 (Park et al. 2024).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Anderson, R. C. and Schelfhout, S. 1980. Phenological patterns among talgrass prairie plants and their implications for pollinator competition. – Am. Midl. Nat. 104: 153–163.

Asch, M. V. and Visser, M. E. 2007. Phenology of forest caterpillars and their host trees: the importance of synchrony. – Annu. Rev. Entomol. 52: 37–55.

Azzalini, A. and Capitanio, A. 1998. Statistical applications of the multivariate skew-normal distribution. – J. R. Stat. Soc. B 61: 579–602.

Bock, A., Sparks, T. H., Estrella, N., Jee, N., Casebow, A., Schunk, C., Leuchner, M. and Menzel, A. 2014. Changes in first flowering dates and flowering duration of 232 plant species on the island of Guernsey. – Global Change Biol. 20: 3508–3519.

Bodley, E. J., Beggs, J. R., Toft, R. and Gaskett, A. C. 2016. Flowers, phenology and pollination of the endemic New Zealand greenhood orchid *Pterostylis brumalis*. – N. Z. J. Bot. 54: 291–310.

CaraDonna, P. J., Iler, A. M. and Inouye, D. W. 2014. Shifts in flowering phenology shape a subalpine plant community. – Proc. Natl Acad. Sci. USA 111: 4916–4921.

Clark, R. M. and Thompson, R. 2011. Estimation and comparison of flowering curves. – Plant Ecol. Divers. 4: 189–200.

Crimmins, T. M., Crimmins, M. A. and Bertelsen, C. D. 2013. Spring and summer patterns in flowering onset, duration, and constancy across a water-limited gradient. – Am. J. Bot. 100: 1137–1147.

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., Seidler, T. G., Sweeney, P. W., Foster, D. R., Ellison, A. M. and Davis, C. C. 2017. Widespread sampling biases in herbaria revealed from large-scale digitization. – New Phytol. 217: 939–955.

Davis, C. C., Willis, C. G., Connolly, B. and Ellison, A. M. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. – Am. J. Bot. 102: 1599–1609.

Dorian, N. N., McCarthy, M. W. and Crone, E. E. 2022. Ecological traits explain long-term phenological trends in solitary bees. – J. Anim. Ecol. 92: 285–296.

Erickson, K. D. and Smith, A. B. 2021. Accounting for imperfect detection in data from museums and herbaria when modeling species distributions: combining and contrasting data-level versus model-level bias correction. – Ecography 44: 1341–1352.

Forrest, J., Inouye, D. W. and Thomson, J. D. 2010. Flowering phenology in subalpine meadows: does climate variation influence community co-flowering patters? – Ecology 91: 431–440.

Gerst, K. L., Rossington, N. L. and Mazer, S. J. 2017. Phenological responsiveness to climate differs among four species of *Quercus* in North America. – J. Ecol. 105: 1610–1622.

Harris, G. A. 1977. Root phenology as a factor of competition among grass seedlings. – J. Range Manage. 30: 172–177.

Inouye, D. W. 2008. Effects of climate change on phenology, frost damage, and floral abundance of montane wildflowers. – Ecology 89: 353–362.

Jones, C. A. and Daehler, C. C. 2018. Herbarium specimens can reveal impacts of climate change on plant phenology; a review of methods and applications. – PeerJ 6: e4576.

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P. and Ripley, B. D. 2018. – In: Koenker, R., (ed.), Quantile regression: package 'quantreg'. CRAN, https://cran.r-project.org/web/packages/quantreg/index.html.

Li, D., Barve, N., Brenskelle, L., Earl, K., Barve, V., Belitz, M. W., Doby, J., Hantak, M. M., Oswald, J. A., Stucky, B. J., Walters, M. and Guralnick, R. P. 2021. Climate, urbanization, and species traits interactively drive flowering duration. – Global Change Biol. 27: 892–903.
Miller-Rushing, A. J. and Primack, R. B. 2008. Global warming and flowering times in Thoreau's Concord: a community perspective. – Ecology 89: 332–341.
Miller-Rushing, A. J., Primack, R. B., Primack, D. and Mukunda, S. 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. – Am. J. Bot. 93: 1667–1674.
Panchen, Z. A., Doubt, J., Kharouba, H. M. and Johnston, M. O. 2019. Patterns and biases in an Arctic herbarium specimen collection: implications for phenological research. – Appl. Plant Sci. 7: e01229.
Park, I. W. and Mazer, S. J. 2018. Overlooked climate parameters best predict flowering onset: assessing phenological models using the elastic net. – Global Change Biol. 24: 5972–5984.
Park, D. S., Breckheimer, I., Williams, A. C., Law, E., Ellison, A. M. and Davis, C. C. 2019. Herbarium specimens reveal substantial and unexpected variation in phenological sensitivity across the eastern United States. – Phil. Trans. R. Soc. B 374: 20170394.
Park, I. W., Ramirez-Parada, T. and Mazer, S. J. 2020. Advancing frost dates have reduced frost risk among most North American angiosperms since 1980. – Global Change Biol. 27: 165–176.
Park, I. W., Ramirez-Parada, T., Record, S., Davis, C., Ellison, A. M. and Mazer, S. J. 2024. Data from: Herbarium data accurately predict the timing and duration of population-level flowering displays. – Dryad Digital Repository, https://doi.org/10.5061/dryad.dbrv15f79.
Pearse, W. D., Davis, C. C., Inouye, D. W., Primack, R. B. and Davies, T. J. 2017. A statistical estimator for determining the limits of contemporary and historic phenology. – Nat. Ecol. Evol. 1: 1876–1882.
Pearson, K. D. 2019. A new method and insights for estimating phenological events from herbarium specimens. – Appl. Plant Sci. 7: e01224.
Primack, D., Imbres, C., Primack, R. B., Miller-Rushing, A. J. and Del Tredici, P. 2004. Herbarium specimens demonstrate earlier flowering times in response to warming in Boston. – Am. J. Bot. 91: 1260–1264.
PRISM Climate Group 2011. 1971–2000 climatology normals. – Oregon State Univ., http://prism.oregonstate.edu.
Ramirez-Parada, T. H., Park, I. W. and Mazer, S. J. 2022. Herbarium specimens provide reliable estimates of phenological responses to climate at unparalleled taxonomic and spatiotemporal scales. – Ecography 2022: e06173.
Rathcke, B. 1988. Flowering phenologies in a shrub community: competition and constraints. – J. Ecol. 76: 975–994.
Rawal, D. S., Kasel, S., Keatley, M. R. and Nitschke, C. R. 2015. Herbarium records identify sensitivity of flowering phenology of eucalypts to climate: implications for species response to climate change. – Austral Ecol. 40: 117–125.
Robbirt, K. M., Davy, A. J., Hutchings, M. J. and Roberts, D. L. 2011. Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid *Ophrys sphegodes*. – J. Ecol. 99: 235–241.
Rosemartin, A. H., Crimmins, T. M., Enquist, C. A. F., Gerst, K. L., Kellermann, J. L., Posthumus, E. E., Denny, E. G., Guertin, P., Marsh, L. and Weltzin, J. F. 2014. Organizing phenological data resources to inform natural resource conservation. – Biol. Conserv. 173: 90–97.
Rossington Love, N. R., Park, I. W. and Mazer, S. J. 2019. A new phenoogical metric for use in pheno-climate models: a case study for using herbarium specimens of *Streptanthus tortuosus*. – Appl. Plant Sci. 7: e11276.
Sherry, R. A., Zhou, X., Gu, S., Arnone, J. A., Johnson, D. W., Schimel, D. S., Verburg, P. S. J., Wallace, L. L. and Luo, Y. 2011. Changes in duration of reproductive phases and lagged phenological response to experimental climate warming. – Plant Ecol. Divers. 4: 23–35.
Singer, M. C. and Parmesan, C. 2010. Phenological asynchrony between herbivorous insects and their hosts: signal of climate change or pre-existing adaptive strategy? – Phil. Trans. R. Soc. B 365: 3161–3176.
Taylor, S. D. 2019. Estimating flowering transition dates from status-based phenological observations: a test of methods. – PeerJ 7: e7720.
Tryjanowski, P., Kuźniak, S. and Sparks, T. H. 2005. What affects the magnitude of change in first arrival dates of migrant birds. – J. Ornithol. 146: 200–205.
Waser, N. M. 1978. Competition for hummingbird pollination and sequential flowering in two Colorado wildflowers. – Ecology 59: 934.
Willis, C. G., Ellwood, E. R., Primack, R. B., Davis, C. C., Pearson, K. D., Gallinat, A. S., Yost, J. M., Nelson, G., Mazer, S. J., Rossington, N. L., Sparks, T. H. and Soltis, P. S. 2017. Old plants, new tricks: phenological research using herbarium specimens. – Trends Ecol. Evol. 32: 531–546.