# Title page

**Article title**: A statistical estimator for determining the limits of contemporary and historic phenology

**Authors:** William D. Pearse[1,2,3,*], Charles C. Davis[4], David W. Inouye[5], Richard B. Primack[6] & T. Jonathan Davies[1]

[1] Department of Biology, McGill University, Montréal, QC, Canada

[2] Department of Biological Sciences, Université du Québec à Montréal, Montréal, Québec, Canada

[3] Department of Biology & Ecology Center, Utah State University, 5305 Old Main Hill, Logan UT, 84322

[4] Department of Organismic and Evolutionary Biology and Harvard University Herbaria, Harvard University, 22 Divinity Ave., Cambridge 02138, MA

[5] Department of Biology, University of Maryland, College Park, MD 20742 USA, and Rocky Mountain Biological Laboratory, PO Box 519, Crested Butte, CO 81224 USA.

[6] Biology Department, Boston University 5 Cummington St. Boston, MA 02215, USA

[*] To whom correspondence should be addressed: `will.pearse@usu.edu` and `j.davies@mcgill.ca`

**Word-count**: 2327 (excluding figure legends)

**Climate change affects not just where species are found, but also when species' key life-history events occur—their phenology. Measuring such changes in timing is often hampered by a reliance on biased survey data: surveys identify that an event has taken place (*e.g.*, the flower is in bloom), but not when that event happened (*e.g.*, the flower bloomed yesterday). Here we show that this problem can be circumvented using statistical estimators, which can provide accurate and unbiased estimates from sparsely sampled observations. We demonstrate that such methods can resolve an ongoing debate about the relative timings of the onset and cessation of flowering, and allow us to reliably place modern observations within the context of the vast wealth of historical data that reside in herbaria, museum collections, and written records. We then analyse large-scale citizen-science data from the USA National Phenology Network, and reveal not just earlier but also potentially more variable flowering in recent years. Evidence for greater variability through time is important because increases in variation are characteristic of systems approaching a state change.**

Anthropogenic climate forcing is likely to increase global temperature by more than 1.5 °C by the end of this century[1]. In response to this rapid environmental shift, species must track favourable conditions by moving or altering the timing of their life-history strategies—their phenology—to flower, breed, or migrate sooner[2,3]. However, predicting species' phenological responses is not straightforward: experimental data often do not match observations[4], and sampling of observational data is frequently limited. Citizen scientists[5] and historical collections[6,7] have emerged as valuable sources of ecological data, and on-going efforts to digitise museum and herbarium collections are making available an unprecedented wealth of historical records[8–11]. Despite their promise, such data present numerous statistical challenges: they are often sparsely sampled spatially and unevenly distributed through time[12], and while they can provide information on the relative timing of events they do not necessarily capture their first occurrence. Compounding this problem, most statistical tools are designed to study changes in species' mean responses, not variation in the onset of events.

Here we present a method derived from the extinction biology literature[13] to address these challenges, and provide three case studies that illustrate the potential of the approach in phenological research. While we focus on plant flowering time here, this approach would also be applicable to other systems, such as the phenology of bird migrations and insect emergence, or the limits of other continuous data such as

2

environmental tolerances. First, we revisit an ongoing debate about shifts in timing of the onset, peak (middle), and cessation of flowering. Second, we show how our approach can reconcile distinct datasets with different sampling (historical collections and field observations), greatly expanding the temporal and climatic ranges across which we can measure change. Third, we apply our method to a sparsely sampled citizen science dataset and find evidence not that climate change is not just altering the timing of plant flowering, but also increasing its variability through time.

## Results and discussion

**Statistically estimating the start of a process.** Estimating the onset of a phenological event is but one instance of the more general problem of determining the absolute limit of a distribution. The tails of distributions are infamously difficult to model because there are fewer data to parameterise them and a single data point can invalidate all previous estimates. This challenge is similar to the 'German Tank Problem', which was faced by Allied forces during World War II who wanted to estimate the number of German tanks (the limit of the distribution of serial numbers) but only had access to the sequential serial numbers of observed (defeated) tanks[14]. We suggest here a solution to this problem that parallels methods first described to determine the date a species went extinct[13]. The general approach is to model the distribution of the earliest observations using a (very flexible) Weibull distribution, which provides an estimate of the start of the observed process (*e.g.*, plants flowering). The joint distribution of the most recent sightings has approximately the same Weibull form irrespective of the distribution from which those sightings were sampled[15], making it well-suited to data collected under different sampling regimes. The estimate for the first occurrence of any event is thus the sum of the times of the first $k$ events, weighted in part according to the joint-Weibull-distribution of all the sightings (following [13] who focused on the last $k$ events). While confidence intervals are defined for this estimate, standard errors must be parametrically bootstrapped as their formula is currently unknown[16]. Figure 1 gives an example of how this approach can provide an estimate of when a process (such as flowering) started, even if the very beginning of that process wasn't directly observed.

Using simulations we demonstrate that our approach has greater power to detect the true onset of a process

than existing methods that use only the first observation (see Methods). This is because our approach draws strength from the first $k$ measurements, not just the single earliest observation. This also allows for confidence intervals and standard errors to be placed around an estimate, which is impossible when working with the first observation alone. Just as any measure of the central tendency of a distribution (*e.g.*, a mean) should not be considered in isolation of the distribution and number of observations underlying it, the same is true of estimates of the limits of a distribution. We note, also, that attempting to estimate the limit of a distribution by averaging across estimates, as is common in phenological studies, is inherently biased: the average of the two (or more) earliest observations must, by definition, be later than the earliest observation. This has implications not just for generating mean estimates of the onset of flowering, but also for commonly-used statistical models that implicitly rely upon averages (*e.g.*, ANOVA and multiple regression). The following case studies illustrate the potential of our approach.

**Relative change in the onset, peak, and cessation of flowering.** First, we re-examined a comprehensive dataset of over two million observations made throughout the last 39 years in the Rocky Mountains of Colorado[17,18] to explore changes in the onset, peak, and cessation of flowering. Previous work on this detailed dataset reported discordance in temporal shifts among phenophases[19]. This finding suggests that communities of co-flowering species may be profoundly altered under climate change, with potentially negative consequences for currently co-occurring pollinator and herbivore communities[20]. Here, using our approach that controls for differences in sampling, we find, surprisingly and to the contrary, a close alignment of change through time among these three aspects of flowering phenology in the same data (Figure 2). Because we are able to measure the confidence in our estimates, our approach allows us to overcome implicit sampling biases in observation data. For example, there is both theoretical and empirical evidence that greater sampling effort increases the chances of observing an event earlier[21]. Such sampling biases are difficult to avoid when using the first (or last) observation as a measurement, but can be corrected for when working with a statistical estimator derived from sampling theory, as used here. While it is uncertain whether these results hold elsewhere, the unprecedented degree of sampling in this system urges a re-assessment of this controversial aspect of plant phenology.

**Reconciling historic herbarium and field observations.** Second, we contrast estimates of first flowering derived from herbarium records with a well-studied historical dataset on flowering times from Mas-

4

sachusetts (USA) initiated by Henry David Thoreau in the 1850s. Despite the age and richness of herbarium data, the records are unevenly sampled through time, making direct comparisons between datasets challenging[21]. While there is a strong correlation between rates of change in herbarium and field observations, herbarium records tend to better sample peak flowering, such that recorded dates of first-flower from the two datasets are not directly comparable[22]. As we show in Figure 3, by applying our approach we directly reconcile estimates of first flowering from these two datasets despite differences in sampling: the two datasets not only show correlated changes through time, but how dates of flowering coincide. This is because our approach can use the collection dates of herbarium records to generate a statistical estimate of the onset of flowering, despite having no direct records of the actual onset. This gives hope that our approach can be used to reconcile modern and historical datasets, increasing our power to detect whether current conditions differ from those in the past and so mitigate shifting baseline syndrome[23]. In addition, by leveraging the vast wealth of data in herbaria, our method allows us to dramatically expand the climate space within which we can study plant phenological responses[22], which is currently strongly biased towards northern temperate biomes[24].

**Increased variation in flowering phenology across North America.** Third, we apply our method to phenological observations from the National Phenology Network[25] (NPN), one of the largest citizen-science monitoring schemes, with more than a million records spanning the continental US over the last decade. In parallel with the increasing appreciation and use of collections data, citizen science has emerged as a powerful tool for collecting large amounts of data across broad taxonomic and spatial scales[5]. However, like herbarium records, such data often suffer from poor sampling for rare or difficult to identify events, potentially biasing estimates for those species most at risk from climate change. Because our method requires relatively few samples (see Methods), it is well-suited for such cases. For our analysis, we calculated an estimate of first flowering for each species, in each year and state, with more than five records. As the potential for sampling error in such a broad dataset is high, we used a hierarchical Bayesian approach that allowed us to propagate error clearly throughout every stage of the analysis. Such models are robust to over-parameterisation[26], and so we can model each species with a hierarchically-drawn intercept and slope of change through time.

Our model has two main components: (1) systematic variation in the date of first-flower as a function

of the species, state where it was observed, and year of an observation and (2) estimated variation in the date of first flower (full details are presented in Methods). Our model finds increases in first-flower date of 2.49 days from 2009 to 2015 on average within New York (the state with most data in our model; see Table 1a and Methods), but average rates of change mask significant variation among species (Table 1b). Flowering date is negatively associated with temperature—warmer temperatures result in earlier flowering—however, estimates of the pooling of the overall mean date of first flowering among species and states suggest that, once climate is accounted for, species' flowering dates are relatively invariant among states (see Table 1b). Taken together, these results indicate that species are responding consistently to climate across the continental US.

There are two reasons to be cautious when interpreting the magnitude of these flowering responses to temperature through time. First, we only used data covering the period 2009–2015, and so our model may not capture decadal dynamics of flowering responses. However, our model is consistent with independent data across the period 2001–2008 (Figure 4) whose mean date of first flower is later than that of 2009–2015 (as predicted by our model; $t_{77} = 4.30$, $p < 0.0001$). Second, our model suggests an increase in the variability of the date of first-flower through time (Table 1), which is also visible in Figure 4. This increase in the variability of the date of first-flower through time likely obscures the degree of phenological change we are already experiencing in North America. Conservatively comparing our modelling results for 2011 and 2015, the variation in first-flower date has increased 13% (coefficient of variation ($\frac{\sigma^2}{\mu}$); see Figure 4).

That variability is increasing through time is important as increases in the unpredictability of, and variation in, a system are thought to be indicative of a system approaching a regime shift[27,28]. There is accumulating evidence that species are approaching the limit of their capacity to adapt their phenology to climate change[29–31], and we suggest that our results are consistent with species being pushed to their limits of phenological adaptation. By using a Bayesian approach to model-fitting, we are able to estimate the relative support for our hypothesis, and found that it is twice as likely that the variance is increasing through time than decreasing (on the basis of posterior densities; see Table 1 and Methods). It is possible that the expansion of the NPN scheme through time might have contributed to this pattern. However, we found a similar tendency for increasing variation in the more detailed and consistently sampled Rocky Mountain dataset,

6

with much greater confidence (99.15% probability of increase through time; see Methods). Detecting such an increase in variation through time would be difficult, if not impossible, in studies using space-for-time substitutions or lacking a hierarchical modelling framework such as ours.

## Conclusion

The dual approach we have presented here of accounting for uncertainty around estimates and using a modelling framework that allows uncertainty to percolate through into predictions, allows for a more robust understanding of climate-driven phenological shifts. By drawing information from the sampled distribution of records and not simply the first observation, our approach accurately estimates the timing of first events from sparsely collected data. We show how this has far-reaching consequences for our understanding of flowering phenology, and allows us to marry historic and modern datasets and so vastly increase the temporal and climatic range over which we can study phenological change. Applying our method to one intensively-studied field dataset, and another continental-scale citizen-science dataset, we find tentative evidence for an increase in the variability of phenology through time. Increases in variation may have profound implications for ecosystems, and additional research is urgently needed to examine whether these patterns generalise beyond the North American continental and local-scale botanical systems we present here.
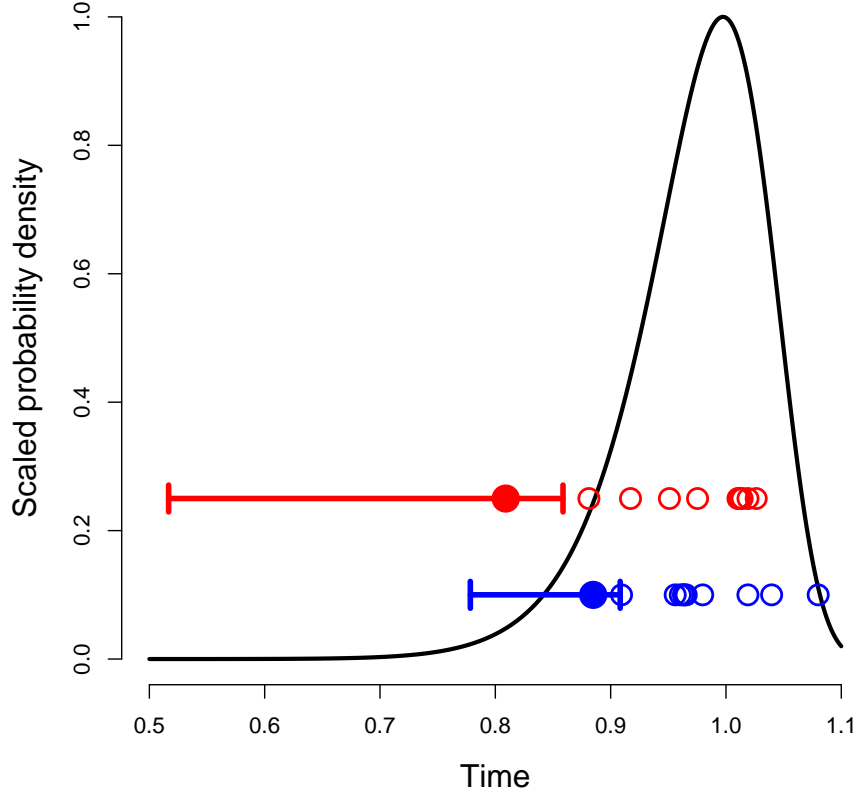
**Figure 1: Example demonstration of the difference between our method and taking first observations at face-value.** Two draws of ten samples (open red and blue circles) from the same Weibull distribution (whose probability density is in black) are shown. Our estimates of the lower limit (start) of the distribution are shown in filled circles, with confidence intervals also shown. Two advantages of this new method are clear in this figure: (1) the estimates have confidence intervals, and (2) the estimates themselves are closer to the true onset of the process (time $0.5$) than the first sample. This results from drawing strength across *all* observations, not simply the single earliest observation. More details and simulations confirming these intuitive properties are given in the Methods.
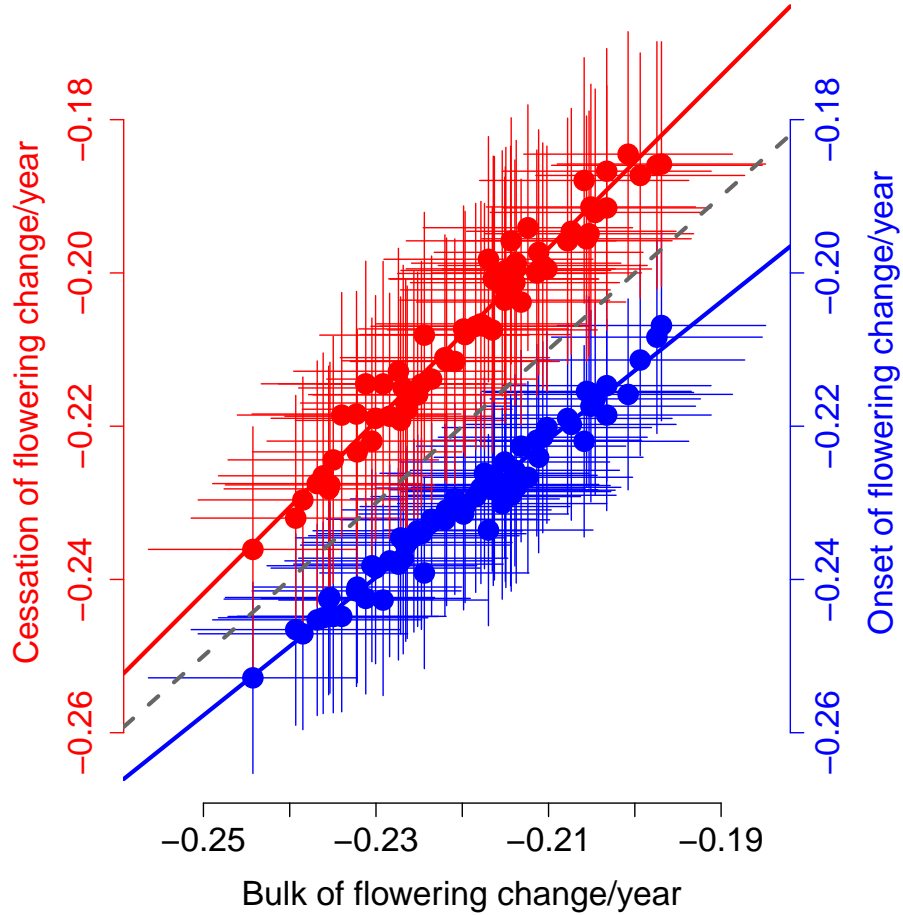
**Figure 2: The rate of change of the onset, bulk, and cessation of flowering through time are tightly correlated in the Rocky Mountain dataset.** This contrasts with a previous analysis not using our approach[19]. Each point represents a species' rate of change (per year) of first (blue) and last (red) flowering, plotted as a function of the change in peak flowering (bottom axis). The coloured lines emanating from each point represent the standard error of each species' change estimate. The thick, solid blue (onset of flowering; slope=0.99, 95% CI 0.90–1.08) and red (cessation of flowering; slope=1.02, 95% CI 0.91–1.13) lines are best-fit lines from a Deming regression accounting for error in both variables; the grey dashed line is a 1:1 line for reference, and is the expectation if the dates of the onset, bulk, and cessation of flowering were changing at the same rate in the data. Species' estimates are taken from an overall model that accounts for species' abundance; each model had an $r^2_{adjusted}$ greater than 74%.
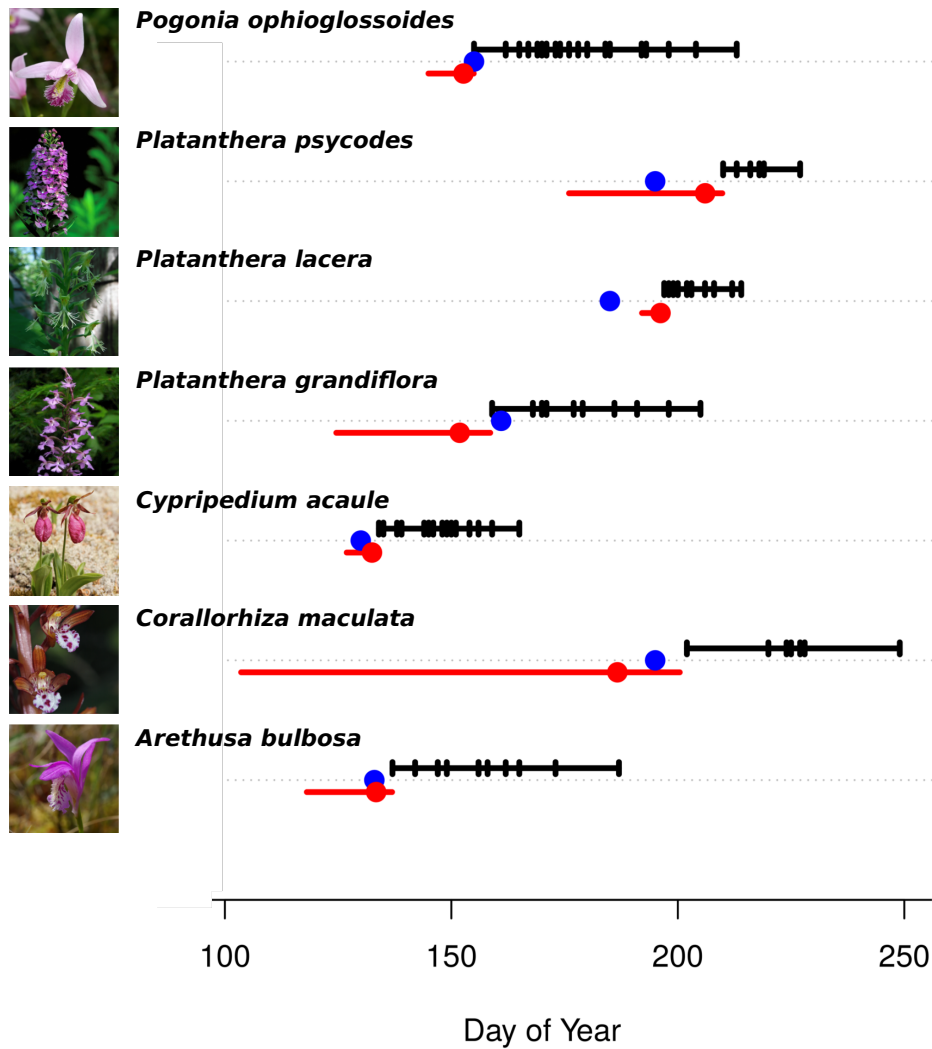
**Figure 3: Reconciling flowering phenology in two historic datasets[21,22].** Data collected between 1858 and 1902 were used, corresponding to the period of greatest overlap between the datasets when Alfred Hosmer and Henry David Thoreau were collecting phenological data. Within the figure, the horizontal black line represents the range of herbarium records (vertical ticks represent each observation), blue circles the earliest field observation, and red closed circles our modelled estimate of onset from the herbarium records (with 95% confidence intervals also in red). Our approach produces estimates that are, on average, almost $4.96$ days closer to the true onset of flowering, as recorded by Thoreau and Hosmer, than the earliest herbarium record (paired test comparing differences between earliest observation and modelled onset: $t_6 = -2.61, p = 0.0399$). We acknowledge that this approach does not account for variation across years, which is mainly driven by annual temperature variation[32,33]. Species with fewer than six herbarium or field observations were excluded from the analyses; see Methods for more details.

|                                      | 0.5%  | 2.5%  | Median | 97.5%  | 99.5%  | SD   |
|--------------------------------------|-------|-------|--------|--------|--------|------|
| Overall mean ($\mu$)                 | 94.80 | 96.99 | 105.03 | 113.58 | 116.33 | 4.27 |
| Yearly change                        | −5.35 | −4.61 | −2.68  | −0.99  | −0.48  | 0.93 |
| Temperature                          | −1.34 | −1.08 | −0.29  | 0.45   | 0.67   | 0.38 |
| Precipitation                        | −0.10 | −0.07 | 0.03   | 0.13   | 0.18   | 0.05 |
| Overall variation                    | 22.92 | 23.30 | 24.48  | 25.86  | 26.30  | 0.66 |
| Yearly variation change ($\sigma^2$) | −0.82 | −0.59 | 0.16   | 0.90   | 1.13   | 0.39 |

**(a)** Selected model coefficients

|                      | Mean | Median | SE    | SD       |
|----------------------|------|--------|-------|----------|
| Species              | 1.00 | 1.00   | 0.013 | 0.001 87 |
| State $\times$ Year  | 0.58 | 0.54   | 0.234 | 0.034 82 |
| State                | 0.12 | 0.13   | 0.052 | 0.006 50 |
| State $\times$ Year  | 0.75 | 0.81   | 0.238 | 0.030 01 |
| $r^2$                | 0.53 | 0.53   | 0.010 | 0.000 18 |

**(b)** Pooling estimates

**Table 1: Modelled estimates of first flowering date in the National Phenology Network (NPN) data.**
See Figure 4 for plots of the model output through time. We modelled the onset of flowering as a function
of species-specific responses and environmental conditions (see Methods). All coefficients are summaries
of Bayesian credible intervals (not frequentist confidence intervals) taken from 3200 samples across 16
Markov Chain Monte Carlo runs with all $n_{eff} > 3000$ and $\hat{R} = 1$ (see Methods for more details and
all model coefficients). (a) shows model coefficients taken from the posterior distribution of the model
(see Methods for all coefficients). The first four rows describe changes in the date of flowering through
time, while the last two rows describe how variation about the average flowering date changes through
time. These provide support for earlier flowering in hotter years and locations, along with more variable
flowering through time. (b) shows estimates of the degree of pooling[34] for species' and states' mean and
change through time ('Year') in the data. Pooling indicates the extent to which estimates at each level
within a multi-level model vary; values close to 0 indicate variation, values close to 1 no variation. Thus
these results suggest that individual species' flowering times varied independently, but that state-level
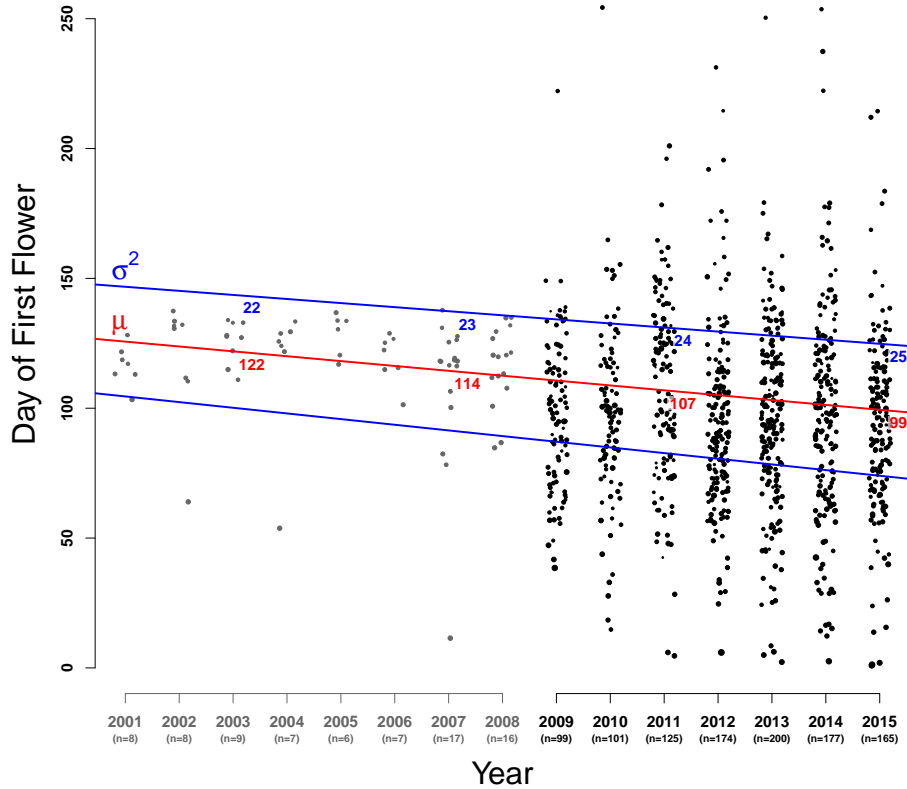effects did not to the same extent.

**Figure 4: Yearly variation in flowering phenology throughout North America in the National Phenology Network (NPN) data.** On the vertical axis, we plot the estimated date of first flower, with point size inversely proportional to the standard error of the estimate. The red line is the average estimate of flowering time through time ($\mu$ in Table 1), while the blue upper and lower lines are the modelled variance of flowering through time ($\sigma^2$ in Table 1). Estimates for particular years are labelled on the graph. The figure shows the trend for earlier flowering through time, as well as indicating the increase in variability of first-flower date through time. We plot data from 2001–2008 that were not used to parameterise the model in grey, to show the predictive power of the model for novel data.

**Author Contributions** WDP, TJD, and CCD conceived of the study; WDP analysed the data. All authors wrote the manuscript and interpreted the results.

**Author information** Reprints and permissions information is available at `www.nature.com/reprints`. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to `will.pearse@usu.edu`.

**Competing interests** The authors declare no competing financial interests

# Methods

All analyses were conducted in `R` version `3.3.2` [35].

## A new approach to estimating the start of a process

Roberts & Solow[13] gave formulae to produce an estimate for the end of a process, and confidence intervals for that estimate. These same formulae can be used to estimate the beginning of a process if the values are sorted ascendingly. In the accompanying supplement, we provide code (`headers.R`) to perform these calculations that re-creates the exact values as reported by Roberts & Solow[13] in their original manuscript. Figure 1 gives a graphical example of the difference between our approach and that of taking the first observation at face value.

We were unable to find an analytical solution for the standard error of the onset or end of events, and so used a parametric bootstrap to estimate its error (code also in `headers.R`). For this we estimated the shape parameter of the joint-Weibull distribution of sighting times, drew 100 samples of the same size as our observed sample from a distribution parameterised by the estimated shape parameter, and calculated the standard deviation of the samples. Note that, as is clear from Figure 1, the confidence intervals generated from this approach are not symmetrical; we therefore caution against the uncritical use of the width of the confidence intervals as an estimate of error of an estimate.

Our approach cannot be used when all observations are made at exactly the same time, or when the first/last onset/cessation observations are exactly identical, so our code removes all such duplicates and issues a warning. When measurements were made on only two or fewer unique dates/times there can be no estimate of onset/end, and so our code returns an '`NA`' value and again issues a warning.

Finally, we note that very large samples of observations are not as informative as might be expected using this method, because the standard Gamma distribution upon which it is based greatly weakens the influence of observations far from the tail of the distribution being estimated. This makes a degree of intuitive sense: when estimating the onset of a process, the end of it has very little information content (and vice-versa). In our experience, the weakening is such that examining more than the earliest/latest 30 observations is

14

<sup>213</sup> unnecessary; the influence of such values is so low that it can go beyond the numerical precision of some

<sup>214</sup> *R* instances and cause errors. In all analyses below, we used a maximum of the 50 earliest observations;

<sup>215</sup> concerned users can alter this using the *k* parameter in our code.

## Our approach vs. the first observation

<sup>216</sup>

<sup>217</sup> To examine our power to detect the true onset of a process, we examined type I error rates: when the

<sup>218</sup> two-tailed 95% confidence intervals of our estimate overlapped the true value of the onset of the process.

<sup>219</sup> Fifty times each, we drew $n$ samples from a uniform distribution ranging from 0 and $m$ across all com-

<sup>220</sup> binations of $n$ and $m$, where $n$ was $4, 5, ..., 49, 50$, and $m$ $20, 21, ..., 349, 350$. We consider these ranges

<sup>221</sup> and sampling regimes to reflect the kinds of phenological data frequently used (*i.e.*, sample size of obser-

<sup>222</sup> vations and the day of the year on which flowering was first observed). For these simulations, 0 was the

<sup>223</sup> true onset of the process: even if a sample was not drawn with a value of 0, that is the statistical limit of

<sup>224</sup> the uniform distribution from which we were sampling. When using 95% confidence intervals ($\alpha_{5\%}$), we

<sup>225</sup> would typically expect an 80% chance of producing confidence intervals that encompass the true value

<sup>226</sup> (*i.e.*, a statistical power—$\beta$—of 80%): we exceeded this expectation in 93% of parameter combinations.

<sup>227</sup> As Supplementary Figure 1 shows, the overwhelming majority of cases where we had poorer power were

<sup>228</sup> when we had fewer than ten samples (the left-hand side of the figure). We thus consider our approach to

<sup>229</sup> have high power.

<sup>230</sup> To contrast our approach with assuming the first observed value as the onset of a process, we also recorded

<sup>231</sup> the least (in our context, earliest) observation while performing the same simulations above. Supplemen-

<sup>232</sup> tary Figure 2 shows the percentage error of the estimate ($\frac{estimate}{range} \times 100$). Note that it is impossible to

<sup>233</sup> perform a direct quantitative comparison of these two approaches: our method produces a statistical es-

<sup>234</sup> timator with an associated degree of error, while the first observation is a single observation for which

<sup>235</sup> there is no meaningful estimate of confidence. The first estimate under-estimates the onset of flowering in

<sup>236</sup> many cases; a log-unit increase or decrease of the range in the sampling results in a log-unit increase in

<sup>237</sup> the percentage error (Supplementary Figure 2). Thus as the duration of a process increases, the amount of

<sup>238</sup> sampling required to accurately estimate the true onset increases. That uncritical use of the first observa-

15

tion is biased is uncontroversial; it is well-known that the first observation of a flower in bloom is strongly affected by sampling effort[21]. Even while keeping the variance of a distribution constant, sampling it more times gives more opportunity for a more extreme event, by chance, to be sampled—the limits of most statistical distributions are infinite. Our approach, which produces a statistical estimator, can account for this, which is not possible when working with the first estimate.

We also note that the accuracy of this method has been empirically verified by Clements *et al.* [36], who examined its ability to accurately detect local extinction under different sampling regimes and experimental conditions.

## Colorado Rocky Mountains—Data

Data are from CaraDonna *et al.* [19], and consist of regular surveys carried out in the Colorado Rocky Mountains (USA); from 1974–2012, thirty square $4m^2$ plots were surveyed, and the number of flowers counted on each individual every two days. Following CaraDonna *et al.*[19], we restricted our analyses to those species for which there were records in at least half of the dataset (19 years). Estimates for each species were calculated for each plot within each year; if such a grouping had fewer than ten measurements we excluded that measurement. We excluded these measurements as we wished to model changes in variability, and we did not want to include less precise estimates which could inflate variation. Our power analyses (see above) suggested that ten samples were sufficient to estimate the onset of a process with reasonable confidence. We included log-transformed abundance as a factor in our analyses.

## Colorado Rocky Mountains—onset vs. peak vs. cessation

The models presented in the Results and Discussion regress onset and cessation of flowering against peak (median) flowering, ignoring variation among species and abundance. To account for these factors following an earlier analysis of this dataset[19], we fitted full linear models incorporating species' identities and their interaction with year, and a separate additive effect of abundance. The model results for the shifts in the onset, peak, and cessation of flowering can be seen in Supplementary Tables 1, 2, and 3 respectively and each model had an $r^2_{adjusted}$ greater than 74%. We then performed Deming regressions of species-

16

level changes in onset (slope=0.99, 95% CI 0.90–1.08) and cessation (slope=1.02, 95% CI 0.91–1.13) of flowering through time as a function of peak flowering. Deming regressions were performed using *deming*[37], and account for error in estimates of change in both predictor and response variables.

## Historical comparisons of phenology

Data were taken from Davis *et al.*[22] and consist of herbarium records and direct field observations from the surroundings of Concord (Massachusetts, USA). These historical collections reflect four main periods of sampling: records collected by Thoreau (1852–1858), Hosmer (1878, 1888–1902), Miller-Rushing & Primack (2003–2006), and Davis & Connolly (2011–2013)[32,38]. We restricted ourselves to only those samples collected before 1903, as this time period overlapped best with the collection of herbarium specimens, and it was the comparison between these two sets of observations that we were most interested in here. The herbarium data themselves were extracted from the Harvard University Herbaria (HUH), New York Botanical Garden's William and Lynda Steere Herbarium (NY), Yale University Herbarium (YU), and University of Connecticut's George Safford Torrey Herbarium (CONN) by Davis *et al.*[22]. A specimen was recorded as flowering if over 75% of its flowers were open (if multiple flowers were present in a specimen); for more details see Davis *et al.*[22]. We analysed species that were common to both datasets and that had (at a minimum) six dated herbarium records. We estimated the onset of flowering and its confidence intervals in these data as described above, and plot the results in Figure 3.

## National Phenology Network—data

Data were downloaded from the National Phenology Network (NPN), including observations from the 1st of January 2001 until the 13th of February 2017 (the date of download); species functional type was set to 'deciduous broadleaf', phenophase category to 'leaves, flowers', and data collected from the continental United States of America. Only events referring to flowers were retained for analysis; specifically, those with 'flower' and 'bloom' (but not 'end' or 'pollen') in their phenophase descriptions. Observations were split according to species, state, and year, and estimates of first-flower (and their standard errors) calculated across these groupings were the basis of analysis. Temperature and precipitation data were taken

from the University of East Anglia's Climatic Research Unit high-resolution gridded historical datasets (v. 3.24.01[39]), and yearly mean values for each state calculated on the basis of state outlines taken from the Global Administrative Areas dataset (v. 2.8; http://www.gadm.org/). Since these temperature data are currently only available from 1901 until 2015, we restricted our analyses to estimates of first flowering between the 1st of January 2005 and the 31st of December 2015.

In the analyses presented in the Results and Discussion, we (conservatively) limited our analyses to species-site-year estimates with at least 5 observations, and excluded species with fewer than ten species-site-year estimates. This provided 1041 observations across a total of 63 species in 45 states, covering the period 2009–2015, and all parameter estimates from these analyses are presented in Supplementary Table 4. Here, we also present results from a model fit to all data from 2009–2015 (1249 observations of 150 species in 46 states), and show that the results are qualitatively identical (Supplementary Table 5). In addition, because the coverage of the data is markedly increased after 2009 (see Figure 4), we fit models to data collected from 2001–2015. Results from 2001–2015 data limited to species-site-year estimates with at least 5 observations and excluding species with fewer than ten species-site-year estimates (1119 observations of 63 species in 45 states) are given in Supplementary Table 6. Results from all data from 2001–2015 (1327 observations of 150 species in 46 states) are given in Supplementary Table 7. All year, temperature, precipitation, longitude (of state centroid), and latitude (also of state centroid) data were scaled to have a mean of zero and standard deviation of one to make model coefficients directly comparable [following 26]. Model coefficients were back-transformed to their original scales in the Results and Discussion, but not in the Supplementary Tables.

## National Phenology Network—Hierarchical modelling

We computed our model using *rstan*[40] in each dataset, running a total of 16 chains for 20,000 iterations, sampling every 50 iterations and discarding the first 10,000 iterations as burn-in. All models were checked graphically for convergence and mixing, and $\hat{r}$ values were all equal to 1. In the Results and Discussion we report that it is twice as likely that the variation in the date of first flower is increasing through time than it is not (*i.e.*, that $\epsilon_\beta > 0$; see below for definitions); we base this upon the observation that 66.67%

of the posterior distribution of $\epsilon_\beta$ was greater than 0.

In Figure 4 we show for reference points from 2001–2008 that were not used to fit models to data. Visual posterior predictive checks were also performed on all model results to ensure model validity. We draw the reader's attention to the greater support for our main result (increased variance through time as measured with the parameter $\epsilon_\beta$; see below) in the model fitted to the longer time series (Supplementary Table 6); we consider it more conservative, and so preferable, to present the more modest coefficients in the main text of the manuscript.

The general structure of our model is described in the Results and Discussion; here we present it more formally. Specifically, the higher-level structure of the model is as follows:

$$DOY \sim N(\alpha_0 + \mu_{spp} + \mu_{env} + \mu_{space} + \mu_{space-time}, \epsilon) \tag{1}$$

Where $DOY$ is the estimated 'Day of Year' of first flower and $\alpha_0$ is the overall first flowering date. For ease of presentation, we have grouped the model parameters together: the terms $\mu_{spp}$ and $\mu_{env}$ describe species' and environmental effects, $\mu_{space}$ and $\mu_{space-time}$ account for spatial and temporal auto-correlation, and $\epsilon$ describes changes in the variance of $DOY$ through time. We describe each below.

Species-specific changes through time, $\mu_{spp}$, is defined as:

$$\mu_{spp} = \alpha_i + \beta_0.Year + \beta_i.Year \tag{2}$$

where $\alpha_i$ is the difference from the overall mean ($\alpha_0$) for each species ($i$), $Year$ is the year of an observation, $\beta_0$ is the slope of the overall change in $DOY$ through time, and $\beta_i$ is the difference in that slope for each species.

Environmental determinants of $DOY$, $\mu_{env}$, is defined as:

$$\mu_{env} = \tau.Temp_{j,k} + \pi.Precip_{j,k} \tag{3}$$

where $\tau$ quantifies the effect of the mean yearly temperature ($Temp_{j,k}$) of an observation's state ($j$) in a given year ($k$), and $\pi$ the effect of the mean yearly precipitation ($Precip_{j,k}$) of an observation's state in a given year.

Each state's residual variation in $DOY$, both overall ($\mu_{space}$) and through time ($\mu_{space-time}$) are expressed similarly. $\mu_{state}$ is defined as:

$$\mu_{state} = \alpha_j + x_\alpha.Long_j + y_\alpha.Lat_j + z_\alpha.Long_j.Lat_j \tag{4}$$

where $\alpha_j$ is the difference from the overall mean ($\mu_0$) for each state, and $x_\alpha$ and $y_\alpha$ measure variation in $DOY$ longitudinally ($Long_j$) and latitudinally ($Lat_j$), respectively. $z_\alpha$ captures the interaction of latitude and longitude. Note that each state's ($j$) latitude and longitude is measured as the centroid of a state, as described above. The influence of each state may also vary through time, as captured in the definition of $\mu_{space-time}$:

$$\mu_{space-time} = Year.(\beta_j + x_\beta.Long_j + y_\beta.Lat_j + z_\beta.Long_j.Lat_j) \tag{5}$$

where $\beta_j$ is the difference from the overall change through time ($\beta_0$) for each state, and $x_\beta$ and $y_\beta$ measure variation in $DOY$ longitudinally and latitudinally through time, respectively. $z_\beta$ captures the interaction of latitude and longitude through time.

Finally, but importantly, the term $\epsilon$ measures the overall variance of $DOY$:

$$\epsilon = \epsilon_0 + \beta_\epsilon.Year \tag{6}$$

where $\epsilon_0$ is the overall variance (error) in our data, and $\beta_\epsilon$ is the change in that variance through time.

The species-specific parameters were drawn from prior distributions centred at 0 with estimated variances. Specifically:

$$\alpha_i \sim Normal(0, \sigma_{\alpha_i}) \tag{7}$$

$$\alpha_j \sim Normal(0, \sigma_{\alpha_j}) \tag{8}$$

$$\beta_i \sim Normal(0, \sigma_{\beta_i}) \tag{9}$$

$$\beta_j \sim Normal(0, \sigma_{\beta_j}) \tag{10}$$

Other parameters were given Normal priors with wide distributions so as to be uninformative, specifically:

$$\alpha_0, \beta_o, x, y, z, x_\beta, y_\beta, z_\beta \sim Normal(0, 1000) \tag{11}$$

With the exception of the variance parameters, for which our priors were:

$$\epsilon_0, \sigma_{\alpha_i}, \sigma_{\alpha_j}, \sigma_{\beta_i}, \sigma_{\beta_j} \sim Uniform(0.0001, Infinity) \tag{12}$$

$$\epsilon_\beta \sim Uniform(-10, 10) \tag{13}$$

## Colorado Rocky Mountains—hierarchical modelling

Within the Results and Discussion, we refer to a hierarchical model of the onset of species' flowering times in the Rocky Mountain dataset, which we describe here in full.

We computed our model using *rstan*[40] in each dataset, running a total of 16 chains for 20,000 iterations, sampling every 50 iterations and discarding the first 10,000 iterations as burn-in. All models were checked graphically for convergence and mixing, and $\hat{r}$ values were all equal to 1.

The structure of our model, which is comparable to that of the NPN model above, is as follows:

$$DOY \sim N(\alpha_i + \beta_i.Year + \gamma.Abundance, \epsilon_0 + \epsilon_\beta.Year) \tag{14}$$

21

Where $DOY$ is the estimated 'Day of Year' of first flower, $\alpha_i$ the mean $DOY$ for each species ($i$), $\beta_i$ the slope of yearly change of $DOY$ for each species, $\gamma$ a slope accounting for abundance-driven changes, $\epsilon_0$ the mean variance of $DOY$, and $\epsilon_\beta$ the rate of change of variance through time. $Year$ and $Abundance$ represent the recorded year and abundance of species within each plot, respectively. These terms are similar to those used for the NPN model (described above).

$\alpha_i$ and $\beta_i$ are species-specific parameters, and are drawn from distributions parameterised as follows:

$$\alpha_i \sim Normal(\alpha_0, \sigma_{\alpha_i}) \tag{15}$$

$$\beta_i \sim Normal(\beta_0, \sigma_{\beta_i}) \tag{16}$$

Most parameters were given Normal priors with wide distributions so as to be uninformative, specifically:

$$\alpha_0, \beta_0, \gamma_0, \epsilon_0, \epsilon_\beta \sim Normal(0, 1000) \tag{17}$$

The only exceptions to this were our hyper-parameters of variance, for which such priors would be inappropriate (negative variances are impossible). Our hyper-parameter priors were:

$$\sigma_{\alpha_i}, \sigma_{\beta_i}, \epsilon_0 \sim Uniform(0.0001, Infinity) \tag{18}$$

All parameter estimates from this model are given in Supplementary Table 8. In the manuscript we refer to evidence that the variance in the onset of flowering in the Rocky Mountain dataset has been increasing through time: this is supported by the estimates of $\epsilon_\beta$ in Supplementary Table 8, whose high-credibility intervals (and standard errors and deviations) suggest a positive (non-zero) change through time. In the Results and Discussion we report a 99.15% probability that the variation in the date of first flower is increasing through time (*i.e.*, that $\epsilon_\beta > 0$); we base this upon the observation that 99.15% of the posterior distribution of $\epsilon_\beta$ was greater than 0.

## Variation among species in flowering time

There is growing evidence that early-flowering species are changing their phenology more strongly in response to climate change. One of the advantages of our hierarchical approach is it permits the examination of variation among species' responses, while propagating uncertainty for each species' response through into the final analysis. In Supplementary Figures 3 and 4, we plot the species-level changes in flowering phenology through time as a function of overall first-flower data for both the Rocky Mountain and NPN data, respectively. We provide these data as a test of the overall validity of our approach, and note that the Rocky Mountain data show some support for two kinds of flowering regime (early vs. late).

## Data availability

All the data we have analysed are publicly available at the references we provide above. The Colorado data are archived through the Open Science Framework at `https://osf.io/jt4n5/`

# Bibliography

[1]   IPCC. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, 2014.

[2]   A. Menzel *et al.* European phenological response to climate change matches the warming pattern. *Global Change Biol.* 12.10 (2006), 1969–1976.

[3]   C. Parmesan. Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.* (2006), 637–669.

[4]   E. M. Wolkovich *et al.* Warming experiments underpredict plant phenological responses to climate change. *Nature* 485.7399 (2012), 494–497.

[5]   J. Silvertown. A new dawn for citizen science. *Trends Ecol. Evol.* 24.9 (2009), 467–471.

[6]   G. H. Pyke & P. R. Ehrlich. Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological Reviews* 85.2 (2010), 247–266.

[7]   K. M. Robbirt, A. J. Davy, M. J. Hutchings, & D. L. Roberts. Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid Ophrys sphegodes. *Journal of Ecology* 99.1 (2011), 235–241.

[8]   R. C. Baird. Leveraging the fullest potential of scientific collections through digitisation. *Biodiversity Informatics* 7.2 (2010), 130–136.

[9]   R. S. Beaman & N. Cellinese. Mass digitization of scientific collections: new opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys* 209 (2012), 7–17.

[10]  M. Balke *et al.* Biodiversity into your hands: a call for a virtual global natural history 'metacollection'. *Front. Zool.* 10.55 (2013).

[11] C. G. Willis, E. Law, *et al.* CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytologist* 215.1 (2017). 2016-23274, 479–488.

[12] C. Meyer, P. Weigelt, & H. Kreft. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* 19 (8 2016), 992–1006.

[13] D. L. Roberts & A. R. Solow. Flightless birds: when did the dodo become extinct? *Nature* 426.6964 (2003), 245–245.

[14] R. Ruggles & H. Brodie. An empirical approach to economic intelligence in World War II. *JASA* 42.237 (1947), 72–91.

[15] P. Cooke. Optimal linear estimation of bounds of random variables. *Biometrika* 67.1 (1980), 257–258.

[16] I. Weissman. Confidence intervals for the threshold parameter. *Commun. Stat. Theory Methods* 10.6 (1981), 549–557.

[17] D. W. Inouye. Effects of climate change on phenology, frost damage, and floral abundance of montane wildflowers. *Ecology* 89.2 (2008), 353–362.

[18] G. Aldridge, D. W. Inouye, J. R. Forrest, W. A. Barr, & A. J. Miller-Rushing. Emergence of a mid-season period of low floral resources in a montane meadow ecosystem associated with climate change. *J. Ecol.* 99.4 (2011), 905–913.

[19] P. J. CaraDonna, A. M. Iler, & D. W. Inouye. Shifts in flowering phenology reshape a subalpine plant community. *Proc. Natl. Acad. Sci. U.S.A.* 111.13 (2014), 4916–4921.

[20] M. E. Visser & C. Both. Shifts in phenology due to global climate change: the need for a yardstick. *Proc. R. Soc. Lond. B* 272.1581 (2005), 2561–2569.

[21] A. J. Miller-Rushing, D. W. Inouye, & R. B. Primack. How well do first flowering dates measure plant responses to climate change? The effects of population size and sampling frequency. *J. Ecol.* 96.6 (2008), 1289–1296.

[22] C. C. Davis, C. G. Willis, B. Connolly, C. Kelly, & A. M. Ellison. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *Am. J. Bot.* 102.10 (2015), 1599–1609.

[23] S. Papworth, J. Rist, L. Coad, & E. Milner-Gulland. Evidence for shifting baseline syndrome in conservation. *Conserv. Lett.* 2.2 (2009), 93–100.

[24] B. I. Cook, E. M. Wolkovich, T. J. Davies, *et al.* Sensitivity of spring phenology to warming across temporal and spatial climate gradients in two independent databases. *Ecosystems* 15.8 (2012), 1283–1294.

[25] *https://www.usanpn.org/*. Accessed on the 1st of June 2017.

[26] A. Gelman, J. B. Carlin, H. S. Stern, & D. B. Rubin. Bayesian Data Analysis. 3rd edition. Taylor & Francis, 2014.

[27] M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, & B. Walker. Catastrophic shifts in ecosystems. *Nature* 413.6856 (2001), 591–596.

[28] M. Scheffer, J. Bascompte, *et al.* Early-warning signals for critical transitions. *Nature* 461.7260 (2009), 53–59.

[29] C. Körner & D. Basler. Phenology under global warming. *Science* 327.5972 (2010), 1461–1462.

[30] B. I. Cook, E. M. Wolkovich, & C. Parmesan. Divergent responses to spring and winter warming drive community level flowering trends. *Proceedings of the National Academy of Sciences* 109.23 (2012), 9000–9005.

[31] S. J. Brooks, A. Self, F. Toloni, & T. Sparks. Natural history museum collections provide information on phenological change in British butterflies since the late-nineteenth century. *Int. J. Biometeorol.* 58.8 (2014), 1749–1758.

[32] A. J. Miller-Rushing & R. B. Primack. Global warming and flowering times in Thoreau's Concord: a community perspective. *Ecology* 89.2 (2008), 332–341.

[33] E. R. Ellwood, S. A. Temple, R. B. Primack, N. L. Bradley, & C. C. Davis. Record-breaking early flowering in the eastern United States. *PloS One* 8.1 (2013), e53788.

[34] A. Gelman & I. Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics* 48.2 (2006), 241–251.

[35] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2016.

[36] C. F. Clements *et al.* Experimentally testing the accuracy of an extinction estimator: Solow's optimal linear estimation model. *Journal of Animal Ecology* 82.2 (2013), 345–354.

[37] T. Therneau. deming: Deming, Thiel-Sen and Passing-Bablock Regression. R package version 1.0-1. 2014.

[38] C. G. Willis, B. Ruhfel, R. B. Primack, A. J. Miller-Rushing, & C. C. Davis. Phylogenetic patterns of species loss in Thoreau's woods are driven by climate change. *Proc. Natl. Acad. Sci. U.S.A.* 105.44 (2008), 17029.

[39] I. Harris, P. Jones, T. Osborn, & D. Lister. Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. *Int. J. Climatol.* 34.3 (2014), 623–642.

[40] B. Carpenter *et al.* Stan: A Probabilistic Programming Language. *J. Stat. Softw.* 76.1 (2017), 1–32.