Methods: Winter temperatures predominate in spring phenological responses to warming

A. K. Ettinger, C. J. Chamberlain, I. Morales-Castilla, D. M. Buonaiuto, D. F. B. Flynn, T. Savas, J. A. Samaha & E. M. Wolkovich

The Observed Spring Phenology Responses in Experimental Environments (OSPREE) database

To conduct this meta-analysis, we followed systematic review methods to facilitate replication and use by other researchers *(e.g.,* we include at least 22/27 items on the PRISMA checklist, as summarized in Appendix 1, **?***)*. We searched the literature for research papers that experimentally addressed controls of temperature, chilling, and/or photoperiod requirements on the spring phenology of woody plant species. To identify phenological experiments that manipulated chilling, forcing, and/or photoperiod, we searched both ISI Web of Science and Google Scholar with the following terms:

1. TOPIC = (budburst OR leaf-out) AND (photoperiod or daylength) AND temperature*, which yielded 85 publications
2. TOPIC = (budburst OR leaf-out) AND dorman*, which yielded 193 publications

The initial searches yielded 201 papers, which we reviewed and assessed for inclusion in the database. To be included, papers needed to focus on woody plants in temperate ecosystems and test for photoperiod and/or temperature effects on budburst, leafout, or flowering, and we needed to be able to quantify the phenological response to chilling, forcing, and/or photoperiod. We used ImageJ to scrape these response data from figures, whenever possible, and added additional relevant information from the tables and text of each manuscript that could not be scraped. Multiple people checked scraping and data-entry, and mis-entered data and other mistakes were cleaned in R.

Our meta-analysis relies on the published literature where positive effects and larger effect sizes may be more likely to be be published *(***?***)*. Methods such as a funnel plot of effect size versus sample size can help diagnose the potential for such biases, but have many drawbacks and complications as well *(***???***)*. We could not use funnel plot methods here for several reasons: (1) our fundamental study design is based on three factors that can influence plant phenology, thus variation in effect size could be due to other levels of each factor, (2) studies had low sample sizes (75% of data came from studies with treatment sample sizes less than 10) and most often sample size was not reported (*i.e.*, in 25 out of 39 studies in the model with well-represented species), and finally (3) our modeling approach (see below) is designed to standardize and regularize data, thus it will pool some extreme effects that may arise from publication bias. Further, we note that these environmental cues have a firm physiological basis—thus, multiple lines of evidence (outside of publication bias) support that most studies should find an effect of (at least) chilling and forcing.

Some species are only represented in one dataset in the OSPREE database. In these instances, it is not possible to statistically differentiate between species, study, and treatment effects. To address this, we combined species found in only one study into "complexes" at the level of genera—such that each taxonomic

unit we use in our model occurs across multiple studies (and treatments). Thus our taxonomic units of analysis are "species complexes," which are either species represented in >1 dataset or complexes combining multiple species within a genus that are each singly represented in the dataset. Species represented in only one dataset with no congenerics in other datasets were excluded from most of our analyses, except when analyzing "all species."

Although all studies measured days to budburst, many communicated results differently (*e.g.*, days to budburst, degree-days to budburst, percent budburst, number of leaves). We standardized papers to common units whenever possible (details below) and further restricted studies to those for which forcing, chilling, and photoperiod treatments could be quantitatively identified. For this paper, we focus on studies measuring days to budburst. This subset of OSPREE includes data across 72 experiments (in 49 papers, Table **??**), 39 years, and 203 species (Table **??**, Fig. **??**). These experiments span a wide range of chilling, forcing, and photoperiod treatment levels (Fig. **??**), and many test for interactions between two of these cues (Table **??**). This subset of OSRPREE is freely available on KNB *(**?**)* and we hope other researchers will find it useful.

Defining budburst

Most studies defined budburst as initial "green tips" (33/49 papers). Select studies defined budburst as a specific increment of growth (*e.g.*, "0.5 cm of new growth") or as bud swell, leaf emergence, leaf unfolded, open bud scales, or petiole emerged. The remaining papers (4/49) did not include a definition of budburst. The majority of papers using the above definitions (34/49) required only one bud to have met the defined criteria of budburst, however, the remaining studies implemented specific thresholds to be met (*i.e.*, 10-100% of all buds on an individual needed to have bursted bud). For studies that quantified multiple measurements of percent budburst over time (days), we extracted one value of "days to budburst" of these multiple measurements to make them comparable to other studies. To extract this summary value, we selected the days to budburst when percent budburst was closest to 90%, including estimates as low as 49.5% budburst.

Estimating chilling

Chilling was reported far less in the OSPREE database than forcing and photoperiod. Although not all studies applied multiple treatments of forcing or photoperiod they generally all maintained and explicitly defined the forcing temperatures and daylengths applied in their treatments. In contrast, we found that most studies did not experimentally apply chilling by manipulating duration or temperature of chilling in controlled environments, nor did most quantify the total chilling imposed in their experiment. We therefore calculated the total chilling imposed by all studies; it would otherwise have been impossible to provide estimates with only experimental chilling given the rarity of such study designs (Fig. **??**).

To estimate total chilling we combined chilling from the field (*i.e.*, chilling before plant material was brought into controlled environment conditions) and experimental chilling (*i.e.*, chilling that plant material experienced in controlled environment conditions) into two widely used metrics of chilling: Utah units (Table **??**) and dynamic chill portions *(**??**)*. We used the `chillR` package (version 0.70.17) in R *(**??**)*, version 3.6.0, to calculate both Utah units and dynamic chill portions from time-series of hourly temperature data. To estimate field chilling, we generated hourly time series from a European-wide gridded climate dataset *(**?**)*, from which we extracted daily minimum and maximum temperature from the grid cells and dates during which experiments were conducted. For experimental chilling, we used reported chilling treatments to generate time-series of hourly temperature data.

In the formulation we used, Utah chilling units accumulated the most at temperatures between 2.4-9.1°C but slightly less at temperatures between 1.4-2.4°C and from 9.1-12.4°C. Utah units were reduced when temperatures fell below or exceeded this range (Table **??**). Chill portions accumulated when temperatures were between 0 and 7.2°C. We note that these models for chilling (both of which were originally developed for peach species) are *hypotheses* for how chilling may accumulate to affect the process of endodormancy release, but are likely to be inaccurate for many species. These models are, however, some of our current best approximations, and versions of them are routinely applied to forest trees *(e.g., **?**)*. We found the effects of

chilling and other cues remained qualitatively consistent across the two methods of estimating total chilling (*i.e.*, 95% uncertainty intervals of estimates for all cues in the standardized models overlapped, Table **??**).

We wished to explore model predictions across a wide range of experimental temperature conditions (*i.e.*, chilling and forcing temperatures) applied by studies included in the OSPREE database (Fig. 3). To do this, we needed to convert chilling temperature to total chilling units, which could be input into our model. There is no straightforward conversion between chilling temperature and total chilling, since the duration a temperature is applied affects chilling (Fig. **??**). We therefore made these conversions using two alternative approaches and we present both. For one approach, we generated daily time series of a range of experimental chilling temperatures for a range of durations spanning those in the OSPREE database (from -10 to 16°C for 7 to 240 days). We averaged across the range of durations for each temperature to get one chilling estimate per chilling temperature (Fig. **??**, **??**). For the alternative approach, we used historical climate data from a gridded climate dataset *(*E-OBS*,* **?***)* to estimate chilling, and used these historical relationships between mean winter temperature and total chilling to convert chilling temperature to a representative amount of total chilling (Fig. 3). We present this alternative approach in the main text as it is more closely related to field chilling conditions, which was by far the most common type of chilling across experiments.

Estimating forcing & photoperiod

Our studies included a diversity of designs for applying forcing and/or photoperiod experimentally, including studies that imposed constant forcing temperatures and forcing temperatures that varied between day and night. Additionally several studies applied forcing or photoperiod using a "ramped" design, such that treatments increased or decreased gradually over time throughout the duration of the application. For all studies we used the daylength of light as our photoperiod estimate (*e.g.*, a study with 8 hours of light and 16 hours of dark was recorded as "8"). For forcing, we used the temperature applied when forcing temperatures were constant (*i.e.*, the same temperature was applied 24 hours per day); if forcing varied with photoperiod, we estimated the mean daily temperature weighted by the hours that temperature was applied. Similarly, for studies that ramped forcing, we calculated a weighted average of forcing temperature over the period from when forcing treatments were applied until budburst day. For studies that ramped photoperiod, we used the photoperiod conditions that individuals initially experienced (*e.g.*, studies with photoperiod lengthening from 6 hours until budburst, we recorded as "6"). When forcing and photoperiod treatments were reported as ambient, we used the E-OBS dataset to estimate mean forcing temperature and the R package `geosphere` to estimate daylength associated with each date and latitude *(***?***)*.

Models

We fit four overall models: the main budburst model fit to species in OSPREE that measured days to budburst, the latitude model, which included only studies that had provenance latitude information, a model to examine how the design of chilling treatments affects estimated effects, and a model to test for life-stage differences in budburst responses. Given the complexity of our meta-analytic data, we fit each model separately, and present the main model in the main text as it was designed to best estimate chilling, forcing, and photoperiod cues (our primary goal here). The other models represent subsets of the data in the main model that allow more direct tests of relevant, related questions.

As our primary goal was to directly compare the effects of chilling, forcing, and photoperiod we standardized these predictor variables *(***?***)*. This was necessary because the range and scale of each predictor varied widely (total chilling ranged from -1304 to 4724 Utah units; forcing ranged from -5.2 to 32°C, photoperiod ranged from 6 to 24 hours). We followed well-established methods of subtracting the mean and dividing by the standard deviation *(***?***)* to yield "z-score" values for all predictor variables (total chilling units, forcing temperatures, and photoperiods in the experiments). In addition to these models with standardized predictors (Table **??**), we also fit models in which predictors were not standardized (Table **??**) so that estimates could be more easily interpreted on their natural scales. For all figures in which predictors are shown on their

natural scales, we use estimates from models in which predictors were not standardized.

All models were fit using the programming language `Stan` *(?)* (`www.mc-stan.org`), accessed via the `rstan` package (version 2.18.0) in R *(??)*, version 3.6.0. `Stan` provides efficient MCMC sampling via a No-U-Turn Hamiltonian Monte Carlo approach (more details can be found in *(?)* and in *(?)*). We validated our models using test data, then fit the models described below. In all models i represents each unique observation, sp is the species or species complex grouping, α terms represent intercepts, β terms represent slope estimates, and y is the days to budburst since forcing conditions were applied.

1. Main budburst model:

$$y_i = \alpha_{sp[i]} + \beta_{forcing_{sp[i]}} + \beta_{photoperiod_{sp[i]}} + \beta_{chilling_{sp[i]}} + \epsilon_i$$

,

$$\epsilon_i \sim N(0, \sigma_y^2)$$

The α and each of the three β coefficients were modeled at the species level, as follows:

$$\begin{aligned} \alpha_{sp} &\sim N(\mu_\alpha, \sigma_\alpha) \\ \beta_{forcing_{sp}} &\sim N(\mu_{forcing}, \sigma_{forcing}) \\ \beta_{photoperiod_{sp}} &\sim N(\mu_{photoperiod}, \sigma_{photoperiod}) \\ \beta_{chillingsp} &\sim N(\mu_{chilling}, \sigma_{chilling}) \end{aligned}$$

We applied this model to both a dataset with 203 species ("all species"), as well as with 65 species grouped into 36 taxa or "species complexes" (Tables **??**, **??**) and a model excluding a single study (**?**) because this study contains 112 species (Table **??**). We present estimates from the model fit to the reduced dataset in the main text (including for Figs. 2-3) as these estimates summarize across species that were more well-represented in multiple papers and study designs, and thus are likely to be more accurate estimates (more details above in section describing the OSPREE database). Based on our modeling approach, species from fewer studies will be pooled towards the overall mean. The reduced dataset model also excluded studies which reported only "ambient" forcing and photoperiod; these studies were included in the dataset with 203 species ("all species" model).

2. Latitude model: Given continuing debate over the role of photoperiod on budburst timing across a species' latitudinal range *(e.g.,* **??***)*, we examined the effect of including provenance latitude in a model similar to our main one, but designed to estimate effects of provenance latitude. This model estimated the effects of each phenological cue (chilling, forcing, photoperiod) on days to budburst (as in the main model), in addition to the effect of provenance latitude (i..e., the latitude of origin of plant material used in the experiment) and the interaction of photoperiod and provenance latitude. We include this interaction because photoperiod effects are expected to vary by latitude and this interaction may have important implications under climate change *(**???**)*.

   We followed the methods above for including species or species complex (see *Observed Spring Phenology Responses in Experimental Environments (OSPREE) database* section above), including only species and species complexes that had multiple provenance locations across different latitudes. This yielded the following model:

$$y_i = \alpha_{sp[i]} + \beta_{forcing_{sp[i]}} + \beta_{photoperiod_{sp[i]}} + \beta_{chilling_{sp[i]}} + \beta_{latitude_{sp[i]}} + \beta_{photoperiod:latitude_{sp[i]}} + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

The α and each of the five β coefficients were modeled at the species level, as follows:

$$\begin{aligned}
\alpha_{sp} &\sim N(\mu_\alpha, \sigma_\alpha) \\
\beta_{forcing_{sp}} &\sim N(\mu_{forcing}, \sigma_{forcing}) \\
\beta_{photoperiod_{sp}} &\sim N(\mu_{photoperiod}, \sigma_{photoperiod}) \\
\beta_{chilling_{sp}} &\sim N(\mu_{chilling}, \sigma_{chilling}) \\
\beta_{latitude_{sp}} &\sim N(\mu_{latitude}, \sigma_{latitude}) \\
\beta_{photoperiod:latitude_{sp}} &\sim N(\mu_{photoperiod:latitude}, \sigma_{photoperiod:latitude})
\end{aligned}$$

The latitude model is summarized in Table **??** and Fig. **??**.

3. Chilling study design model: As we found chilling to be the strongest cue, and given how few studies directly manipulate it (Fig. **??**), we also used a subset of our data to estimate how a study's experimental design for chilling impacts model estimates. For this, we included only species or species complexes used in both experiments that employed the Weinberger method (in this method plant tissue is sequentially removed from the field and then exposed to "forcing" conditions, with the assumption that tissues collected later experience more field chilling *(**?**)* and those that experimentally manipulated chilling (*i.e.*, by varying chilling temperatures and/or the duration of chilling conditions). We defined Weinberger studies as those with two or more field sample dates, each two or more weeks apart, that did not otherwise manipulate chilling. The chilling study-design model was thus:

$$\begin{aligned}
y_i = \alpha_{sp[i]} + \beta_{forcing} + \beta_{photoperiod} + \beta_{chilling} + \beta_{chillmethod} + \\
\beta_{forcing:chillmethod} + \beta_{photoperiod:chillmethod} + \beta_{chilling:chillmethod} + \epsilon_i,
\end{aligned}$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

The α coefficients were modeled at the species level, as follows:

$$\alpha_{sp} \sim N(\mu_\alpha, \sigma_\alpha)$$

The chilling design model is summarized in Table **??** and and Fig. **??**.

4. Life stage model: Previous research has found differences in spring phenology across life stages *(e.g.*, juvenile versus adult trees **?***)*. We tested for differences in days to budburst across life stages.

   We followed the guidelines above for including species or species complex (see *Observed Spring Phenology Responses in Experimental Environments (OSPREE) database* section above), including only the species and species complexes used in experiments involving plant material from adult trees as well as juvenile life stages (seedlings or saplings). The life-stage model was thus:

$$y_i = \alpha_{sp[i]} + \beta_{forcing_{sp[i]}} + \beta_{photoperiod_{sp[i]}} + \beta_{chilling_{sp[i]}} + \beta_{stage_{sp[i]}} + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

The α and each of the five β coefficients were modeled at the species level, as follows:

$$\begin{aligned}
\alpha_{sp} &\sim N(\mu_{\alpha}, \sigma_{\alpha}) \\
\beta_{forcing_{sp}} &\sim N(\mu_{forcing}, \sigma_{forcing}) \\
\beta_{photoperiod_{sp}} &\sim N(\mu_{photoperiod}, \sigma_{photoperiod}) \\
\beta_{chilling_{sp}} &\sim N(\mu_{chilling}, \sigma_{chilling}) \\
\beta_{stage_{sp}} &\sim N(\mu_{stage}, \sigma_{stage})
\end{aligned}$$

The life-stage model is summarized in Table **??**.

For all models, we chose weakly informative priors; increasing the priors three-fold did not change the model results. We ran four chains simultaneously, each with 2 500 sampling iterations (1 500 of which were used for warm-up), yielding 4 000 posterior samples for each parameter. We assessed model performance through $\hat{R}$ close to 1 and high n_{eff} (4 000 for most parameters, but as low as 713 for a few parameters in the latitude model), as well as visual consideration of chain convergence and posteriors *(**?**)*.

In our figures we show means ± 50% uncertainty intervals from our models (Figs. 2, 3, **??**, **??**, **??**, **??**), because our focus here is on the most likely value for each parameter (*e.g.*, estimated response to forcing) and because they are computationally stable *(**??**)*. See Tables **??**- **??** for 95% uncertainty intervals.

Modeling limitations based on experimental designs

An ideal model to predict budburst would potentially include (but is not limited to): interactions between cues, sigmoidal or other non-linearities to assess potential threshold effects, provenance location, methodological details (*e.g.*, if plant material was whole plant versus twigs, or whether temperatures were constant or varied each day, etc.), and measurement error. As with all models, though, we were limited in how many parameters we could estimate given available data. Thus we focused on species differences and used additional models to assess some of the potentially largest other effects (latitude, methods of estimating chilling, life stage). We were unable to estimate interactions between cues in our meta-analysis because very few studies design experiments to test for interactions between chilling, forcing, and photoperiod (Table **??**). Most experiments in our dataset, however, did include interactions between at least two cues (Table **??**); we fit our main budburst model to this subset of experiments (Table **??**), which resulted in qualitatively similar estimates to those of the model fit to the full set of experiments (Table **??**).

As our focus is on experiments, which—by design—often impose high variation in phenological cues, we expected a linear model for chilling, forcing, and photoperiod would be most appropriate. Non-linear models, however, are often appropriate for phenological cues, especially in nature, where chilling may typically be very high or extremely short photoperiods are rare. Thus we tested a non-linear (sigmoidal) model on the OSPREE data *(**?**)*. As chilling was the least experimentally manipulated in our database, we examined whether a sigmoidal curve for chilling would be more appropriate, but found that it was a poorer fit than a comparable all-linear model (R-squared = 0.53 versus 0.57), did not qualitatively alter estimates of forcing (-0.83 versus -0.79) or photoperiod (-0.25 versus -0.54) and led to non-biologically relevant estimates of chilling. Fitting meaningful non-linear models to experimental data may require more data, and/or data at very high and low chilling, forcing, and photoperiod values than are currently available.

The few studies that did incorporate interactions generally used the Weinberger method, which is not designed to robustly tease apart the effects of multiple cues (Table **??**, Fig. **??**). Similarly we found variation in thermoperiodicity and variation in study material were too infrequent to test for effects with current data (though our life-stage model found no large differences in days to budburst between material from adults [μ_{α} = 25.29] versus juveniles [β_{stage} = 24.2; 50% uncertainty intervals overlap], Table **??**). Our estimated effects therefore average over interactions *(**?**)*, but identifying them in future research will be critical to understanding and predicting budburst. This will be particularly challenging for forcing and chilling, as

a lack of information on endodormancy requirements makes disentangling forcing from chilling conditions impossible with current data *(?)*.

Our model does not include measurement error because these data were not possible to include for 25 out of 39 experiments included in the OSPREE model with well-represented species. For those studies that did report measurement error, the error was generally small relative to the magnitude of the responses (*e.g.*, standard deviation was, on average 12.07 % of the response variable for studies for which standard deviation was extracted). Thus, it is unlikely that adding measurement error to our analyses would have a large effect on our estimates *(?)*.