Received: 6 December 2021 | Accepted: 22 August 2022

DOI: 10.1111/1365-2435.14173

Functional Ecology | BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Leveraging Natural History Collections to Understand the Impacts of Global Change

Phenological research based on natural history collections: Practical guidelines and a lepidopteran case study

Michael W. Belitz[1] | Elise A. Larsen[2] | Vaughn Shirey[2] | Daijiang Li[3,4] | Robert P. Guralnick[1]

[1]Florida Museum of Natural History, University of Florida, Gainesville, Florida, USA

[2]Department of Biology, Georgetown University, Washington, District of Columbia, USA

[3]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA

[4]Center for Computation & Technology, Louisiana State University, Baton Rouge, Louisiana, USA

Correspondence
Michael W. Belitz
Email: mbelitz@ufl.edu

Funding information
National Science Foundation, Grant/Award Number: DEB # 1702664 and DEB # 1703048

Handling Editor: Madhav P. Thakur

Abstract

1. Natural history collections (NHCs) have been indispensable to understanding longer-term trends of the timing of seasonal events. Massive-scale digitization of specimens promises to further enable phenological research, especially the ability to move towards a deeper understanding of drivers of change and how trait–environment interactions shape phenological sensitivity.
2. Despite the promise of NHCs to answer fundamental phenology questions, the use of these data resources presents unique and often overlooked challenges requiring specialized workflow steps, such as assembling multisource data, accounting for date imprecision and making decisions about trade-offs between data density and spatial resolution.
3. We provide a set of key best practice recommendations and showcase these via a case study that utilizes NHC data to test hypotheses about spatiotemporal trends in adult Lepidoptera (i.e. butterflies and moths) flight timing across North America. Our case study is a worked example of these best practices, helping practitioners recognize and overcome potential pitfalls at each step, from data acquisition and cleaning, to delineating spatial units and proper estimation of phenological metrics and associated uncertainty, to building appropriate models.
4. We confirm and extend the critical importance of voltinism and diapause strategy, but less-so daily activity patterns, for predicting Lepidoptera phenology spatiotemporal trends. Our case study also showcases the unique power of NHC data to test existing hypotheses and generate new insights about temporal phenological trends. Specifically, migratory species and species that enter diapause as adults are advancing the start of flight periods in more recent years, even after accounting for climate context. These results highlight the physiological and adaptive differences between species with different overwintering strategies.
5. We close by noting the value of partnerships between data scientists, museum experts and ecological modellers to fully harness the power of digital data resources to address pressing global change challenges. These partnerships can

[Correction added on 30 September 2022, after first online publication: the Special Feature title has been corrected.]

extend approaches for integrating multiple data types to fully unlock our understanding of the tempo, mode, drivers and outcomes of phenological changes at greater spatial, temporal and taxonomic scales.

KEYWORDS
best practices, climate change, ecological traits, interactive effects, museum, seasonal timing

1 | INTRODUCTION

Natural history collections (NHCs) have been instrumental in understanding temporal trends in changes to the timing of seasonal events. In particular, herbarium specimens have been reliably used to characterize phenological responses to changing climate (reviewed in Willis et al., 2017). More recently, researchers are attempting to move beyond simply documenting changes in phenology and are examining more complex phenological responses such as determining phenological cueing mechanisms (Davis et al., 2015; Park & Mazer, 2018) and determining if changes in bird migration phenology are related to changes in body size (Zimova et al., 2021). Massive-scale digitization efforts, including specimen imaging, while still mostly incomplete (Cobb et al., 2019), promises to further enable macrophenological research (sensu Gallinat et al., 2021) across multiple branches of the tree of life (Soltis, 2017). However, challenges still abound in proper use of NHCs in phenological research. A myriad of sampling biases and data gaps, spanning geographic, temporal, taxonomic and trait dimensions often complicate their proper use (Meineke & Daru, 2021). Dealing with these issues requires data exploration and filtering, whose customization and testing differ depending on the taxonomic group, study extent and data sources (Zizka et al., 2019). Even after data cleaning, phenological estimators and downstream modelling steps that fail to account for incidental sampling biases across space and time can produce biased results (Isaac et al., 2014; Larsen & Shirey, 2021).

Unlike species distribution modelling, which has an extensive literature documenting best practices for use of NHC data (e.g. Chauvier et al., 2021; Elith et al., 2006; Mateo et al., 2010; Merow et al., 2013; Myers et al., 2015; Pearce & Boyce, 2006), similar efforts for phenological practices have been lacking. Previous efforts have been made to make phenology estimates more accurate (e.g. Pearse et al., 2017; Pearson, 2019), but guidelines to assist researchers along an entire project are scant. Here, we detail a workflow to support best practices in using digital NHC data in phenological research. We identify and discuss areas where practitioners may face potential pitfalls along the entire workflow, from data acquisition and cleaning, to delineating spatial units and proper estimation of phenological metrics and associated uncertainty, to building appropriate models that can test hypotheses about spatiotemporal phenological variation and underlying drivers. Our case study focuses on Lepidoptera (i.e. butterflies and moths), but throughout, we broadly discuss cross-cutting issues, including challenges with reconciling taxonomy, capturing collector information needed to understand observation effort, dealing with imprecise and missing data, and accounting for spatiotemporal sampling bias and phylogenetic relationships.

While our intent is to provide a set of best practice recommendations, we also address key questions about temporal trends in phenology for Lepidoptera across North America. Lepidoptera are particularly well-suited for phenology studies, given that they are well collected, highly diverse and have temperature-dependent developmental rates (Buckley et al., 2017). Structured monitoring networks such as the UK Butterfly monitoring scheme and the Rothamsted Insect Survey have been critical resources for documenting temporal trends in phenology (e.g. Macgregor et al., 2019; Roy & Sparks, 2000), but these are limited in spatial scope. Incidental citizen science efforts, such as iNaturalist, can increase the breadth of analyses (Barve et al., 2020), but data are only available for more recent years and have potential biases that differ from NHC data. Therefore, the key means to examine patterns and drivers of Lepidoptera phenology at broad spatial, temporal and taxonomic scales will be careful use of NHCs.

Natural history collection-based multi-species phenology studies of especially butterflies (e.g. Brooks et al., 2014, 2017; Kharouba et al., 2014) have showcased the utility of examining longer-term phenology trends in the face of global change, and the importance of trait-based approaches. In North America, key studies using NHCs have found overall evidence of advancing phenology for a subset of Lepidoptera species (e.g. Kharouba et al., 2014; Maurer et al., 2018). Traits mediate responsiveness, and species flying earlier in the season showed the strongest responses to temperature (Kharouba et al., 2014 for Canadian butterflies) or shifted phenology most over time (Maurer et al., 2018 for macromoths in the Pacific Northwest). These previous studies both focused on median flight periods, missing both onset or termination dynamics, which likely shift independently and have different environmental cues, given the complex shapes seasonal phenology curves may take for insects (Duchenne et al., 2020).

Here, we examine spatiotemporal trends in Lepidoptera phenology, focusing not only on midpoint dynamics, but also on onset, termination and adult flight duration. In particular, we explore whether voltinism and winter diapause, two key traits known to impact phenology sensitivities, drive differential onset and termination dynamics over time. We account for key climate and spatial contexts, and thus determine whether these temporal trends vary spatially and environmentally. Finally, we examine whether diurnal (mainly butterflies) and nocturnal species show similar trends in their phenological responses over time, given that night-time global temperature is increasing faster than daytime temperature (Fu et al., 2016). At

all points from conceptualizing the design of the study to interpretation, we highlight challenges, pitfalls and best practices towards deriving robust results and inferences.

2 | MATERIALS AND METHODS

2.1 | Downloading digital data and removing duplicate records

Research using NHC data should strive to include records from all key repositories because they may each contain unique data. While the Global Biodiversity Information Facility (GBIF) is often exclusively used, not all digitized data, especially in North America, are also published globally. Therefore, we downloaded all Lepidoptera data from GBIF (2021) as well as iDigBio and Symbiota Collections of Arthropods Network (SCAN) which are the other main data aggregators of digitized specimen data for museums based in the United States. We only downloaded records that were identified by the Darwin Core term 'PreservedSpecimen' to ensure human observations were not included in our dataset. In total, we had a dataset of 6,691,163 records (Belitz, 2022a).

After aggregating records across repositories, we next determined sets of records which shared a unique collector, date of collection and location, removing all but one for downstream analysis. Removing duplicate records is essential for proper phenometric fitting, since phenometric estimators are sensitive to record densities. Furthermore, duplicates artificially inflate sampling intensity. Even if only using one source (e.g. GBIF), duplicate records are not uncommon and should be filtered. For this case study, the bulk of our unique records that could not be downloaded from other data sources were from GBIF, although SCAN also contributed a significant number of unique records (Figure 1a,b).

2.2 | Further data cleaning

Digitized museum records downloaded from biodiversity discovery platforms such as GBIF or iDigBio must be further vetted and cleaned, post-deduplication, before being used in phenological research. Since our analysis focused on adult Lepidopteran phenology, we removed records that were of specimens of non-adult life stages such as egg, larvae or pupae. Many records did not note the life stage of the record, but in the case of Lepidoptera, the vast majority (>99%) of NHC specimens are adults, based on examination of images available on GBIF, and thus records were kept that did not note life stage. Next, we removed records with values missing for the date of collection and geographical coordinates, that is, latitude

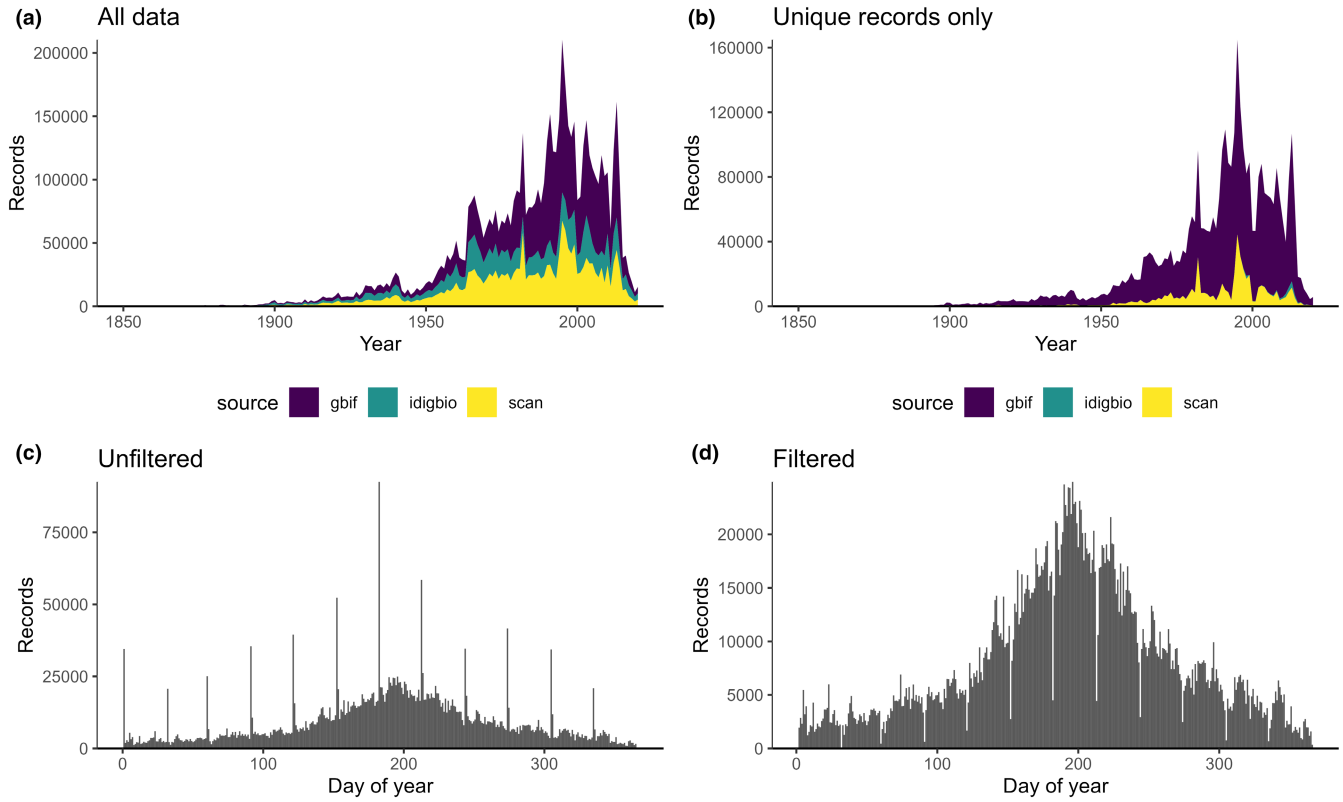

FIGURE 1 The distribution of records over years (panels a and b) and intra-annually (panels c and d). Panel a shows the total volume of records available from different digitized natural history collection (NHC) record platforms (i.e. GBIF, iDigBio and SCAN) and panel b shows those distributions after cleaning duplicates. Panel c shows a clear spike of records on the first of each month, representing date imprecision in NHC data, and panel d shows data distribution after removing specimen collected on the first day of the month (note first of month day of year differ in leap years)

and longitude value, since these records are unusable downstream. Still, many records have imprecise dates that only list the month or year that the specimen was collected. These records are often given default 'eventDate' values in databases that correspond to the first of the month or year that the specimen was collected, leading to far more observations on the first of the month (and year) than would be expected by chance (Figure 1c). We dealt with this bias by removing all records that were noted as collected as the first day of every month (Figure 1d). One could determine whether a specimen was truly collected on the first day of a month by examining the 'verbatimEventDate' of the record or, if an image is available, the original label of the specimen. However, this would involve significant manual vetting of records. When possible, it is best to avoid manual vetting as it is time-consuming, error prone and difficult to reproduce (Zizka et al., 2019). Depending on the spatial resolution of the analysis, it may also be necessary to filter points with imprecise geographical coordinates. However, our analysis is conducted at coarse resolution (see below), limiting any concern about georeferencing precision.

We generated an initial species list by first gridding our study area into 250×250-km equal area cells using the North America Albers Equal Area Conic projection and used a spatial join to assign specimens to grid cells. Our choice of this coarse spatial grain reflects a series of compromises related to data availability and spatial localization of phenological response. We discuss this choice in more detail below, but critical here is that our interest is in broad-scale, continental patterns across wide environmental gradients, rather than local-scale phenomena.

Taxonomic harmonization is an important step in working with NHCs, as taxonomic revisions are common, and names must be harmonized to avoid pseudoreplicated species in our final taxonomic list. Our initial species list included those that had at least three cell by year combinations, with at least three distinct collectors, four distinct days and five observations. For each of these species, we used the package taxotools (Barve, 2020) to report an accepted name and all associated synonyms. Using this set of accepted names and synonyms, we generated a species list of accepted names and aggregated all records of synonyms to these accepted names.

2.3 | Fitting phenometrics

Ensuring phenometrics are estimated using a sufficient number of records is necessary for deriving meaningful phenology insights; too few records and phenometrics are highly biased and uncertain (Belitz et al., 2020). However, it may not be sufficient to simply count the total number of collected records used to estimate a phenometric. Rather, two other metrics—the total number of independent collectors and number of distinct days of observations—may be more relevant. These metrics better assess intra-annual sampling density from independent sources. Appropriate thresholds may vary depending on the context of a study. Here, we applied different requirements for univoltine and multivoltine species. For univoltine species, we fit phenometrics for species–cell–year combinations that had at least five observations, four distinct days of collecting and three distinct collectors. Three days with observations have been demonstrated to produce useful phenoestimates of unimodal species using survey data (Edwards & Crone, 2021). However, multimodal species may have longer durations and more complex seasonal abundances, making their phenometrics harder to estimate (Belitz et al., 2020). Therefore, the data requirements were increased for species that were not obligate univoltine species. Specifically, at least 10 total observations, 8 distinct collecting days and 3 unique collectors were needed to retain multivoltine phenometrics. Simulation experiments have shown phenometrics estimated from 10 total observations can produce useful phenometrics, although more records typically lead to more robust phenometrics (Belitz et al., 2020).

Estimating the start and end of phenological events is challenging given that there are fewer data to parameterize these bounds (Pearse et al., 2017). We choose to estimate quantiles near the bounds of a distribution, since these have been demonstrated to be more accurate and less biased estimates of phenology (Belitz et al., 2020). Phenometrics and their associated 95% confidence intervals for cell-by-year-by-species combinations with enough data were estimated using 0.05, 0.5 and 0.95 continuous sample quantiles using the quantile_ci() function within the *phenesse* R package (Belitz et al., 2020). These quantiles were used to represent approximations of onset, midpoint and termination of adult flight periods. Flight period duration was calculated as the difference between termination and onset values.

Despite taking steps to ensure phenology estimates were being made using sufficient data and methods, estimated phenometrics still occur outside of the range of expected values. Outlier estimates should be examined at a species-specific level. Doing so requires first attempting to account for expected spatial variation in estimates to make phenoestimates more comparable. For example, a phenoestimate of a particular species found in a northerly cell is expected to be different than a phenoestimate in a southerly cell. We accounted for geographical differences in our outlier detection protocol by taking a data-driven approach, where we fit a simple linear mixed model using estimated onset values as our response variable and mean annual temperature (see below for temperature data information) as the fixed effect predictor variable. Random intercepts were fit for each species and cell to account for differences across species and location, and we also fit random slopes for each species to allow each species to have a different response to temperature. Model residuals were calculated and phenoestimates with residuals greater than three times the standard deviation of the overall model residuals were deemed outliers and removed from further analyses (Figure 2).

2.4 | Gathering trait data

We collected trait information from literature and web sources for the 480 species that formed our initial species list to be included as

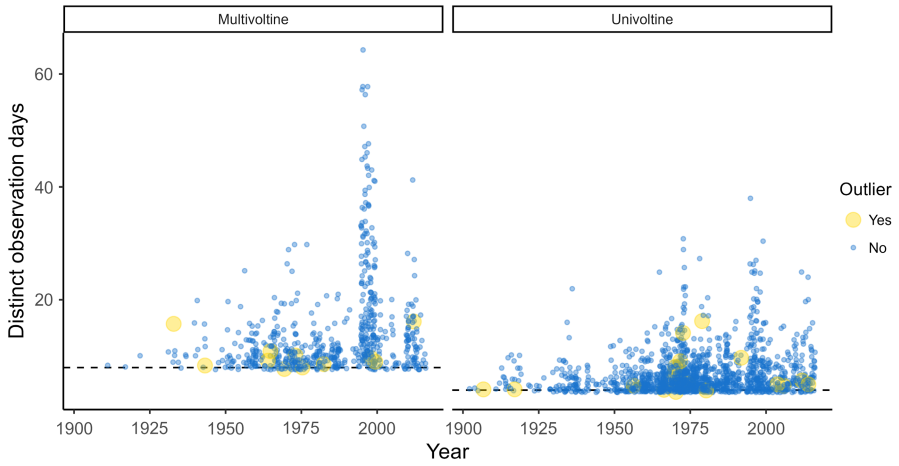

FIGURE 2 Phenological estimates flagged as an outlier and removed. The horizontal line shows the minimum number of distinct observation days for fitting phenometrics per voltinism trait (four observation days for univoltine and eight for multivoltine). Points are jittered to show data density

categorical predictor variables in statistical analyses. We focused on gathering three traits thought to be relevant to Lepidoptera phenology: (1) diapause stage (i.e. the life stage at which diapause or overwintering occurs), (2) voltinism and (3) diurnality. Diapause stage, which is known to affect the phenology of insects (Diamond et al., 2011; Roy et al., 2015), was categorized as egg, larva, pupa, adult, none or migratory, and species that diapause at multiple life stages or do not diapause were removed from further analyses. Voltinism was categorized as either obligate univoltine, facultative or multivoltine, with species that take more than 1 year to develop categorized as univoltine. Diurnality was categorized as diurnal, nocturnal or both, depending on when the adults are active. Sources that were particularly beneficial to gathering these traits were the Butterflies and Moths of North America database (sources used to develop the database include; Covell, 2005; Opler, 1999; Powell & Opler, 2009) and The Caterpillars of North America (Wagner, 2010). Many species that met minimum requirements for generating phenometrics were dropped from further analyses because of lack of available trait data. Given that species traits regulate insect phenological responses (Belitz et al., 2021; Diamond et al., 2011), modelling frameworks addressing interspecific variation in temperature sensitivity must incorporate trait data (Kharouba et al., 2018). We recognize the challenges of assembling complete trait data and therefore emphasize the importance of choosing traits that are documentable and are expected to impact phenology. Restricting the number of traits in an analysis will limit the challenges due to incomplete trait matrices.

We also generated a categorical trait for seasonal flight timing since early versus late flying species can show different phenological responses to climate change (Kharouba et al., 2014; Maurer et al., 2018). To generate this trait, we classified the seasonality of Lepidoptera species by relativizing estimated onset values in a particular cell to the beginning date of the frost free period (averaged from 1991–2020; Wang et al., 2016) of a given cell. The number of days a species began flying after the start of the frost free period was averaged across cells and years. These values were then used to determine the 15th and 85th percentile values, which were used to assign a species as a 'spring' (0–15), 'summer' (15–85) or 'fall' (85–100) species.

2.5 | Gathering climate data

Predictor variables must reflect appropriate spatial and temporal scales given biological questions of interest. We generated four climate variables to include in our modelling framework: annual temperature, annual precipitation, temperature seasonality and precipitation seasonality. We did so by obtaining monthly estimated maximum temperature and cumulative precipitation values for 1901–2016 at approximately 1-km spatial resolution from the Chelsea data product (Karger & Zimmermann, 2018). These data were used to calculate mean annual temperature, annual precipitation, temperature seasonality and precipitation seasonality values for every grid in our analysis from 1901 to 2016. Temperature seasonality was calculated as the standard deviation of the monthly maximum temperatures across a year and precipitation seasonality values were the coefficient of variation of monthly precipitation values across a year.

2.6 | Determining phenological trends in the context of climate and traits

Careful modelling techniques that can account for biases and autocorrelation are needed to ensure that robust conclusions can be drawn from phenological studies using NHCs. Simple analyses using incidental data sources have been found to produce biased estimates or have low statistical power (Isaac et al., 2014). We used linear mixed models (LMMs) to examine the effects of climate, traits and year covariates on the onset, midpoint, termination and duration of adult Lepidoptera across the study area. LMMs were fit using the R package lme4 (Bates et al., 2015). Species-specific phenoestimates of onset, midpoint, termination and duration were the response variables of our models, while climate variables, traits and year were fixed predictor variables. We also included the number of distinct days and number of distinct collectors as fixed effects in our global model as a proxy of observation effort. Two-way interactions chosen based on previous biological knowledge were also tested to examine (1) whether the effects of temperature change along a precipitation gradient, (2) whether temporal effects are more prominent in warm

or wet regions and (3) whether climate and temporal effects are conditioned by life-history traits. Variables were scaled to a mean of zero and a standard deviation of one to make the relative influence of predictor variables easy to interpret. We included grid cell identity as a random intercept in our model to control for uneven sampling in space and other unmeasured environmental variables (Isaac et al., 2014). Species identity was also included as a random intercept to account for unmeasured traits.

Model selection is essential in macrophenology since manipulative experiments cannot be conducted at relevant temporal and spatial scales. To select our top models, we used three modelling approaches. First, we fit a global model with all variables and interactions together in a single weighted mixed effects model, where model inputs were weighted by the inverse of the phenometric confidence interval size. We used the step function from the R package *lmerTest* (Kuznetsova et al., 2017) to reduce model complexity via backward selection. Each top model was evaluated for potentially problematic collinearity by calculating variance inflation factors (VIFs). If variables had VIFs greater than five, one variable was removed based on which variable had smaller effect sizes. The backward selection process was rerun until all VIFs were below five, which has been suggested as a useful value for detecting multicollinearity potentially harmful for inference (Neter et al., 1996). The second modelling approach mirrored the first approach, except that we did not weigh model inputs. The third approach was also a non-weighted mixed model but differed in that we first fit a model with only climate variables and the two-way interaction between temperature and precipitation. Then, after completing our backward selection process and finding the top climate model, we introduced the trait and year covariates to the global model. Backward model selection was performed as described above to reduce model variables and select a top model. To select our overall best model, we choose the model with the lowest Akaike information criterion (AIC) (Anderson & Burnham, 2004) value among our three candidate top models. Selecting models linking weather to biological responses remains a challenging and unsettled topic (Tredennick et al., 2021). Our decision to use these three modelling approaches was that it generated competing top models following multiple approaches used in the literature (Belitz et al., 2021; Li et al., 2021) and encapsulates best practices for ecological modelling with the goal of data exploration (Tredennick et al., 2021).

Modelling frameworks incorporating multiple species in a single analysis can lead to erroneous biological conclusions if the phylogenetic relationship among the species is ignored by inflating the chances of type I errors (Li & Ives, 2017). We converted our top LMMs to phylogenetic linear mixed models (PLMMs) by incorporating a covariance matrix containing the phylogenetic distances between the species as a random intercept term. To build our phylogenetic covariance matrix, we generated a subtree from the Open Tree of Life for the species in our analysis (Michonneau et al., 2016). The divergence time of each internal node was estimated by searching the TimeTree of Life database (Kumar et al., 2017), and the branch lengths were scaled from these ages using the ph_bladj function from the R package *phylocomr* (Ooms & Chamberlain, 2019). The R package *phyr* (Li et al., 2020) was used to fit our top LMMs as PLMMs using this phylogeny and a Bayesian framework with default uninformative INLA priors (Rue et al., 2009). Our results differed between the PLMM and LMM (Supporting Information 2), and therefore, we present the results based on PLMM in the main text. We calculated pseudo-R^2 values to measure the goodness of fit of our top PLMM by calculating the variance of the difference between the observed and predicted values of our fitted PLMM using the following equation, partial $R^2 = 1 - \text{var}(y - \text{y.fitted.f})/\text{var}(y - \text{y.fitted.r})$, where y is the observed data, and y.fitted.f and y.fitted.r are the predicted values from the full and reduced models, respectively (Ives, 2019; Ives & Li, 2018). Spatial autocorrelation in model residuals can also influence model effect sizes and increase type I error rates. We examined the degree of autocorrelation in our top PLMM residuals by calculating Moran's *I* values across different spatial lags and did not detect significant Moran's *I* values across any spatial lags (Figure S1).

3 | RESULTS

Data filtering to ensure sufficient data were used to generate phenoestimates reduced both the number of records in our dataset and the number of species that could be used in our analysis. Taxonomic harmonization was a critical first step, leading to nearly 20% of reported names (95 species names total) being lumped into other accepted names. Removing records of specimens of non-adult life stages eliminated 0.06% of records in our dataset. Filtering records that occur on the first day of the month removed 17.3% of remaining records. Of these remaining records, 480 species had enough data to generate phenoestimates (3 cell by year combinations with at least 3 distinct collectors, four distinct days and five observations). Another 148 species were removed due to lack of comprehensive trait information, and 67 multivoltine species were dropped due to not meeting the more stringent multivoltine data requirements. After applying all of our data filtering steps, we had a final dataset of 105,072 records used to calculate 2025 phenoestimates spanning 265 species (Figure 3), 114 years (1902–2016) and 104 grid cells (Figure 4).

The effects of climate and trait predictors on the onset, midpoint, termination and duration were consistent with previous studies examining the drivers of Lepidoptera phenology (Kharouba et al., 2014; Roy et al., 2015; Zografou et al., 2021). Onset of flight period was earlier and flight periods terminated later when temperatures were warmer (Table 1). Midpoint phenometrics are earlier in warmer locations for spring and summer flying species, whereas fall species do not shift their midpoint flight period across a temperature gradient (Figure 5a). Diapause stage influenced the timing of all phenometrics, typically via interactions with temperature (Figure S2), as documented previously by insect phenology studies (Belitz et al., 2021; Stemkovski et al., 2020). Compared to species that diapause as adults, species that diapause as pupae had shorter flight periods (Table 2) with earlier midpoints and terminations

(Table 1). Of particular note, especially given that flight termination timing is much less examined using NHCs, is our finding that both univoltine and multivoltine species delay termination under warm conditions, although this response is much greater for multivoltine species (Figure 5b).

Year effects were included in all top models except the duration model, and these effects were conditioned by the diapause strategy of a species in the onset and termination model (Table 1).

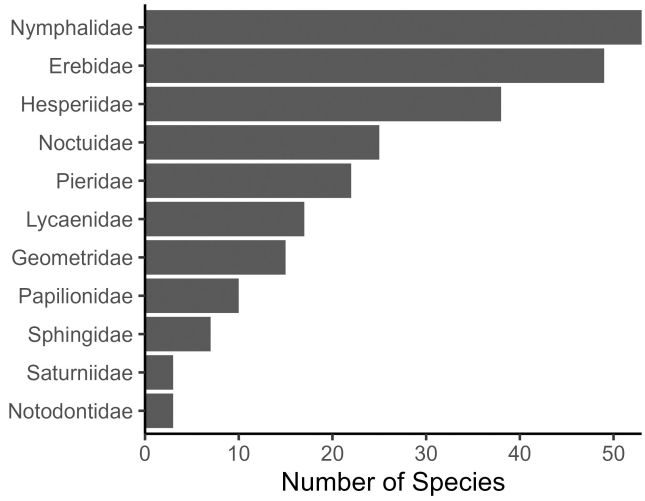

FIGURE 3 Number of species included in our statistical analysis for families with >2 species represented. Eight additional families (Depressariidae, Elachistidae, Gelechiidae, Gracillariidae, Plutellidae, Pyralidae, Crambidae and Lasiocampidae) had 1–2 species included in our analysis

Migratory species have earlier flight period onset and termination in more recent years, even after factoring in the effects of climate (e.g. temperature) and trait variables (Figure 5c,d). For species that diapause as adults, onset of flight periods is advancing and termination of flight periods is delaying even after accounting for climate and trait effects (Figure 5c,d). Year effects were also conditioned by climate variables, although these effects were relatively weak with large credible intervals (Figure S3).

The number of distinct days with observations, which we included as a measure of effort, was included in the top model for all phenometrics except the model predicting mid-flight period (Tables 1 and 2). More distinct days with observation records led to earlier onset, later termination and longer duration of flight periods (Figure 6). The number of distinct collectors was never retained in a top model. Finally, we found no evidence that diurnality affected Lepidoptera phenology as the 95% Bayesian credible intervals of this variable included zero (Tables 1 and 2). Top models explained a significant portion of the variation in our data with pseudo-R^2 (see above in Section 2) values of 0.79, 0.75, 0.80 and 0.70 for the onset, midpoint, termination and duration models, respectively.

4 | DISCUSSION

We provide a framework and methodological checklist (Box 1) to address key questions of macrophenology with a particular focus on examining the trends in Lepidoptera across North America. Our results confirm several findings of previous studies using structured survey

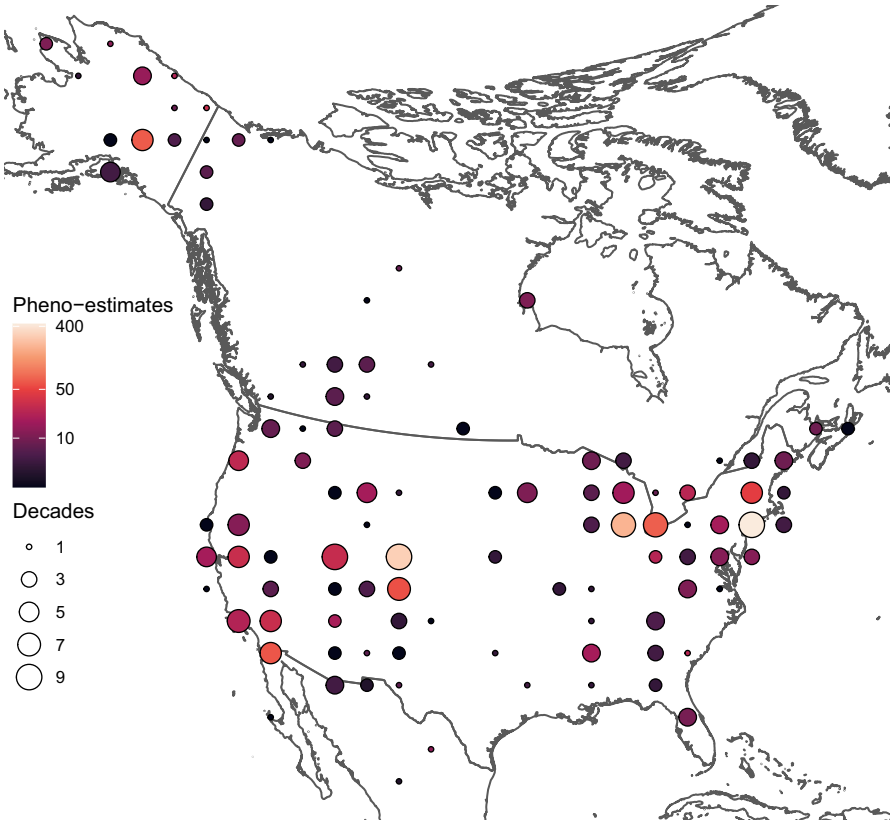

FIGURE 4 Locations at the centroids of the 250×250 km grids with enough data to produce species-by-year phenological estimates. Number of decades with a phenoestimate at each grid is represented by circle size, and the total number of phenoestimates at each grid is represented by circle colour

TABLE 1 Fixed effects coefficients for top onset, midpoint and termination models

Predictors	Onset	Midpoint	Termination
(Intercept)	**166.9 (1512.0 to 181.9)**	**235.2 (213.8 to 256.4)**	**280.9 (256.8 to 305.0)**
Temp	**−12.8 (−20.6 to −5.0)**	6.4 (−3.6 to 16.4)	**10.6 (0.4 to 20.7)**
Temp Seas		**−5.0 (−8.5 to −1.6)**	**−4.6 (−7.7 to −1.6)**
Prec	2.1 (−0.3 to 4.5)	**4.4 (1.9 to 7.0)**	
Prec Seas	**−3.0 (−5.1 to −0.9)**		**−2.9 (−5.0 to −0.8)**
Temp:Prec	**−5.2 (−7.7 to −2.7)**		
voltinism [Uni]		**−20.8 (−26.2 to −15.5)**	**−33.1 (−39.3 to −26.9)**
Diapause stage (DS) [Egg]	**18.1 (0.2 to 35.8)**	6.8 (−8.1 to 21.8)	−3.8 (−20.2 to 12.7)
DS [Larvae]	−6.6 (−22.5 to 9.0)	−9.6 (−22.7 to 3.4)	**−16.7 (−31.2 to −2.2)**
DS [Migratory]	10.5 (−11.1 to 31.9)	5.0 (−13.9 to 23.9)	9.7 (−12.2 to 31.5)
DS [Pupae]	**−17.6 (−34.1 to −1.3)**	**−18.5 (−31.9 to −5.0)**	**−29.4 (−44.3 to −14.4)**
Seasonality (Seas) [Spring]		**−52.7 (−63.2 to −42.1)**	**−46.5 (−58.6 to −34.3)**
Seas [Summer]		**−23.4 (−31.5 to −15.3)**	**−24.3 (−33.6 to −15.1)**
Diurnality [Diurnal]		−9.7 (−28.1 to 8.8)	−15.3 (−35.4 to 4.7)
Diurnality [Nocturnal]		0.2 (−17.5 to 17.9)	1.4 (−18.3 to 21.1)
Temp:DS [Egg]	**−13.1 (−22.1 to −4.1)**	**−26.1 (−35.6 to −16.7)**	**−27.2 (−37.0 to −17.4)**
Temp:DS [Larvae]	−3.3 (−11.0 to 4.3)	−7.8 (−15.7 to 0.1)	−1.9 (−10.1 to 6.3)
Temp:DS [Migratory]	**−19.9 (−34.7 to −5.2)**	−10.2 (−25.0 to 4.5)	−8.1 (−23.5 to 7.3)
Temp:DS [Pupae]	−5.5 (−13.2 to 2.2)	**−9.0 (−16.9 to −1.0)**	−3.8 (−11.9 to 4.3)
Temp:Seas [Spring]		**−19.6 (−27.2 to −12.1)**	**−17.6 (−25.4 to −9.9)**
Temp:Seas [Summer]		**−13.5 (−19.9 to −7.2)**	**−8.9 (−15.3 to −2.5)**
Temp:voltinism [Uni]			**−5.7 (−10.4 to −1.0)**
DS [Adult]:Year	**−9.4 (−16.9 to −1.9)**		5.1 (−2.6 to 12.7)
DS [Egg]:Year	−2.1 (−4.3 to 0.1)		0.1 (−2.3 to 2.4)
DS [Larvae]:Year	1.0 (−0.7 to 2.7)		1.2 (−0.5 to 3.0)
DS [Migratory]:Year	−7.3 (−15.9 to 1.4)		−6.0 (−15.0 to 3.0)
DS [Pupae]:Year	−0.0 (−2.7 to 2.6)		−0.6 (−3.3 to 2.1)
Prec:Year	**1.4 (0.2 to 2.7)**		
Prec Seas:Year			−0.8 (−2.0 to 0.5)
Temp:Year			−0.5 (−2.8 to 1.7)
Temp Seas:Year		**−1.4 (−2.8 to −0.0)**	−1.1 (−3.0 to 0.8)
Distinct observation days	**−3.0 (−4.3 to −1.8)**		**4.5 (3.4 to 5.6)**

Note: Bold values denote coefficients whose 95% Bayesian credible interval does not include zero.

data, highlighting that NHCs can be used to make biologically meaningful conclusions across broad spatial, temporal and taxonomic scales. We also showcase especially flight termination and duration findings that extend our knowledge of how global changes may continue to impact butterflies and moths into the future. Below we briefly discuss the biological insights our study provides before discussing more general recommendations about use of NHCs in broad-scale phenology studies.

4.1 | Key trends in lepidoptera phenology

The unique value of NHCs is their ability to capture trends over longer periods, broader regions and more taxa than are typically covered by surveys. Here we showcase the use of a modelling framework leveraging this temporal depth while also capturing spatially localized phenometric estimates, associated traits and climatic drivers. Our framework provides the means to test questions about the tempo and mode of phenological shifts, dependent on life-history strategy and regional climate contexts, critical for determining which species may be able to adapt phenologically in the face of global change (Hällfors et al., 2021; Macgregor et al., 2019). This framework confirms much of what has been shown previously, but also unveiled unexpected findings that do not conform to current hypotheses. As an example, we were surprised to find a temporal trend of earlier arrival of migratory species, even after accounting for climate warming. If migratory

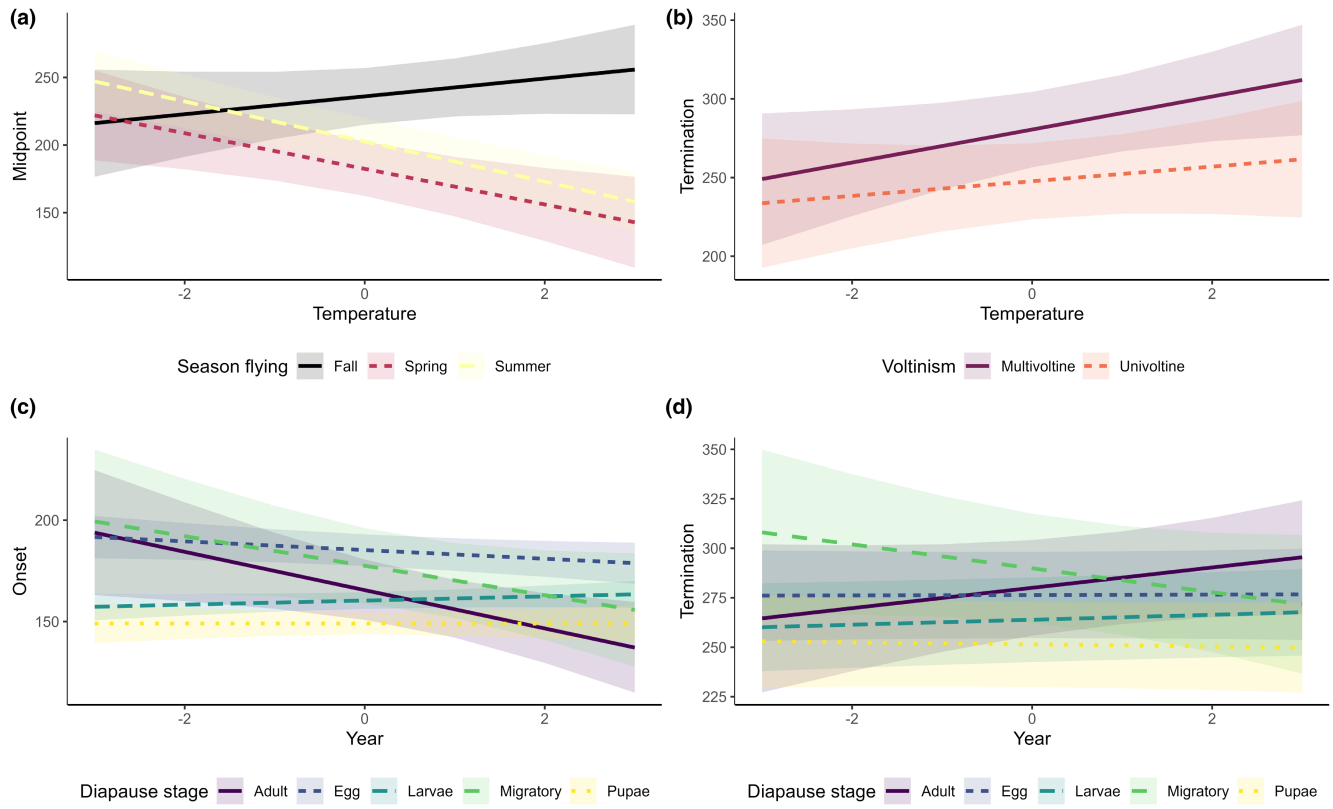

FIGURE 5 Two-way interactions displaying how species with different life-history traits have different phenological responses to temperature (a, b) and year (c, d) effects

species arrive at breeding sites too early to capture peak food sources, this may impact population viability (Visser et al., 2006). We also document that after accounting for climate effects, migratory species are leaving summering grounds earlier in more recent years, while species that diapause as adults are delaying their terminations in later years, even after accounting for changing climate. These results suggest that our climate variables are not encapsulating the temporal variation of important drivers of the onset and termination of flight period for species that overwinter as adults. Migratory species likely respond differently than resident species because they are moving through several climatically variable regions (Zipkin et al., 2012), whereas species diapausing as adults may more directly influenced by photoperiod since it is a primary cue for insects to initiate diapause (Denlinger, 2002).

We find trait and temperature effects on Lepidoptera phenology consistent with previous studies using more structured data, for example, earlier adult emergence in warmer temperatures (e.g. Roy et al., 2015). This extends to the importance of key traits, for example, our finding that diapause stage influences both the absolute timing of flight periods (Diamond et al., 2011) and sensitivity of response to temperature (Forrest, 2016; Roy et al., 2015). Our results also corroborate that seasonal flight timing is an important determinant of phenological sensitivity. Early season species have been predicted to be more sensitive to climate and have the most phylogenetically conserved phenologies since early season species will risk increased likelihood of frost or asynchrony with hostplants if phenology is mistimed (Pau et al., 2011). Our top midpoint model results, mirroring those of Kharouba et al. (2014), also support this hypothesis.

Finally, our broad taxonomic sampling, spanning multiple families of Lepidoptera, provided a means to examine if day and night-flying Lepidoptera differ in their phenological responses. This trait is strongly phylogenetically stratified across Lepidoptera, and we did not find evidence for a significant difference in phenological response between day and night-flying Lepidoptera when controlling for phylogenetic covariance. When phylogenetic covariance was not included in the model, diurnal species are found to terminate their flight periods earlier than nocturnal species, highlighting the importance of incorporating phylogenetic relationships into phenological modelling. These results in sum confirm the critical importance of winter diapause and voltinism traits, and less-so daily activity patterns, for predicting Lepidoptera phenology spatiotemporal trends.

4.2 | Challenges and opportunities

The workflows and modelling frameworks for NHC phenology studies require special attention, some of which are needed for any incidental record-based studies and some fully unique to digitized NHC data. Here, we briefly summarize the most critical challenges and

TABLE 2 Fixed effects coefficients for top duration models

Predictors	Phylogenetic linear mixed model
(Intercept)	**103.9 (82.9 to 124.8)**
Temp	**19.1 (11.3 to 26.8)**
Prec	1.4 (−1.2 to 4.0)
Prec Seas	2.2 (−0.1 to 4.6)
Voltinism [Univoltine]	**−29.8 (−35.5 to −24.1)**
Diapause stage (DS) [Egg]	**−36.1 (−50.0 to −21.9)**
DS [Larvae]	**−31.5 (−43.5 to −19.2)**
DS [Migratory]	−11.6 (−28.6 to 5.5)
DS [Pupae]	**−31.8 (−44.0 to −19.2)**
Diurnality [Diurnal]	−5.9 (−23.6 to 12.0)
Diurnality [Nocturnal]	0.8 (−16.5 to 18.2)
Seasonality (Seas) [Spring]	5.1 (−5.6 to 16.0)
Seas [Summer]	−3.3 (−11.6 to 5.0)
Temp:Voltinism [Univoltine]	**−14.8 (−19.4 to −10.2)**
Temp:Seas [Spring]	2.7 (−4.9 to 10.2)
Temp:Seas [Summer]	**7.5 (0.9 to 14.0)**
Temp:Prec	**4.5 (2.0 to 6.9)**
Distinct observation days	**7.3 (5.7 to 8.8)**

Note: Bold values denote coefficients whose 95% Bayesian credible interval does not include zero.

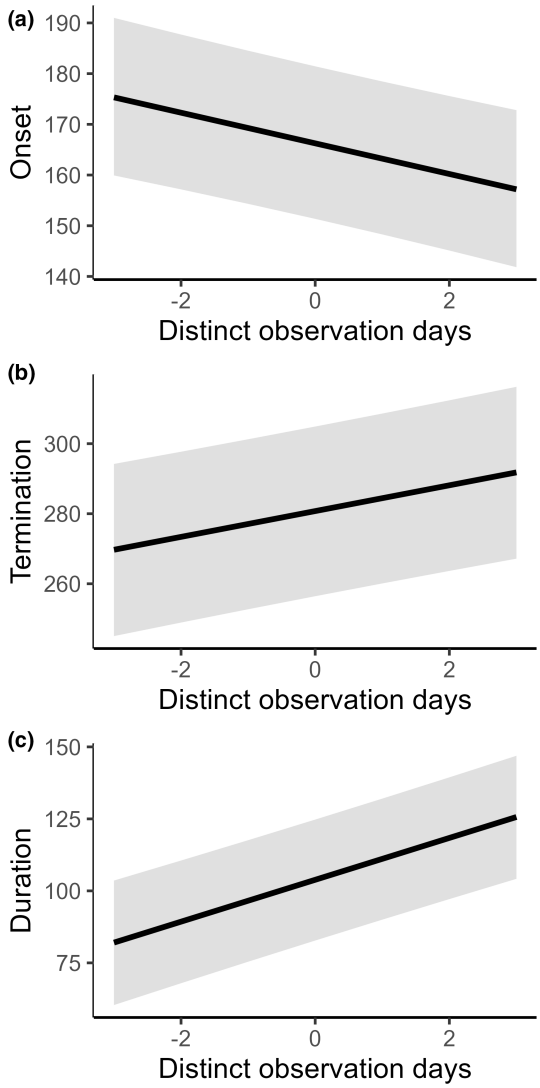

FIGURE 6 More observation days lead to significantly early onsets (a), later offsets (b) and longer durations (c)

potential solutions, drawing from this case study and the broader literature.

Taxonomic standardization is a critical challenge, as taxonomic names change due to revisions at multiple taxonomic levels, and there is no guarantee that data aggregators are either up to date with these changes or mapping synonymies. For this reason, researchers need to understand how aggregators manage names as GBIF, SCAN and iDigBio all differ in their name management approach. In nearly all cases, synonyms must be harmonized to a standardized species list to reduce potential for pseudoreplication. We found that doing so removed 95 synonyms out of 360 species names from our raw dataset. While our study is smaller, this parallels what has also been found in plants, where Weiser et al. (2007) dropped nearly half of the verbatim names in a dataset that originally contained 22,100 binomials via harmonization. In addition to reducing pseudoreplication, harmonizing taxonomic names will also add records for some species, increasing precision of the phenoestimates. In this study, we used the R package taxotools (Barve, 2020) to generate a standardized species list. Other options to generate standardized taxonomic names include the taxonomic name resolution service (Boyle et al., 2013) and the R package taxize (Chamberlain & Szöcs, 2013).

Choosing an appropriate spatial resolution is important to ensure that analyses are tractable given appropriate data density but still capture meaningful biological processes (Park et al., 2021). While finer-grain estimates are often preferable to localize processes driving phenology, these are impractical given the sparseness of NHC data. One approach to optimize data density versus cell coarseness trade-offs is to test alternate spatial grains and examine the number of cell–year–species combinations with enough data to estimate reliable phenometrics. In this case, given the focus on broad, continental-scale patterns using climatic predictors, coarse-grain cells were an acceptable trade-off for upping the number of phenometrics we could estimate. However, studies incorporating land-use changes may need to sacrifice quantity to model these more localized phenomena at finer resolutions. Similarly, temporal resolution is critical. Most robust phenological research will use annual or seasonal phenological metrics, most commonly measured in days, although data biases, precision or sufficiency sometimes leads to phenometrics to be reported in weeks.

Deciding on the minimum data requirements necessary for fitting phenometrics is also a key trade-off between number of estimates versus their reliability. While the minimum number of records needed to develop accurate species distribution models has been discussed in the literature for years (Mateo et al., 2010; van Proosdij et al., 2016; Wisz et al., 2008), minimum data requirements

BOX 1 Methodological checklist for phenological workflows using NHCs data

1. *Harmonize taxonomy* to generate a standardized list of species for use in analysis. This is especially important if the specimen records were downloaded from a source without a backbone taxonomy or if multiple data sources are being combined.
2. *Filter data* to remove records that are imprecise or incorrect. These may include records with coordinate precision or temporal precision too coarse to be included in further analyses.
3. *Annotate records* if necessary to ensure specimen records of only the phenophase of interest (e.g. fruiting, flowering) are included in the analysis. Human annotations can greatly be sped up using the software ImageAnt (https://gitlab.com/stuckyb/imageant).
4. *Select appropriate spatial and temporal grain* where analyses are tractable from a data perspective but are still meaningful from a biological perspective. Appropriate scales will depend on data densities, the life history of the group of interest, and the spatial and temporal scale at which predictor variables operate.
5. *Decide on minimum data requirements* for fitting phenological estimates. These requirements will be species specific and differ depending on the length of the phenophase, the complexity of phenological distribution and phenological metric of interest. Longer phenophases, multimodal phenological distributions and phenometrics closer to the bound of a distribution (i.e. onset and offset metrics) will require more data.
6. *Flag outlier phenometrics*. Even with many records, phenoestimates can be inaccurate and outlier estimates should be examined at a species-specific level. In this study, we propose a data-driven approach to flag potential outlier phenoestimates.
7. *Account for biases and autocorrelative structures* in the modelling framework. We include the number of days with an observation as a fixed effect in our models but other Bayesian approaches that quantify uncertainty in phenoestimates and propagate this uncertainty hierarchically exist. Additionally, phylogenetic signal and spatial autocorrelation must be accounted for to reduce false-positive rate.

have been less explored in phenological research using NHCs, although Belitz et al. (2020) demonstrated that modest increases in the number of incidental data records used to generate phenoestimates lead to significant improvements in accuracy. Additionally, that study showed that estimating phenometrics of species with long phenophases or bimodal phenological distributions was more difficult and may require greater sample size (Belitz et al., 2020), corroborating studies showing that minimum data requirements in distribution modelling must be based on species-specific biological information (van Proosdij et al., 2016). Here, we required more records to estimate phenometrics for multivoltine species, given the adult flight season is likely to be longer and more complex.

Simply setting minimum data requirements is not enough because incidental-based phenoestimates can still be biased if effort is highly skewed. Therefore, estimates should be examined for outliers before advancing to downstream analyses. We used a data-driven model residual approach, where onset estimates were considered outliers if associated residual values were too high. By including temperature values as a predictor variable, along with random effects for cell and species, we can identify inaccurate phenoestimates without needing the geographical locations of phenoestimates to be uniformly sampled across the range of a particular species.

Phenology analyses using incidental records need to account for observation effort, which can vary over time and space and thus lead to spurious conclusions. Here we used number of distinct collectors and days with distinct observations as two key effort measures. Distinct observation days was included as a fixed effect in the top onset, termination and duration model but dropped from midpoint models, supporting the supposition that onset and termination metrics are most sensitive to sampling variation (Miller-Rushing et al., 2008; Moussus et al., 2010). While simply including effort metrics as covariates in linear models is a commonly used method (Belitz et al., 2021; Johnston et al., 2021; Linden, 2011), better still is using a fully Bayesian approach that quantifies uncertainty in phenoestimates and hierarchically propagates this uncertainty (see Youngflesh et al., 2021).

Finally, multiple forms of autocorrelation in model residuals can lead to spurious model results and type I errors, which is especially problematic if the modelling goal is inference rather than exploration or prediction (Tredennick et al., 2021). We used a phylogenetic covariance matrix as a random effect in our model to account for phylogenetic signal and found that several model coefficients significant in the LMM no longer differed from zero in the PLMM. For example, LMMs predicted species that diapause as eggs or are migratory have onsets approximately a month later than species that diapause as adults (Supporting Information 2). However, when accounting for phylogenetic signal, the phenological delay was predicted to only be 2 weeks and had credible intervals that encompassed zero (Supporting Information 2). It is now easier than ever to generate phylogenetic synthesis trees utilizing resources such as the Open Tree of Life (Hinchliff et al., 2015). These synthesis trees, while still approximations, can also be used to directly test new questions such as whether timing and sensitivity of response to drivers are phylogenetically conserved. While we did not find evidence for spatial autocorrelation in our analyses, a variety of methods from using a spatial covariance matrix to include as a random effect (see Belitz et al., 2021) to simultaneous autoregressive approaches (see Du et al., 2020) can help account for that structure.

Given the multitude of biases and pitfalls found in digitized NHCs data, how can phenological research best move forward? Partnerships between data scientists, NHC and taxon-specific experts, and ecological modellers may be the surest route towards new innovations that can further harness NHCs to address pressing global change challenges. Frontier areas, such as multiple uses of machine learning approaches, both for phenology annotation, and for multiscale analysis, is one such track. Another is the integration of NHC and recent structured or unstructured citizen science data, which will likely require new methodological approaches. Our hope is that continued sharing of these innovations, aligned with clear descriptions of best practices, can help to fully unlock our understanding of the tempo, mode, drivers and outcomes of phenological changes at the broadest spatial, temporal and taxonomic scales.

AUTHOR CONTRIBUTIONS

Robert P. Guralnick and Michael W. Belitz conceived the idea for the manuscript and all authors contributed substantially to study design. Michael W. Belitz curated the occurrence data and conducted the analyses. Michael W. Belitz and Robert P. Guralnick led the writing of the manuscript. All authors contributed critically to manuscript drafts.

ACKNOWLEDGEMENTS

We thank the countless individuals who have contributed to making specimen records freely available to download. No funding directly supported this study, but MWB, EAL and RPG developed phenological tools and concepts used in this study that were supported by the National Science Foundation DEB # 1703048 and DEB # 1702664.

CONFLICT OF INTEREST

The authors have no competing or conflicting interests to report.

DATA AVAILABILITY STATEMENT

The data and code to reproduce the results and figures presented can be found on GitHub (https://github.com/mbelitz/LepPheno_BestPractices) and are archived on Zenodo (https://doi.org/10.5281/zenodo.6585090; Belitz, 2022b). Raw occurrence records needed to fully replicate workflow can be downloaded and unzipped from our Open Science Framework project (https://osf.io/wdzay/; Belitz, 2022a).

ORCID

Michael W. Belitz https://orcid.org/0000-0002-8162-5998
Elise A. Larsen https://orcid.org/0000-0002-9238-6777
Vaughn Shirey https://orcid.org/0000-0002-3589-9699
Daijiang Li https://orcid.org/0000-0002-0925-3421

REFERENCES

Anderson, D., & Burnham, K. (2004). *Model selection and multi-model inference* (2nd ed.). Springer-Verlag.

Barve, V. (2020). *Taxotools: Tools to handle taxonomic lists*. https://CRAN.R-project.org/package=taxotools

Barve, V. V., Brenskelle, L., Li, D., Stucky, B. J., Barve, N. V., Hantak, M. M., McLean, B. S., Paluh, D. J., Oswald, J. A., Belitz, M. W., Folk, R. A., & Guralnick, R. P. (2020). Methods for broad-scale plant phenology assessments using citizen scientists' photographs. *Applications in Plant Sciences*, *8*(1), e11315. https://doi.org/10.1002/aps3.11315

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Belitz, M. W. (2022a). *Phenological research based on natural history collections*. OSF. https://doi.org/10.17605/OSF.IO/WDZAY

Belitz, M. W. (2022b). *LepPheno_BestPractices: Code for manuscript submission (0.1.1)*. Zenodo. https://doi.org/10.5281/zenodo.6585090

Belitz, M. W., Barve, V., Doby, J. R., Hantak, M. M., Larsen, E. A., Li, D., Oswald, J. A., Sewnath, N., Walters, M., Barve, N., Earl, K., Gardner, N., Guralnick, R. P., & Stucky, B. J. (2021). Climate drivers of adult insect activity are conditioned by life history traits. *Ecology Letters*, *24*(12), 2687–2699. https://doi.org/10.1111/ele.13889

Belitz, M. W., Larsen, E. A., Ries, L., & Guralnick, R. P. (2020). The accuracy of phenology estimators for use with sparsely sampled presence-only observations. *Methods in Ecology and Evolution*, *11*(10), 1273–1285. https://doi.org/10.1111/2041-210X.13448

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., Matasci, N., Narro, M. L., Piel, W. H., Mckay, S. J., Lowry, S., Freeland, C., Peet, R. K., & Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*(1), 16. https://doi.org/10.1186/1471-2105-14-16

Brooks, S. J., Self, A., Powney, G. D., Pearse, W. D., Penn, M., & Paterson, G. L. J. (2017). The influence of life history traits on the phenological response of British butterflies to climate variability since the late-19th century. *Ecography*, *40*(10), 1152–1165. https://doi.org/10.1111/ecog.02658

Brooks, S. J., Self, A., Toloni, F., & Sparks, T. (2014). Natural history museum collections provide information on phenological change in British butterflies since the late-nineteenth century. *International Journal of Biometeorology*, *58*(8), 1749–1758. https://doi.org/10.1007/s00484-013-0780-6

Buckley, L. B., Arakaki, A. J., Cannistra, A. F., Kharouba, H. M., & Kingsolver, J. G. (2017). Insect development, thermal plasticity and fitness implications in changing, seasonal environments. *Integrative and Comparative Biology*, *57*(5). https://doi.org/10.1093/icb/icx032

Chamberlain, S. A., & Szöcs, E. (2013). taxize: Taxonomic search and retrieval in R. *F1000Research*, *2*, 191. https://doi.org/10.12688/f1000research.2-191.v2

Chauvier, Y., Zimmermann, N. E., Poggiato, G., Bystrova, D., Brun, P., & Thuiller, W. (2021). Novel methods to correct for observer and sampling bias in presence-only species distribution models. *Global Ecology and Biogeography*, *30*(11), 2312–2325. https://doi.org/10.1111/geb.13383

Cobb, N. S., Gall, L. F., Zaspel, J. M., Dowdy, N. J., McCabe, L. M., & Kawahara, A. Y. (2019). Assessment of north American arthropod collections: Prospects and challenges for addressing biodiversity research. *PeerJ*, *7*, e8086. https://doi.org/10.7717/peerj.8086

Covell, C. V., Jr. (2005). *A field guide to the moths of eastern North America*. Houghton Mifflin Co. https://www.cabdirect.org/cabdirect/abstract/19850526301

Davis, C. C., Willis, C. G., Connolly, B., Kelly, C., & Ellison, A. M. (2015). Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany*, *102*(10), 1599–1609. https://doi.org/10.3732/ajb.1500237

Denlinger, D. L. (2002). Regulation of diapause. *Annual Review of Entomology*, *47*, 93–122.

Diamond, S. E., Frame, A. M., Martin, R. A., & Buckley, L. B. (2011). Species' traits predict phenological responses to climate change in butterflies. *Ecology*, *92*(5), 1005–1012. https://doi.org/10.1890/10-1594.1

Du, Y., Mao, L., Queenborough, S. A., Primack, R., Comita, L. S., Hampe, A., & Ma, K. (2020). Macro-scale variation and environmental predictors of flowering and fruiting phenology in the Chinese angiosperm flora. *Journal of Biogeography*, *47*(11), 2303–2314. https://doi.org/10.1111/jbi.13938

Duchenne, F., Thébault, E., Michez, D., Elias, M., Drake, M., Persson, M., Rousseau-Piot, J. S., Pollet, M., Vanormelingen, P., & Fontaine, C. (2020). Phenological shifts alter the seasonal structure of pollinator assemblages in Europe. *Nature Ecology & Evolution*, *4*(1), 115–121. https://doi.org/10.1038/s41559-019-1062-4

Edwards, C. B., & Crone, E. E. (2021). Estimating abundance and phenology from transect count data with GLMs. *Oikos*, *130*(8), 1335–1345. https://doi.org/10.1111/oik.08368

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Peterson, A. T., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Forrest, J. R. (2016). Complex responses of insect phenology to climate change. *Current Opinion in Insect Science*, *17*, 49–54. https://doi.org/10.1016/j.cois.2016.07.002

Fu, Y. H., Liu, Y., De Boeck, H. J., Menzel, A., Nijs, I., Peaucelle, M., Peñuelas, J., Piao, S., & Janssens, I. A. (2016). Three times greater weight of daytime than of night-time temperature on leaf unfolding phenology in temperate trees. *New Phytologist*, *212*(3), 590–597. https://doi.org/10.1111/nph.14073

Gallinat, A. S., Ellwood, E. R., Heberling, J. M., Miller-Rushing, A. J., Pearse, W. D., & Primack, R. B. (2021). Macrophenology: Insights into the broad-scale patterns, drivers, and consequences of phenology. *American Journal of Botany*. https://doi.org/10.1002/ajb2.1793

GBIF. (2021). *Occurrence download*. 10.15468/dl.v84n8j

Hällfors, M. H., Pöyry, J., Heliölä, J., Kohonen, I., Kuussaari, M., Leinonen, R., Schmucki, R., Sihvonen, P., & Saastamoinen, M. (2021). Combining range and phenology shifts offers a winning strategy for boreal lepidoptera. *Ecology Letters*, *24*(8), 1619–1632. https://doi.org/10.1111/ele.13774

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., & Gude, K. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 12764–12769.

Isaac, N. J. B., Strien, A. J., van August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*(10), 1052–1060. https://doi.org/10.1111/2041-210X.12254

Ives, A. R. (2019). R^2s for correlated data: Phylogenetic models, LMMs, and GLMMs. *Systematic Biology*, *68*(2), 234–251. https://doi.org/10.1093/sysbio/syy060

Ives, A. R., & Li, D. (2018). 'rr2': An R package to calculate R^2s for regression models. *Journal of Open Source Software*, *3*(30), 1028. https://doi.org/10.21105/joss.01028

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Ruiz Gutierrez, V., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S. T., & Fink, D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, *27*(7), 1265–1277. https://doi.org/10.1111/ddi.13271

Karger, D., & Zimmermann, N. E. (2018). *CHELSAcruts—High resolution temperature and precipitation timeseries for the 20th century and beyond*. EnviDat. https://doi.org/10.16904/envidat.159

Kharouba, H. M., Lewthwaite, J. M. M., Guralnick, R., Kerr, J. T., & Vellend, M. (2018). Using insect natural history collections to study global change impacts: Challenges and opportunities. *Philosophical Transactions of the Royal Society B*, *374*, 20170405. https://doi.org/10.1098/rstb.2017.0405

Kharouba, H. M., Paquette, S. R., Kerr, J. T., & Vellend, M. (2014). Predicting the sensitivity of butterfly phenology to temperature over the past century. *Global Change Biology*, *20*(2), 504–514. https://doi.org/10.1111/gcb.12429

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A resource for timelines, Timetrees, and divergence times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(1), 1–26 10.18637/jss.v082.i13

Larsen, E. A., & Shirey, V. (2021). Method matters: Pitfalls in analysing phenology from occurrence records. *Ecology Letters*, *24*(6), 1287–1289. https://doi.org/10.1111/ele.13731

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., & Guralnick, R. P. (2021). Climate, urbanization, and species traits interactively drive flowering duration. *Global Change Biology*, *27*(4), 892–903.

Li, D., Dinnage, R., Nell, L. A., Helmus, M. R., & Ives, A. R. (2020). Phyr: An R package for phylogenetic species-distribution modelling in ecological communities. *Methods in Ecology and Evolution*, *11*(11), 1455–1463.

Li, D., & Ives, A. R. (2017). The statistical need to include phylogeny in trait-based analyses of community composition. *Methods in Ecology and Evolution*, *8*(10), 1192–1199. https://doi.org/10.1111/2041-210X.12767

Linden, A. (2011). Using first arrival dates to infer bird migration phenology. *Boreal Environment Research*, *16*, 49–60.

Macgregor, C. J., Thomas, C. D., Roy, D. B., Beaumont, M. A., Bell, J. R., Brereton, T., Bridle, J. R., Dytham, C., Fox, R., Gotthard, K., Hoffmann, A. A., Martin, G., Middlebrook, I., Nylin, S., Platts, P. J., Rasteiro, R., Saccheri, I. J., Villoutreix, R., Wheat, C. W., & Hill, J. K. (2019). Climate-induced phenology shifts linked to range expansions in species with multiple reproductive cycles per year. *Nature Communications*, *10*(1), 4455. https://doi.org/10.1038/s41467-019-12479-w

Mateo, R. G., Croat, T. B., Felicísimo, Á. M., & Muñoz, J. (2010). Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions*, *16*(1), 84–94. https://doi.org/10.1111/j.1472-4642.2009.00617.x

Maurer, J. A., Shepard, J. H., Crabo, L. G., Hammond, P. C., Zack, R. S., & Peterson, M. A. (2018). Phenological responses of 215 moth species to interannual climate variation in the Pacific northwest from 1895 through 2013. *PLoS One*, *13*(9), e0202850. https://doi.org/10.1371/journal.pone.0202850

Meineke, E. K., & Daru, B. H. (2021). Bias assessments to expand research harnessing biological collections. *Trends in Ecology & Evolution*, *36*(12), 1071–1082. https://doi.org/10.1016/j.tree.2021.08.003

Merow, C., Smith, M. J., & Silander, J. A., Jr. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Michonneau, F., Brown, J. W., & Winter, D. J. (2016). Rotl: An R package to interact with the open tree of life data. *Methods in Ecology and Evolution*, *7*(12), 1476–1481. https://doi.org/10.1111/2041-210X.12593

Miller-Rushing, A. J., Inouye, D. W., & Primack, R. B. (2008). How well do first flowering dates measure plant responses to climate change? The effects of population size and sampling frequency. *Journal of Ecology*, *96*(6), 1289–1296. https://doi.org/10.1111/j.1365-2745.2008.01436.x

Moussus, J.-P., Julliard, R., & Jiguet, F. (2010). Featuring 10 phenological estimators using simulated data. *Methods in Ecology and Evolution*. https://doi.org/10.1111/j.2041-210X.2010.00020.x

Myers, C. E., Stigall, A. L., & Lieberman, B. S. (2015). PaleoENM: Applying ecological niche modeling to the fossil record. *Paleobiology*, *41*(2), 226–244. https://doi.org/10.1017/pab.2014.19

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*. McGraw-Hill/Irwin.

Ooms, J., & Chamberlain, S. (2019). *phylocomr: Interface to "Phylocom."* https://CRAN.R-project.org/package=phylocomr

Opler, P. A. (1999). *A field guide to western butterflies*. Houghton Mifflin Harcourt.

Park, D. S., Newman, E. A., & Breckheimer, I. K. (2021). Scale gaps in landscape phenology: Challenges and opportunities. *Trends in Ecology & Evolution*, *36*(8), 709–721. https://doi.org/10.1016/j.tree.2021.04.008

Park, I. W., & Mazer, S. J. (2018). Overlooked climate parameters best predict flowering onset: Assessing phenological models using the elastic net. *Global Change Biology*, *24*(12), 5972–5984. https://doi.org/10.1111/gcb.14447

Pau, S., Wolkovich, E. M., Cook, B. I., Davies, T. J., Kraft, N. J. B., Bolmgren, K., Betancourt, J. L., & Cleland, E. E. (2011). Predicting phenology by integrating ecology, evolution and climate science. *Global Change Biology*, *17*(12), 3633–3643. https://doi.org/10.1111/j.1365-2486.2011.02515.x

Pearce, J. L., & Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, *43*(3), 405–412. https://doi.org/10.1111/j.1365-2664.2005.01112.x

Pearse, W. D., Davis, C. C., Inouye, D. W., Primack, R. B., & Davies, T. J. (2017). A statistical estimator for determining the limits of contemporary and historic phenology. *Nature Ecology & Evolution*, *1*(12), 1876–1882. https://doi.org/10.1038/s41559-017-0350-0

Pearson, K. D. (2019). A new method and insights for estimating phenological events from herbarium specimens. *Applications in Plant Sciences*, *7*(3), e01224. https://doi.org/10.1002/aps3.1224

Powell, J. A., & Opler, P. A. (2009). *Moths of Western North America*. University of California Press. https://doi.org/10.1525/9780520943773

Roy, D. B., Oliver, T. H., Botham, M. S., Beckmann, B., Brereton, T., Dennis, R. L. H., Harrower, C., Phillimore, A. B., & Thomas, J. A. (2015). Similarities in butterfly emergence dates among populations suggest local adaptation to climate. *Global Change Biology*, *21*(9), 3313–3322. https://doi.org/10.1111/gcb.12920

Roy, D. B., & Sparks, T. H. (2000). Phenology of British butterflies and climate change. *Global Change Biology*, *6*(4), 407–416. https://doi.org/10.1046/j.1365-2486.2000.00322.x

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Soltis, P. S. (2017). Digitization of herbaria enables novel research. *American Journal of Botany*, *104*(9), 1281–1284. https://doi.org/10.3732/ajb.1700281

Stemkovski, M., Pearse, W. D., Griffin, S. R., Pardee, G. L., Gibbs, J., Griswold, T., Neff, J. L., Oram, R., Rightmyer, M. G., Sheffield, C. S., Wright, K., Inouye, B. D., Inouye, D. W., & Irwin, R. E. (2020). Bee phenology is predicted by climatic variation and functional traits. *Ecology Letters*, *23*(11), 1589–1598. https://doi.org/10.1111/ele.13583

Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, *102*(6), e03336. https://doi.org/10.1002/ecy.3336

van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, *39*(6), 542–552. https://doi.org/10.1111/ecog.01509

Visser, M. E., Holleman, L. J. M., & Gienapp, P. (2006). Shifts in caterpillar biomass phenology due to climate change and its impact on the breeding biology of an insectivorous bird. *Oecologia*, *147*(1), 164–172. https://doi.org/10.1007/s00442-005-0299-6

Wagner, D. L. (2010). *Caterpillars of eastern North America: A guide to identification and natural history*. Princeton University Press. https://doi.org/10.1515/9781400834143

Wang, T., Hamann, A., Spittlehouse, D., & Carroll, C. (2016). Locally downscaled and spatially customizable climate data for historical and future periods for North America. *PLoS One*, *11*(6), e0156720. https://doi.org/10.1371/journal.pone.0156720

Weiser, M. D., Enquist, B. J., Boyle, B., Killeen, T. J., Jørgensen, P. M., Fonseca, G., Jennings, M. D., Kerkhoff, A. J., Lacher, T. E., Jr., Monteagudo, A., Vargas, M. P. N., Phillips, O. L., Swenson, N. G., & Martínez, R. V. (2007). Latitudinal patterns of range size and species richness of New World woody plants. *Global Ecology and Biogeography*, *16*(5), 679–688. https://doi.org/10.1111/j.1466-8238.2007.00323.x

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. (2017). Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology and Evolution*, *32*, 531–546.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & Group, N. P. S. D. W. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Youngflesh, C., Socolar, J., Amaral, B. R., Arab, A., Guralnick, R. P., Hurlbert, A. H., LaFrance, R., Mayor, S. J., Miller, D. A. W., & Tingley, M. W. (2021). Migratory strategy drives species-level variation in bird sensitivity to vegetation green-up. *Nature Ecology & Evolution*, *5*(7), 987–994. https://doi.org/10.1038/s41559-021-01442-y

Zimova, M., Willard, D. E., Winger, B. M., & Weeks, B. C. (2021). Widespread shifts in bird migration phenology are decoupled from parallel shifts in morphology. *Journal of Animal Ecology*, *90*(10), 2348–2361. https://doi.org/10.1111/1365-2656.13543

Zipkin, E. F., Ries, L., Reeves, R., Regetz, J., & Oberhauser, K. S. (2012). Tracking climate impacts on the migratory monarch butterfly. *Global Change Biology*, *18*(10), 3039–3049. https://doi.org/10.1111/j.1365-2486.2012.02751.x

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744–751. https://doi.org/10.1111/2041-210X.13152

Zografou, K., Swartz, M. T., Adamidis, G. C., Tilden, V. P., McKinney, E. N., & Sewall, B. J. (2021). Species traits affect phenological responses to climate change in a butterfly community. *Scientific Reports*, *11*(1), 3283. https://doi.org/10.1038/s41598-021-82723-1

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Belitz, M. W., Larsen, E. A., Shirey, V., Li, D., & Guralnick, R. P. (2023). Phenological research based on natural history collections: Practical guidelines and a lepidopteran case study. *Functional Ecology*, *37*, 234–247. https://doi.org/10.1111/1365-2435.14173