

¹ Comparison of large-scale citizen science data and
² long-term study data for phenology modeling

³ Shawn D. Taylor¹, Joan M. Meiners¹, Kristina Riemer², Michael C. Orr³, and
⁴ Ethan P. White^{2,4}

⁵ ¹School of Natural Resources and Environment, University of Florida
⁶ Gainesville, FL, United States

⁷ ²Department of Wildlife Ecology and Conservation, University of Florida,
⁸ Gainesville, FL, United States

⁹ ³Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology,
¹⁰ Chinese Academy of Sciences, Beijing 100101, P.R. China

¹¹ ⁴Informatics Institute, University of Florida, Gainesville, FL, United States

¹² Corresponding author: Shawn D. Taylor - shawntaylor@weecology.org

13

Abstract

Large-scale observational data from citizen science efforts are becoming increasingly common in ecology, and researchers often choose between these and data from intensive local-scale studies for their analyses. This choice has potential trade-offs related to spatial scale, observer variance, and inter-annual variability. Here we explored this issue with phenology by comparing models built using data from the large-scale, citizen science USA National Phenology Network (USA-NPN) effort with models built using data from more intensive studies at Long Term Ecological Research (LTER) sites. We built statistical and process based phenology models for species common to each dataset. From these models we compared parameter estimates, estimates of phenological events, and out-of-sample errors between models derived from both USA-NPN and LTER data. We found that model parameter estimates for the same species were most similar between the two datasets when using simple models, but parameter estimates varied widely as model complexity increased. Despite this, estimates for the date of phenological events and out-of-sample errors were similar, regardless of the model chosen. Predictions for USA-NPN data had the lowest error when using models built from the USA-NPN data, while LTER predictions were best made using LTER-derived models, confirming that models perform best when applied at the same scale they were built. This difference in the cross-scale model comparison is likely due to variation in phenological requirements within species. Models using the USA-NPN dataset can integrate parameters over a large spatial scale while those using an LTER dataset can only estimate parameters for a single location. Accordingly, the choice of dataset depends on the research question. Inferences about species-specific phenological requirements are best made with LTER data, and if USA-NPN or similar data are all that is available, then analyses should be limited to simple models. Large-scale predictive modeling is best done with the larger-scale USA-NPN data, which has high spatial representation.

37 tation and a large regional species pool. LTER datasets, on the other hand, have high site
38 fidelity and thus characterize inter-annual variability extremely well. Future research aimed at
39 forecasting phenology events for particular species over larger scales should develop models
40 which integrate the strengths of both datasets.

41 Keywords: LTER, USA-NPN, forecasting, budburst, flowering, data integration, scale

42 Introduction

43 Plant phenology, the timing of recurring biological events such as flowering, plays an important
44 role in ecological research extending from local to global scales (Cleland et al., 2007; Richardson
45 et al., 2013; Tang et al., 2016). At large scales the timing of spring leaf out and fall senescence in-
46 fluence the carbon budget of earth system models, which has implications for correctly accounting
47 for biosphere-atmosphere feedbacks in long-term climate forecasts (Richardson et al., 2012). At
48 smaller scales, species-specific responses to temperature and precipitation can alter flower commu-
49 nities (Diez et al., 2012; CaraDonna et al., 2014; Theobald et al., 2017) and affect the abundance
50 and richness of both pollinators (Ogilvie and Forrest, 2017; Ogilvie et al., 2017) and organisms at
51 higher trophic levels (Tylianakis et al., 2008). Plant phenology models that are robust at multiple
52 ecological scales, or deemed appropriate for a particular scale, are needed to better understand and
53 forecast the timing of key biological events.

54 Many plant phenology studies use intensively collected datasets from a single location over a long
55 time-period by a single research group (Cook et al., 2012; Wolkovich et al., 2012; Iler et al., 2013;
56 Roberts et al., 2015). These datasets have regular sampling and large numbers of samples over
57 long periods of time. As a result, the biological and climatic variability at that site is well rep-
58 resented. It is common for phenology models built with observations from a single site to not
59 transfer well to other sites (García-Mozo et al., 2008; Xu and Chen, 2013; Olsson and Jönsson,
60 2014; Basler, 2016). This lack of transferability can be driven by plasticity in phenology require-
61 ments, local adaptation, microclimates, or differences in plant age or population density (Kramer,
62 1995; Diez et al., 2012). For these reasons, data from a single location are not adequate for larger
63 scale phenology modeling. Accurately forecasting phenology at larger scales will require models

64 that account for the full range of variation across a species' range (Richardson et al., 2013; Tang
65 et al., 2016; Chuine and Régnière, 2017), which will necessitate the use of data sources beyond
66 traditional single-site studies.

67 Data from citizen science projects are becoming increasingly important for ecological research
68 (Kelling et al., 2009; Dickinson et al., 2010; Tulloch et al., 2013). Because these data are often
69 collected by large numbers of volunteers, it is possible to gather data at much larger scales than with
70 individual research teams. A relatively new citizen science project started in 2009, Nature's Note-
71 book run by The USA National Phenology Network (USA-NPN), collects phenology observations
72 from volunteers throughout the United States and makes the data openly available (Schwartz et al.,
73 2012). Data from this project have already been used to study variation in oak phenology at a conti-
74 nental scale (Gerst et al., 2017), develop large-scale community phenology models (Melaas et al.,
75 2016), and forecast long-term phenology trends (Jeong et al., 2013). Large-scale datasets from
76 China and Europe have already contributed considerably to phenological research (Xu and Chen,
77 2013; Olsson and Jönsson, 2014; Basler, 2016; Zhang et al., 2017), and the USA-NPN dataset has
78 the potential to meet these needs for North American plant species and communities. However,
79 the features that allow citizen science projects to collect data at large scales can also introduce spa-
80 tial biases toward cities and easily-accessible areas, and variation in sampling effort and observer
81 skill (Dickinson et al., 2010). With thousands of participants, the potential for variation among
82 observers in their determination of species identification and dating of phenological events is high.
83 While volunteers have been shown to be accurate at distinguishing different leaf and flower stages
84 for plants (Fuccillo et al., 2015) and can have high agreement on abundance estimates (Feldman
85 et al., 2018), contributions to USA-NPN are sometimes made sporadically across seasons, years,

86 and locations. This means that the quantity and quality of data at a specific site will typically be
87 more variable for citizen science efforts than for intensive, long-term studies.

88 In order to accurately model and forecast phenology, it is important to understand how the strengths
89 and weaknesses of intensive local studies and large-scale citizen science projects influence both our
90 inferences about biological processes driving phenology (e.g. warming requirements for a specific
91 plant) and our ability to predict future phenology events (e.g. forecasting when flowering or leaf
92 out occurs). Here, we fit a suite of plant phenology models for the budburst and first-flowering
93 phenophases of 24 plant species to data from both the USA-NPN and a set of intensive long-term
94 studies from the Long Term Ecological Research (LTER) network. We compare the resulting mod-
95 els based on both inference about models and parameters and predictions for unobserved events.
96 We then use this comparison to assess the best methods for both local- and large-scale phenology
97 modeling and to point the way forward for integrating large-scale and local-scale data to determine
98 the best possible models across scales.

99 Methods

100 Datasets

101 The USA National Phenology Network (USA-NPN) protocol uses status-based monitoring, where
102 via a phone app or web based interface observers answer 'yes,' 'no,' or 'unsure' when asked if
103 an individual plant has a specific phenophase present (Denny et al., 2014). Phenophases refer to
104 specific phases in the annual cycle of a plant, such as the presence of emerging leaves, flowers,

¹⁰⁵ fruit, or senescent leaves. Sites in the USA-NPN datasets are located across the U.S. and generally
¹⁰⁶ clustered around populated areas (Figure 1). To represent long-term, intensive phenology studies
¹⁰⁷ we used four datasets from North America representing three major ecosystem types (Table 1,
¹⁰⁸ Figure 1). All four long-term studies are located in the U.S. and are part of the Long Term Ecolog-
¹⁰⁹ ical Research network (LTER). The Harvard Forest and Hubbard Brook Long Term Experimental
¹¹⁰ Forest are located in the northeastern U.S. and are dominated by deciduous broadleaf species. The
¹¹¹ H.J. Andrews Experimental Forest is a coniferous forest in the coastal range of the western U.S.
¹¹² The Jornada Experimental Range is in the Chihuahua desert of the southwestern U.S.

¹¹³ We downloaded all USA-NPN observations from 2009, when collections began, to 2016 for the
¹¹⁴ following phenophases: Breaking Leaf Buds, Breaking Needle Buds, Emerging Needles, and Open
¹¹⁵ Flowers (USA National Phenology Network, 2017). The first three phenophases apply to the
¹¹⁶ 'leaf out' phase for deciduous broadleaves, evergreen conifers, and pines, respectively. The 'Open
¹¹⁷ Flowers' phenophase refers to fully-open flowers and applies to all angiosperms. Hereafter, we
¹¹⁸ will refer to these as either 'Flowers' for the Open Flower phenophase, or 'Budburst' for all other
¹¹⁹ phenophases. We subset the USA-NPN observations similar to methods outlined in Crimmins et al.
¹²⁰ (2017). First, 'yes' observations for individual plants were kept only if they were preceded by a
¹²¹ 'no' observation within 30 days. Observations for 'Budburst' that were past day of year (DOY)
¹²² 172, and for 'Flowers' that were past DOY 213 were dropped to minimize any influence from
¹²³ outliers. We inferred the observed DOY of each phenophase as the midpoint between each 'yes'
¹²⁴ observation and the preceding 'no' observation. Finally, only species that had greater than 30 total
¹²⁵ observations were kept. Crimmins et al. (2017) only kept observations that were preceded by a 'no'
¹²⁶ within 15 days, and also grouped multiple individuals at single sites to a single observation. We

₁₂₇ used 30 days to allow for a greater number of species to be compared. We tested the sensitivity of
₁₂₈ this choice by also performing the analysis using a 15 day cutoff. We chose not to group multiple
₁₂₉ individuals at a single site to better incorporate intra-site variability.

₁₃₀ In the LTER datasets observation metrics varied widely due to different protocols. To match the
₁₃₁ USA-NPN data we converted all metrics to binary 'yes' and 'no' observations for each phenophase
₁₃₂ (see Appendix S1 for details). Three of the LTER datasets (Hubbard Brook, Harvard Forest, and
₁₃₃ H.J. Andrews) had a sampling frequency of 3-7 days during the growing season. The Jornada
₁₃₄ dataset had a sampling frequency of 30 days. As with the USA-NPN data, we inferred the date for
₁₃₅ each phenophase as the midpoint between the first 'yes' observation and most recent 'no' observa-
₁₃₆ tion, and only kept species and phenophase combinations which had at least 30 total observations.
₁₃₇ After data processing there were 38 species and phenophase combinations (with 24 unique species)
₁₃₈ common to both the USA-NPN and LTER datasets to use in the analysis (Table 1 & Appendix S2:
₁₃₉ Table S1). Using a 15 day cutoff in the USA-NPN dataset resulted in 35 unique combinations with
₁₄₀ 23 species.

₁₄₁ Models

₁₄₂ It is common to fit multiple plant phenology models to find the one that best represents a specific
₁₄₃ species and phenophase (Chuine et al., 2013). For each of the 38 species and phenophase com-
₁₄₄ binations in the five datasets (USA-NPN and four LTER datasets), we fit eight phenology models
₁₄₅ (Table 2). The *Naive model* uses the mean DOY from prior observations as the estimated DOY.
₁₄₆ The *Linear model* uses a regression with the mean spring (Jan. 1 - March 31) temperature as the
₁₄₇ independent variable and DOY as the response variable. For the six remaining models, the general

¹⁴⁸ form is based on the idea that a phenological event will occur once sufficient thermal forcing units,
¹⁴⁹ F^* , accumulate from a particular start day of the year (t_1). Forcing units are a transformation of
¹⁵⁰ the daily mean temperature and are calculated differently for each model (Table 2). The start day
¹⁵¹ can either be estimated or fixed. For the *Growing Degree Day (GDD)* model, forcing units are
¹⁵² the total degrees above the threshold T_{base} (Réaumur, 1735; Wang, 1960; Hunter and Lechowicz,
¹⁵³ 1992). The *Fixed GDD model* uses the same form but has fixed values for start day ($t_1 = \text{Jan 1}$)
¹⁵⁴ and temperature threshold ($T_{base} = 0^\circ\text{C}$). The *Alternating model* has a variable number of required
¹⁵⁵ forcing units defined as a function of the total number of days below 0°C since Jan. 1 (number of
¹⁵⁶ chill days - NCD). The *Uniforc model* is like the *GDD model* but with the forcing units transformed
¹⁵⁷ via a sigmoid function (Chuine, 2000).

¹⁵⁸ We also fit two models that attempt to capture spatial variation in phenological requirements. The
¹⁵⁹ first spatial model, *M1*, is an extension of the *GDD model* which adds a correction in the required
¹⁶⁰ forcing using the photoperiod (L) (Blümel and Chmielewski, 2012). The second, the *Macroscale
161 Species-specific Budburst model (MSB)*, uses the mean spring temperature as a linear correction
¹⁶² on the total forcing required in the *Alternating model* (Jeong et al., 2013). Since there is little to
¹⁶³ no spatial variation in the LTER datasets, we fit the two spatial models to data from the USA-NPN
¹⁶⁴ only. We compared the resulting parameters, estimates, and errors for the USA-NPN derived *M1*
¹⁶⁵ and *MSB* models to their non-spatial analogs (the *GDD* and *Alternating models*, respectively) for
¹⁶⁶ each species and phenophase in the LTER data.

¹⁶⁷ We extracted corresponding daily mean temperature for all USA-NPN and LTER observations
¹⁶⁸ from the gridded PRISM dataset using the latitude and longitude of the site associated with each
¹⁶⁹ observation (PRISM Climate Group, 2004). We parameterized all models using differential evo-

¹⁷⁰ lution to minimize the root mean square error (RMSE) of the estimated DOY of the phenological
¹⁷¹ event. Differential evolution is a global optimization algorithm which uses a population of ran-
¹⁷² domly initialized models to find the set of parameters that minimize the RMSE (Storn and Price,
¹⁷³ 1997). Confidence intervals for parameters were obtained by bootstrapping, in which individual
¹⁷⁴ models were re-fit 250 times using a random sample, with replacement, of the data. We made
¹⁷⁵ predictions by taking the mean DOY estimated from the 250 bootstrapped iterations. A random
¹⁷⁶ subset consisting of 20% of observations from each species and phenophase combination was held
¹⁷⁷ out from model fitting for later evaluation.

¹⁷⁸ Analysis

¹⁷⁹ As described above, we fit two sets of models for each species and phenophase: one set of models
¹⁸⁰ parameterized using only USA-NPN data, and one set parameterized using only LTER data (with
¹⁸¹ the exception of the *M1* and *MSB* models, see above). We performed three primary analyses from
¹⁸² these model outputs by comparing: 1) the model parameters, 2) estimates from the models, and 3)
¹⁸³ out-of-sample errors from each model.

¹⁸⁴ To compare the inferences about process made by the two datasets, we compared the distribution
¹⁸⁵ of each parameter between LTER and USA-NPN derived models for each species and phenophase
¹⁸⁶ combination. Using the mean value of each bootstrapped parameter, we also calculated the coeffi-
¹⁸⁷ cient of determination (R^2) between LTER and USA-NPN derived models among the 38 species-
¹⁸⁸ phenophases. In three cases where a species phenophase combination occurred in two LTER sites
¹⁸⁹ (Budburst for *Acer saccharum*, *Betula alleghaniensis*, and *Fagus grandifolia* in the Harvard and
¹⁹⁰ Hubbard Brook datasets) they were compared separately to the USA-NPN data.

191 Next we compared the estimates of phenological events between models. Models with different pa-
192 rameter values, and even entirely different structures, can produce similar estimates for the date of
193 phenological events (Basler, 2016). Therefore, to compare the predictions and potential forecasts
194 for models fit to the different datasets, we compared the estimated DOY predicted by the LTER
195 and USA-NPN derived models for all held out observations. For each of the eight models, we
196 calculated the coefficient of determination (R^2) between LTER and USA-NPN derived estimates
197 for estimates made at the four LTER sites and across all USA-NPN sites.

198 Finally, we directly evaluated performance using out-of-sample errors from the four combinations
199 of models and observed data: A) LTER-derived models predicting LTER observations, B) USA-
200 NPN derived models predicting LTER observations, C) LTER-derived models predicting USA-
201 NPN observations, and D) USA-NPN derived models predicting USA-NPN observations. Using
202 the RMSE values from held out observations, we compared the performance of LTER and USA-
203 NPN derived models on different data types in two different ways. First, we focused on local-scale
204 prediction by calculating the difference in the RMSE of LTER and USA-NPN derived models
205 solely with LTER observations. Secondly, to focus on large-scale prediction we calculated the
206 difference in RMSE using solely USA-NPN data. These differences were calculated for each of
207 the model types and 38 species-phenophase combinations. Negative values indicate that LTER-
208 derived models perform better, while positive values indicate that the USA-NPN derived model
209 performed better. We used a t-test to test the difference from zero in these values. In the three
210 cases where the same species and phenophase combination occurred in two LTER sites, we made
211 the LTER-LTER comparison within each site, not across sites, to focus on local scale prediction
212 when LTER data are available. Absolute RMSE values as well as Pearson correlation coefficients

213 are provided in the supplement for specific species (Appendix S2: Figure S5-S7) and with all
214 observations aggregated together (Appendix S2: Table S2).

215 We performed all analysis using both the R and Python programming languages (R Core Team,
216 2017; Python Software Foundation, 2018). Primary R packages used in the analysis included
217 dplyr (Wickham et al., 2017), tidyr (Wickham and Henry, 2018), ggplot2 (Wickham, 2016), lubri-
218 date (Grolemund and Wickham, 2011), prism (Hart and Bell, 2015), raster (Hijmans, 2017), and
219 sp (Pebesma and Bivand, 2005). Primary Python packages included SciPy (Jones et al., 2001),
220 NumPy (Oliphant, 2006), Pandas (McKinney, 2010), and MPI for Python (Dalcin et al., 2011).
221 Code to fully reproduce this analysis is available on GitHub (https://github.com/sdtaylor/phenology_dataset_study)
222 and archived on Zenodo (<https://doi.org/10.5281/zenodo.1256705>)

223 Results

224 Throughout the analysis there were no qualitative differences between a 30-day or 15-day threshold
225 between the first 'yes' and most recent 'no' observation in the USA-NPN dataset. Results presented
226 here reflect the 30 day cutoff; see the supplementary figures S2-S4 in Appendix S2 for matching
227 figures using a 15 day cutoff.

228 The best matches between parameter estimates based on USA-NPN and LTER data were the *Fixed*
229 *GDD model* ($R^2 = 0.49$) and the *Linear model* ($R^2 = 0.39$ for β_1 and -0.05 for β_2). The param-
230 eters for all other models had R^2 values <0 indicating that the relationship was worse than no
231 relationship between the parameters (but with matching mean parameter values across the two sets
232 of models) (Figure 2). The *Naive model* showed a distinct late bias in mean DOY estimates for

phenological events, likely resulting from the LTER datasets being mostly in the northern United States compared to the site locations of the USA-NPN dataset (Figure 2). The large outlier for the *Fixed GDD model* is *Larrea tridentata*; this species' flower phenology is largely driven by precipitation, which is not considered in the Fixed GDD model (Beatley, 1974). While the *Fixed GDD* and *Linear* models showed reasonable correspondence between parameter estimates, all parameters for individual species and phenophase combinations had different distributions between USA-NPN and LTER-derived models (Appendix S2: Figure S10-S11).

When comparing estimates of phenological events between the two sets of models, many USA-NPN and LTER models produced similar estimates (Figure 3). The *Fixed GDD model* had the highest correlation between the two model sets at USA-NPN sites ($R^2 = 0.82$), while the *GDD*, *M1*, and *Uniforc* models had the highest correlation at LTER sites ($R^2 = 0.51$, 0.52, and 0.51, respectively). Comparing models with spatial corrections to the non-spatial alternatives, the *MSB* (an extension of the *Alternating model* with a spatial correction based on mean spring temperature, see Table 2 and Methods) improved the correlation between the two datasets over the *Alternating model*. The *MSB model* improved the R^2 from 0.36 to 0.45 at LTER sites, and from -0.23 to -0.15 at USA-NPN sites. The *M1 model* (an extension of the *GDD model* with a spatial correction based on day length) improved the correlation over the *GDD model* only slightly at LTER sites (from 0.51 to 0.52) and did not improve the correlation at USA-NPN sites.

When comparing the predictive performance using out-of-sample errors, USA-NPN derived models made more accurate predictions for held-out USA-NPN observations, and LTER-derived models performed better on held-out LTER observations (all $p < 0.001$, Figure 4). The *Naive* and *Linear* models had the largest differences between the two model sets, while the *Fixed GDD model*

had relatively similar errors when evaluated on both USA-NPN and LTER held-out observations.

Although the *Fixed GDD model* had the highest agreement in accuracy between USA-NPN and LTER-derived models, it was not the best performing model overall. The *GDD* and *Uniforc* models made the best out of sample predictions, having the lowest RMSE and Pearson coefficient when aggregating all observations together (Appendix S2: Table S2). One exception was that the Fixed GDD model had a slightly higher Pearson value when using LTER-derived models to make predictions for USA-NPN data. The best model for specific species and phenophases varied, but was commonly the *Uniforc* or *GDD* models (Appendix S2: Figs. S5-S6).

Discussion

Data used to build phenology models typically fall into two categories: intensive long-term data with long time-series at a small number of locations (e.g., LTER data in this study), and large-scale data with less intensive sampling at hundreds of locations (e.g., USA-NPN data) (Table 3).

This data scenario—a small amount of intensive data and a large amount of less intensive data—is common in many areas of science and makes it necessary to understand how to choose between, or combine, data sources (Hanks et al., 2011). We explored this issue for phenology modeling in relation to making predictions and inferring process from models. For inference we found that models based on different data sources resulted in different parameter estimates for all but the simplest models. For prediction we found that models fit to different data sources tended to make similar predictions, but that models better predicted out-of-sample data from the data type to which they were fit. These results are consistent with other research showing that phenology model

²⁷⁵ performance decreases when transferring single-site models to other locations (García-Mozo et al.,
²⁷⁶ 2008; Xu and Chen, 2013; Basler, 2016), and with the call for models that better incorporate
²⁷⁷ spatial variation in phenology requirements (Richardson et al., 2013; Chuine and Régnière, 2017).
²⁷⁸ Understanding and making predictions for the phenology of a single location is best served by
²⁷⁹ intensive local-scale data, when available, but large-scale datasets work better for extrapolating
²⁸⁰ phenology predictions across a species range. Thus, the best choice of both data and models
²⁸¹ depends on the desired research goals.

²⁸² In this study, parameter estimates differed widely within the same phenology model when fit to the
²⁸³ two different types of data, except for the simplest process-oriented model: the *Fixed GDD* (Figure 2). These differences may be caused by a variety of factors that have different implications for
²⁸⁴ interpreting process-oriented models and their parameters. First, the differences could result from
²⁸⁵ limitations in the sampling of the USA-NPN dataset, such as irregular sampling of the same loca-
²⁸⁶ tion within or between seasons, leading to less accurate parameter estimates. If this is the case, it
²⁸⁷ would suggest that using LTER data is ideal for making inferences about plant physiology, and that
²⁸⁸ focusing on the *Fixed GDD model* is best for making inferences when USA-NPN data are all that
²⁸⁹ is available. Second, spatial variation (e.g. from local adaptation, acclimation, microclimates, or
²⁹⁰ plant age) in phenology requirements and drivers could contribute to these differences (Diez et al.,
²⁹¹ 2012; Zhang et al., 2017). Models built using USA-NPN data integrate over that spatial variation,
²⁹² while models built using LTER data only estimate the phenological requirements for a specific site.
²⁹³ In this case, USA-NPN data would provide a better estimate of the general phenological require-
²⁹⁴ ments of a species, but LTER data would provide a more accurate understanding for a single site.
²⁹⁵ The best solution to this issue would be the development of models that accurately incorporate

²⁹⁷ spatial variation, such as including genetic variation between different populations (Chuine and
²⁹⁸ Régnière, 2017), although localized models could also be generated when large-scale predictions
²⁹⁹ are unnecessary. Third, these differences could result from issues with model identifiability: since
³⁰⁰ different parameter values can yield nearly identical estimates of phenological events, parameter
³⁰¹ estimates can differ between datasets even when the underlying processes generating the data are
³⁰² the same. Information about which of these issues may be causing the differences between datasets
³⁰³ can be explored using the analyses in the current study, as will be explained below.

³⁰⁴ Despite substantial differences in parameter estimates, LTER and USA-NPN derived models pro-
³⁰⁵ duced similar estimates for phenological events in most cases (Figure 3). This greater correspon-
³⁰⁶ dence between predictions than parameters suggests that more complex models may have iden-
³⁰⁷ tifiability issues. For example, two *GDD* models with parameters of $t_1=1$, $F=10$, $T_{base}=0$ and
³⁰⁸ $t_1=5$, $F=5$, $T_{base}=0$ produce nearly identical estimates in many scenarios. This possibility is sup-
³⁰⁹ ported by the fact that the highest correlation between parameter estimates is seen in models with
³¹⁰ only 1 or 2 parameters. In addition, bootstrap results for more complex models suggest a high
³¹¹ degree of variability in parameter estimates and potentially multiple local optima in fits to both
³¹² USA-NPN and LTER data (Appendix S2: Figure S10-S11). Finally, parameter estimates of more
³¹³ complex models are also not consistent among models for the same species when comparing mul-
³¹⁴ tiple LTER datasets (Appendix S2: Figure S8-S9). These results are consistent with research
³¹⁵ showing that models failed to estimate the starting day of warming accumulation solely from bud-
³¹⁶ break time-series, thus producing parameter estimates that were not biologically realistic (Chuine
³¹⁷ et al., 2016). Basler (2016) suggests that the key component in phenology models is the thermal
³¹⁸ forcing, with additional parameters being sensitive to over-fitting. Here, our simplest model, the

³¹⁹ *Fixed GDD model* which uses only a warming component, had the highest correlation among pa-
³²⁰ rameters between LTER and USA-NPN datasets. In combination with the aforementioned studies,
³²¹ our results indicate that caution is warranted in interpreting parameter estimates from complex
³²² phenology models regardless of the data source used for fitting the models.

³²³ While more complex phenology models appear to have identifiability issues, there is also evidence
³²⁴ that they capture useful information, beyond the *Fixed GDD model*, based on their ability to make
³²⁵ out-of-sample predictions. Based on the RMSE, the *GDD* and *Uniforc* models produce the best
³²⁶ out-of-sample predictions for the majority of species and phenophases at both USA-NPN and
³²⁷ LTER datasets (Appendix S2: Figure S5 & S6). This demonstrates that the more complex models
³²⁸ are capturing additional information about phenology, and that some of the differences between
³²⁹ datasets result from differences in either the scales or the sampling of the data. Spatial variation in
³³⁰ phenological requirements is known to exist in plants (Zhang et al., 2017). In combination with our
³³¹ results showing observed differences in parameter estimates between LTER sites (Appendix S2:
³³² Figure S8-S9), this suggests that variation in phenological requirements across the range is likely
³³³ important. However, the models that attempted to address this by incorporating spatial variation
³³⁴ did not yield improvements over their base models in our analyses. Specifically, correspondence
³³⁵ between parameter estimates (Figure 2), estimates of phenological events (Figure 3), and out-
³³⁶ of-sample error rates (Figure 4) for the *MSB* and *M1* models were essentially the same as the
³³⁷ *Alternating* and *GDD* models, respectively. This lack of improvement from incorporating spatial
³³⁸ variation could be caused either by models not adequately capturing the process driving the spatial
³³⁹ variation, the USA-NPN dataset having biases from variation in sampling effort and/or spatial
³⁴⁰ auto-correlation, or some combination of these factors. Basler (2016) used the *M1* model to predict

³⁴¹ budburst for six species across Europe and found it was generally among the best models in terms
³⁴² of RMSE, albeit never by more than a single day. Their result was strengthened by having a
³⁴³ 40-year time-series across a large region. Chuine and Régnière (2017) listed the incorporation
³⁴⁴ of spatial variation in warming requirements in models as a primary issue in future phenology
³⁴⁵ research. Large-scale phenology datasets, like USA-NPN, will be key in addressing this and other
³⁴⁶ phenological research needs.

³⁴⁷ In addition to exploring differences between phenology datasets, our analyses provide guidance on
³⁴⁸ which models to use when making predictions at a local scale using models built from large-scale
³⁴⁹ data, or vice versa. Among the eight models tested, the *Uniforc* and *GDD* models performed the
³⁵⁰ best overall in the cross dataset comparison in terms of Pearson correlation and RMSE (Appendix
³⁵¹ S2: Figure S5-S6, Table S2). The *GDD* model has one less parameter than the *Uniforc* model, thus
³⁵² the *GDD* model is a suitable choice for making predictions when there is little to no information
³⁵³ at the location of interest (e.g. making phenology forecasts at a new location distant from any
³⁵⁴ observed data). This guidance can vary between species, though, and model testing should still be
³⁵⁵ performed when suitable data are available.

³⁵⁶ In conclusion, our results suggest that both LTER and USA-NPN data provide valuable information
³⁵⁷ on plant phenology. Models built using both data sources yield effective predictions for phenolog-
³⁵⁸ ical events, but parameter estimates from the two data sources differ and models from each source
³⁵⁹ best predict that data source's phenology events. The primary difference in the datasets is spatial
³⁶⁰ scale, but due to trade-offs in data collection efforts, the larger scale USA-NPN data have shorter
³⁶¹ time-series, less site fidelity and other differences from the intensively collected LTER data (Table
³⁶² 3). These differences can be strengths or potential limitations. Observers sampling opportunisti-

363 cally allows the USA-NPN dataset to have a large spatial scale, but also leads to low site fidelity,
364 which limits the ability to measure long-term trends at local scales (Gerst et al., 2016). Tracking
365 long-term trends is the major strength of LTER data, but having a relatively small species pool
366 limits their use in species-level predictive modeling. Due to these differences, the best data source
367 for making predictions depends on the scale at which the predictions are being made. Identifying
368 the most effective data sources for different types and scales of analysis is a useful first step, but
369 the ultimate solution to working with diverse data types is to focus on integrating all types of data
370 into analyses and forecasts (Hanks et al., 2011; Melaas et al., 2016). Our results suggest that meth-
371 ods that can learn from the intensive information available in LTER data in regions where they
372 are available, and simultaneously use large-scale data to capture spatial variation in phenological
373 requirements will help improve our ability to understand and predict phenology. Data integration
374 efforts should also leverage data from remote sensing sources such as the PHENOCAM network or
375 satellite imagery, which have both a large spatial extent and high temporal resolution (Peng et al.,
376 2017; Richardson et al., 2018a,b). Data integration provides the potential to use data from many
377 sources to produce the best opportunity for accurate inference about, and forecasting of, the timing
378 of biological events.

379 Acknowledgments

380 This research was supported by the Gordon and Betty Moore Foundation's Data-Driven Discovery
381 Initiative through Grant GBMF4563 to E.P. White. J.M. Meiners was supported by the Univer-
382 sity of Florida Biodiversity Institute Graduate Research Fellowship. We thank the developers and

383 providers of the data that made this research possible including: the USA National Phenology
384 Network and the many participants who contribute to its Nature's Notebook program; the H.J. An-
385 drews Experimental Forest research program, funded by the National Science Foundation's (NSF)
386 LTER Program (DEB-1440409), US Forest Service Pacific Northwest Research Station, and Ore-
387 gon State University; the Jornada Basin LTER project (NSF Grant DEB-1235828); the Hubbard
388 Brook Experimental Forest, which is operated and maintained by the USDA Forest Service, North-
389 ern Research Station, Newtown Square, PA.; the Harvard Forest LTER; and the PRISM Climate
390 Group at Oregon State University.

391 References

- 392 Bailey, A. (2018). Hubbard brook experimental forest (us forest service): Rou-
393 tine seasonal phenology measurements, 1989 - present. environmental data initiative.
394 <https://doi.org/10.6073/pasta/765084e2b4a5ec389403238c58784572>.
- 395 Basler, D. (2016). Evaluating phenological models for the prediction of leaf-out dates in six tem-
396 perate tree species across central europe. *Agricultural and Forest Meteorology*, 217:10–21.
- 397 Beatley, J. C. (1974). Effects of rainfall and temperature on the distribution and behavior of larrea
398 tridentata (creosote-bush) in the mojave desert of nevada. *Ecology*, 55(2):245–261.
- 399 Blümel, K. and Chmielewski, F. M. (2012). Shortcomings of classical phenological forcing models
400 and a way to overcome them. *Agricultural and Forest Meteorology*, 164:10–19.
- 401 Cannell, M. G. R. and Smith, R. I. (1983). Thermal time, chill days and prediction of budburst in
402 picea sitchensis. *The Journal of Applied Ecology*, 20(3):951.

- 403 CaraDonna, P. J., Iler, A. M., and Inouye, D. W. (2014). Shifts in flowering phenology reshape a
404 subalpine plant community. *Proceedings of the National Academy of Sciences*, 111(13):4916–
405 4921.
- 406 Chuine, I. (2000). A unified model for budburst of trees. *Journal of Theoretical Biology*,
407 207(3):337–347.
- 408 Chuine, I., Bonhomme, M., Legave, J. M., García de Cortázar-Atauri, I., Charrier, G., Lacointe,
409 A., and Améglio, T. (2016). Can phenological models predict tree phenology accurately in the
410 future? the unrevealed hurdle of endodormancy break. *Global Change Biology*, 22(10):3444–
411 3460.
- 412 Chuine, I., de Cortazar-Atauri, I. G., Kramer, K., and Hänninen, H. (2013). Plant development
413 models. In Schwartz, M. D., editor, *Phenology: An Integrative Environmental Science*, pages
414 275–293. Springer Netherlands, Dordrecht.
- 415 Chuine, I. and Régnière, J. (2017). Process-Based Models of Phenology for Plants and Animals.
416 *Annual Review of Ecology, Evolution, and Systematics*, 48(1):159–182.
- 417 Cleland, E., Chuine, I., Menzel, A., Mooney, H., and Schwartz, M. (2007). Shifting plant phenol-
418 ogy in response to global change. *Trends in Ecology Evolution*, 22(7):357–365.
- 419 Cook, B. I., Wolkovich, E. M., and Parmesan, C. (2012). Divergent responses to spring and winter
420 warming drive community level flowering trends. *Proceedings of the National Academy of
421 Sciences*, 109(23):9000–9005.
- 422 Crimmins, T. M., Crimmins, M. A., Gerst, K. L., Rosemartin, A. H., and Weltzin, J. F. (2017).

- 423 Usa national phenology network's volunteer-contributed observations yield predictive models
424 of phenological transitions. *PLOS ONE*, 12(8):e0182919.
- 425 Dalcin, L. D., Paz, R. R., Kler, P. A., and Cosimo, A. (2011). Parallel distributed computing using
426 python. *Advances in Water Resources*, 34(9):1124–1139.
- 427 Denny, E. G., Gerst, K. L., Miller-Rushing, A. J., Tierney, G. L., Crimmins, T. M., Enquist, C.
428 A. F., Guertin, P., Rosemartin, A. H., Schwartz, M. D., Thomas, K. A., and Weltzin, J. F. (2014).
429 Standardized phenology monitoring methods to track plant and animal activity for science and
430 resource management applications. *International Journal of Biometeorology*, 58(4):591–601.
- 431 Dickinson, J., Zuckerberg, B., and Bonter, D. (2010). Citizen science as an ecological research
432 tool: Challenges and benefits. *Annual Review of Ecology, Evolution and Systematics*, 41(1):149–
433 172.
- 434 Diez, J. M., Ibáñez, I., Miller-Rushing, A. J., Mazer, S. J., Crimmins, T. M., Crimmins, M. A.,
435 Bertelsen, C. D., and Inouye, D. W. (2012). Forecasting phenology: from species variability to
436 community patterns. *Ecology Letters*, 15(6):545–553.
- 437 Feldman, R. E., Žemaitė, I., and Miller-Rushing, A. J. (2018). How training citizen scientists af-
438 fects the accuracy and precision of phenological data. *International Journal of Biometeorology*,
439 62(8):1421–1435.
- 440 Fuccillo, K. K., Crimmins, T. M., de Rivera, C. E., and Elder, T. S. (2015). Assessing accuracy
441 in citizen science-based plant phenology monitoring. *International Journal of Biometeorology*,
442 59(7):917–926.
- 443 García-Mozo, H., Chuine, I., Aira, M., Belmonte, J., Bermejo, D., Díaz de la Guardia, C., Elvira,

- 444 B., Gutiérrez, M., Rodríguez-Rajo, J., Ruiz, L., Trigo, M., Tormo, R., Valencia, R., and Galán,
445 C. (2008). Regional phenological models for forecasting the start and peak of the quercus pollen
446 season in spain. *Agricultural and Forest Meteorology*, 148(3):372–380.
- 447 Gerst, K. L., Kellermann, J. L., Enquist, C. A. F., Rosemartin, A. H., and Denny, E. G. (2016).
448 Estimating the onset of spring from a complex phenology database: trade-offs across geographic
449 scales. *International Journal of Biometeorology*, 60(3):391–400.
- 450 Gerst, K. L., Rossington, N. L., and Mazer, S. J. (2017). Phenological responsiveness to climate
451 differs among four species of quercus in north america. *Journal of Ecology*, 38(1):42–49.
- 452 Grolemund, G. and Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of*
453 *Statistical Software*, 40(3):1–25.
- 454 Hanks, E. M., Hooten, M. B., and Baker, F. A. (2011). Reconciling multiple data sources to
455 improve accuracy of large-scale prediction of forest disease incidence. *Ecological Applications*,
456 21(4):1173–1188.
- 457 Hart, E. M. and Bell, K. (2015). prism: Download data from the oregon prism project.
458 <http://github.com/ropensci/prism>.
- 459 Hijmans, R. J. (2017). raster: Geographic data analysis and modeling. r package version 2.6-7.
460 <https://CRAN.R-project.org/package=raster>.
- 461 Hunter, A. F. and Lechowicz, M. J. (1992). Predicting the Timing of Budburst in Temperate Trees.
462 *The Journal of Applied Ecology*, 29(3):597.
- 463 Iler, A. M., Høye, T. T., Inouye, D. W., and Schmidt, N. M. (2013). Nonlinear flowering re-

- 464 sponses to climate: are species approaching their limits of phenological change? *Philosophical*
465 *Transactions of the Royal Society of London*, 368(1624):20120489.
- 466 Jeong, S.-J., Medvigy, D., Shevliakova, E., and Malyshev, S. (2013). Predicting changes in tem-
467 perate forest budburst using continental-scale observations and models. *Geophysical Research*
468 *Letters*, 40(2):359–364.
- 469 Jones, E., Oliphant, T., Peterson, P., and Others (2001). Scipy: Open source scientific tools for
470 python. <http://www.scipy.org/>.
- 471 Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker,
472 G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*,
473 59(7):613–620.
- 474 Kramer, K. (1995). Phenotypic plasticity of the phenology of seven european tree species in
475 relation to climatic warming. *Plant, Cell and Environment*, 18(2):93–104.
- 476 McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the*
477 *9th Python in Science Conference*, pages 51–56.
- 478 Melaas, E. K., Friedl, M. A., and Richardson, A. D. (2016). Multiscale modeling of spring phenol-
479 ogy across deciduous forests in the eastern united states. *Global Change Biology*, 22(2):792–
480 805.
- 481 Ogilvie, J. E. and Forrest, J. R. (2017). Interactions between bee foraging and floral resource
482 phenology shape bee populations and communities. *Current Opinion in Insect Science*, 21:75–
483 82.
- 484 Ogilvie, J. E., Griffin, S. R., Gezon, Z. J., Inouye, B. D., Underwood, N., Inouye, D. W., and Irwin,

- 485 R. E. (2017). Interannual bumble bee abundance is driven by indirect climate effects on floral
486 resource phenology. *Ecology Letters*, 20(12):1507–1515.
- 487 O'Keefe, J. (2015). Phenology of woody species at harvard forest since 1990. harvard forest data
488 archive: Hf003.
- 489 Oliphant, T. (2006). A guide to numpy. USA: Trelgol Publishing.
- 490 Olsson, C. and Jönsson, A. M. (2014). Process-based models not always better than empirical
491 models for simulating budburst of norway spruce and birch in europe. *Global Change Biology*,
492 20(11):3492–3507.
- 493 Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*,
494 5(2):9–13.
- 495 Peng, D., Wu, C., Li, C., Zhang, X., Liu, Z., Ye, H., Luo, S., Liu, X., Hu, Y., and Fang, B. (2017).
496 Spring green-up phenology products derived from MODIS NDVI and EVI: Intercomparison,
497 interpretation and validation using National Phenology Network and AmeriFlux observations.
498 *Ecological Indicators*, 77:323–336.
- 499 PRISM Climate Group (2004). Oregon state university. <http://prism.oregonstate.edu>.
- 500 Python Software Foundation (2018). Python Language Reference Manual, version 3.6.
501 <http://www.python.org>.
- 502 R Core Team (2017). R: a language and environment for statistical computing.
- 503 Réaumur, R. (1735). Observations du thermomètres, faites a Paris pendant l'année 1735, com-

504 parées avec celles qui ont été faites sous la ligne, a l'isle de France, a Alger et quelques unes de
505 nos isles de l'Amérique. *Mem Paris Acad Sci*, 1735(545).

506 Richardson, A. D., Anderson, R. S., Arain, M. A., Barr, A. G., Bohrer, G., Chen, G., Chen, J. M.,
507 Ciais, P., Davis, K. J., Desai, A. R., Dietze, M. C., Dragoni, D., Garrity, S. R., Gough, C. M.,
508 Grant, R., Hollinger, D. Y., Margolis, H. A., McCaughey, H., Migliavacca, M., Monson, R. K.,
509 Munger, J. W., Poulter, B., Racza, B. M., Ricciuto, D. M., Sahoo, A. K., Schaefer, K., Tian,
510 H., Vargas, R., Verbeeck, H., Xiao, J., and Xue, Y. (2012). Terrestrial biosphere models need
511 better representation of vegetation phenology: results from the north american carbon program
512 site synthesis. *Global Change Biology*, 18(2):566–584.

513 Richardson, A. D., Hufkens, K., Milliman, T., Aubrecht, D. M., Chen, M., Gray, J. M., Johnston,
514 M. R., Keenan, T. F., Klosterman, S. T., Kosmala, M., Melaas, E. K., Friedl, M. A., and Frol-
515 king, S. (2018a). Tracking vegetation phenology across diverse north american biomes using
516 phenocam imagery. *Scientific Data*, 5:180028.

517 Richardson, A. D., Hufkens, K., Milliman, T., and Frolking, S. (2018b). Intercomparison of phe-
518 nological transition dates derived from the PhenoCam Dataset V1.0 and MODIS satellite remote
519 sensing. *Scientific Reports*, 8(1):5679.

520 Richardson, A. D., Keenan, T. F., Migliavacca, M., Ryu, Y., Sonnentag, O., and Toomey, M.
521 (2013). Climate change, phenology, and phenological control of vegetation feedbacks to the
522 climate system. *Agricultural and Forest Meteorology*, 169:156–173.

523 Roberts, A. M. I., Tansey, C., Smithers, R. J., and Phillimore, A. B. (2015). Predicting a change in
524 the order of spring phenology in temperate forests. *Global Change Biology*, 21(7):2603–2611.

- 525 Schulze, M. D. (2017). Vegetative phenology observations at the andrews experimental for-
526 est, 2009 - present. long-term ecological research. forest science data bank. corvallis, or.
527 <http://andlter.forestry.oregonstate.edu/data/abstract.aspx?dbcode=TV075>.
- 528 Schwartz, M. D., Betancourt, J. L., and Weltzin, J. F. (2012). From caprio's lilacs to the usa
529 national phenology network. *Frontiers in Ecology and the Environment*, 10(6):324–327.
- 530 Storn, R. and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global
531 optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359.
- 532 Tang, J., Körner, C., Muraoka, H., Piao, S., Shen, M., Thackeray, S. J., and Yang, X. (2016).
533 Emerging opportunities and challenges in phenology: a review. *Ecosphere*, 7(8):e01436.
- 534 Theobald, E. J., Breckheimer, I., and HilleRisLambers, J. (2017). Climate drives phenological
535 reassembly of a mountain wildflower meadow community. *Ecology*, 98(11):2799–2812.
- 536 Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J., and Martin, T. G. (2013). Realising
537 the full potential of citizen science monitoring programs. *Biological Conservation*, 165:128–
538 138.
- 539 Tylianakis, J. M., Didham, R. K., Bascompte, J., and Wardle, D. A. (2008). Global change and
540 species interactions in terrestrial ecosystems. *Ecology Letters*, 11(12):1351–1363.
- 541 USA National Phenology Network (2017). Plant and animal phenology data. data type: Sta-
542 tus and intensity. 01/01/2009-04/31/2017 for region: 49.9375, -66.4791667 (ur); 24.0625,
543 -125.0208333 (ll). USA-NPN, Tucson, Arizona, USA. Data set accessed 04/20/2017 at
544 <http://doi.org/10.5066/F78S4N1V>.

- 545 Wang, J. Y. (1960). A Critique of the Heat Unit Approach to Plant Response Studies. *Ecology*,
- 546 41(4):785–790.
- 547 Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- 548 Wickham, H., Francois, R., Henry, L., and Müller, K. (2017). *dplyr: A grammar of data manipu-*
- 549 *lation.*
- 550 Wickham, H. and Henry, L. (2018). *tidyr: Easily Tidy Data with 'spread()' and 'gather()'* Func-
- 551 *tions.*
- 552 Wolkovich, E. M., Cook, B. I., Allen, J. M., Crimmins, T. M., Betancourt, J. L., Travers, S. E., Pau,
- 553 S., Regetz, J., Davies, T. J., Kraft, N. J. B., Ault, T. R., Bolmgren, K., Mazer, S. J., McCabe,
- 554 G. J., McGill, B. J., Parmesan, C., Salamin, N., Schwartz, M. D., and Cleland, E. E. (2012).
- 555 Warming experiments underpredict plant phenological responses to climate change. *Nature*,
- 556 485(7399):494–497.
- 557 Xu, L. and Chen, X. (2013). Regional unified model-based leaf unfolding prediction from 1960 to
- 558 2009 across northern china. *Global Change Biology*, 19(4):1275–1284.
- 559 Zhang, H., Liu, S., Regnier, P., and Yuan, W. (2017). New insights on plant phenological response
- 560 to temperature revealed from long-term widespread observations in china. *Global Change Biol-*
- 561 *ogy*, 12(10):3218–3221.

Dataset Name	Habitat	Phenological Event (Num. Species)	Reference
562	Harvard Forest	Budburst (17) Flowering (7)	(O'Keefe, 2015)
	Jornada Experimental Range	Flowering (2)	
	H.J. Andrews Experimental Forest	Budburst (5) Flowering (4)	(Schulze, 2017)
	Hubbard Brook	Budburst (3)	(Bailey, 2018)

563 **Table 1:** LTER datasets used in the analysis

Name	DOY Estimator	Forcing Equations	Total Parameters	Reference
564	Naive	\overline{DOY}	-	-
	Linear	$DOY = \beta_1 + \beta_2 T_{mean}$	-	-
	GDD	$\sum_{t=t_1}^{DOY} R_f(T_i) \geq F^*$	$R_f(T_i) = \max(T_i - T_{base}, 0)$	3 (Réaumur, 1735; Wang, 1960; Hunter and Lechowicz, 1992)
	Fixed GDD	$\sum_{t=1}^{DOY} R_f(T_i) \geq F^*$	$R_f(T_i) = \max(T_i, 0)$	1 (Réaumur, 1735; Wang, 1960; Hunter and Lechowicz, 1992)
	Alternating	$\sum_{t=1}^{DOY} R_f(T_i) \geq a + b e^{c NCD(t)}$	$R_f(T_i) = \max(T_i - 5, 0)$	3 (Cannell and Smith, 1983)
	Uniforc	$\sum_{t=t_1}^{DOY} R_f(T_i) \geq F^*$	$R_f(T_i) = \frac{1}{1+e^{b(T_i-c)}}$	4 (Chuine, 2000)
	M1	$\sum_{t=t_1}^{DOY} R_f(T_i) \geq (\frac{L_i}{24})^k F^*$	$R_f(T_i) = \max(T_i - T_{base}, 5)$	4 (Blümel and Chmielewski, 2012)
	MSB	$\sum_{t=1}^{DOY} R_f(T_i) \geq a + b e^{c NCD_i} + d T_{mean}$	$R_f(T_i) = \max(T_i - 5, 0)$	4 (Jeong et al., 2013)

564 **Table 2:** Phenology models used in the analysis. For all models, except the Naive and Linear models, the daily mean temperature T_i is first
 565 transformed via the specified forcing equation. The cumulative sum of forcing is then calculated from a specific start date (either $DOY = 1$ or
 566 using the fitted parameter t_1). The phenological event is estimated as the DOY in which cumulative forcing is greater than or equal to the specified
 567 total required forcing (either F^* or the specified equation). Parameters for each model are as follows: For the Naive model \overline{DOY} is the mean day
 568 of year (ie. the Julian date) of a phenological event; for the Linear model β_1 and β_2 are the intercept and slope, respectively and T_{mean} is the
 569 average daily temperature between January 1 and March 31; for the GDD model F^* is the total accumulated forcing required, t_1 is the start date of
 570 forcing accumulation, and T_{base} is the threshold daily mean temperature above which forcing accumulates; for the Fixed GDD model F^* is the total
 571 accumulated forcing required; for the Alternating model NCD is the number of chill days (daily mean temperature below 0°C) from $DOY = 0$ to
 572 the DOY of the phenological event, a , b , and c are the three fitted model coefficients; for the Uniforc model, is F^* is the total accumulated forcing
 573 required, t_1 is the start date of forcing accumulation, and b and c are two additional fitted parameters which define the sigmoid function; the M1
 574 model is the same as the GDD model, but with the additional fitted parameter k that adjusts the total forcing accumulation according to day length;
 575 the MSB model is the same as the Alternating model, but with the additional fitted parameter d to correct the model according to mean spring
 576 temperature.

	<u>LTER</u>	<u>USA-NPN</u>
Time-series length	High	Low
Spatial extent	Low	High
579 Local species representation	High	Low
Regional/Continental species representation	Low	High
Number of observers	Low	High
Site fidelity	High	Low

580 **Table 3:** Attributes of the two datasets used in this study. Bold text indicates an attribute is expected to increase over time.

581 **Figure 1:** Locations of U.S.A. National Phenology Network sites used (black points) and Long Term Ecological Research sites (labeled circles),
582 with greyscale showing elevation.

583 **Figure 2:** Comparisons of parameter estimates between USA-NPN and LTER derived models. Each point represents a parameter value for a
584 specific species and phenophase, and is the mean value from 250 bootstrap iterations. The black line is the 1:1 line. The R^2 is the coefficient
585 of determination, which can be negative if the relationship between the two parameter sets is worse than no relationship but with the same mean
586 values.

587 **Figure 3:** Comparison of predicted day of year (DOY) of all phenological events between USA-NPN and LTER-derived models. Top panels show
588 comparisons at LTER sites and bottom panels show comparisons at USA-NPN sites. Each point is an estimate for a single held-out observation.
589 Colors indicate observations for a single species and phenophase combination.

590 **Figure 4:** Differences in prediction error between USA-NPN and LTER-derived models. Density plots for comparisons of predictions on LTER data
591 (top row) and USA-NPN data (bottom row). Each plot represents the difference between the RMSE for LTER-derived model and the USA-NPN
592 derived model, meaning that values less than zero indicate more accurate prediction by LTER-derived models and values greater than zero indicate
593 more accurate prediction by NPN-derived models. $p < 0.001$ for all t-tests. Differences are calculated pairwise for the 38 species/phenophase
594 comparisons.

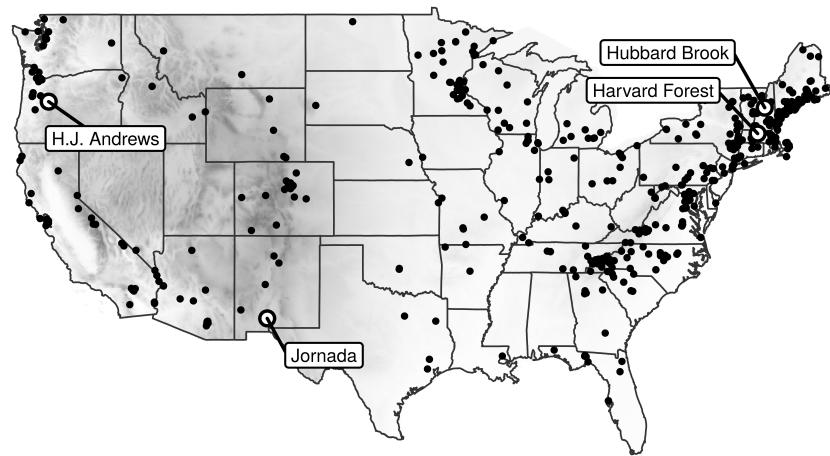


Figure 1: .

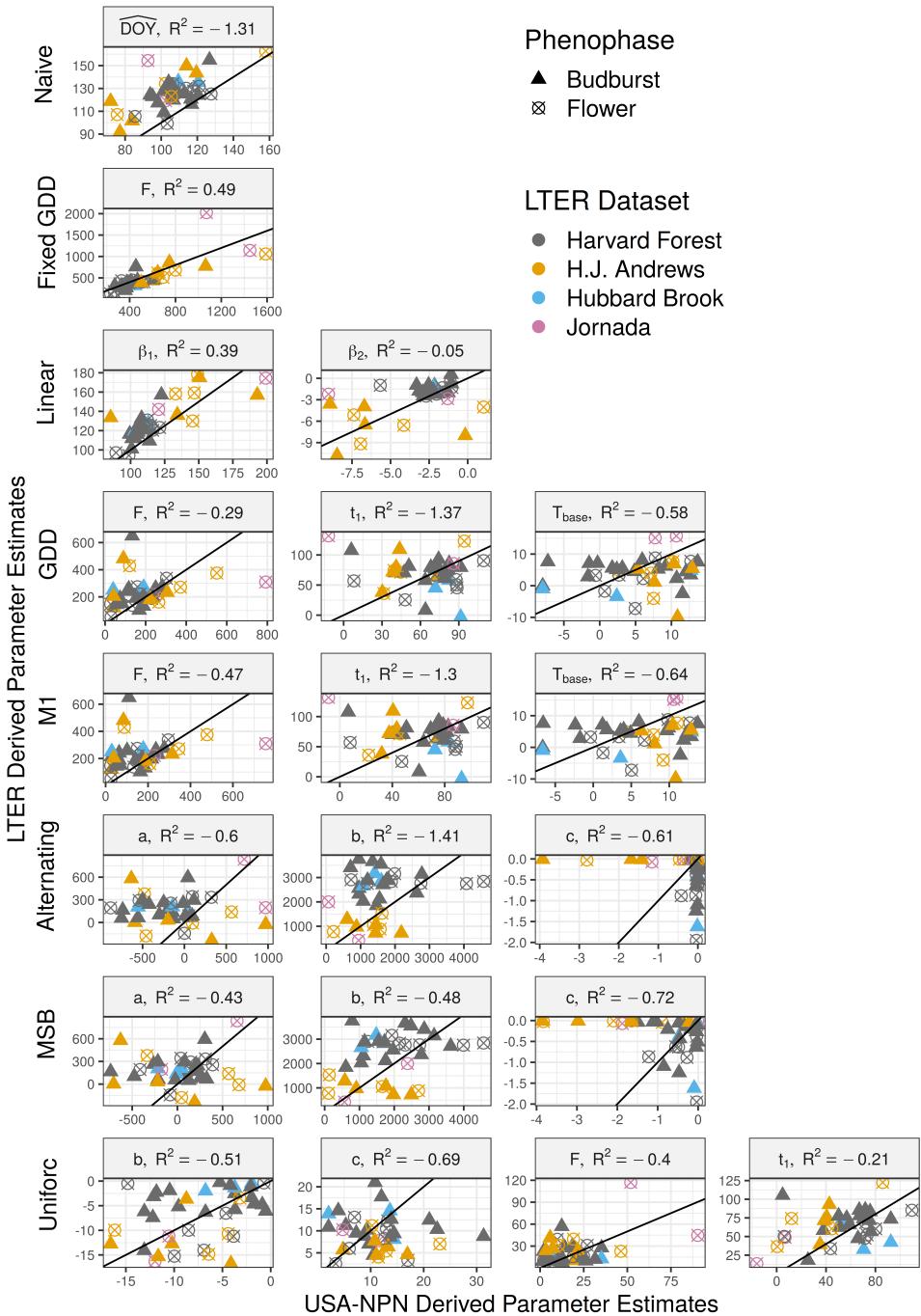


Figure 2: .

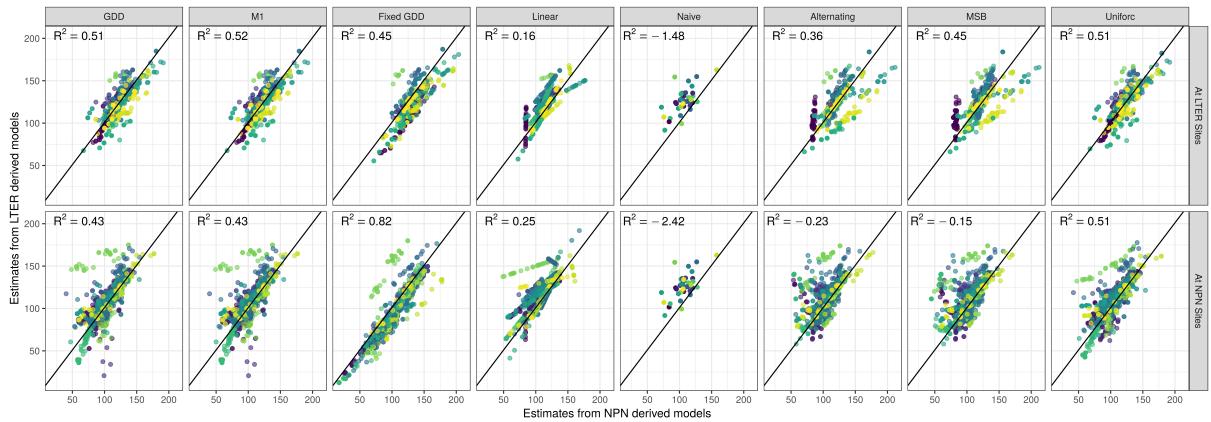


Figure 3: .

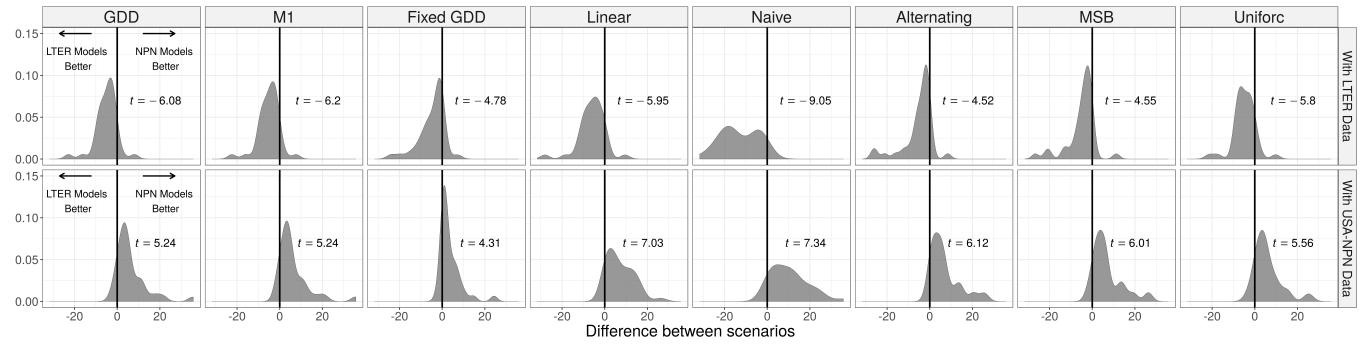


Figure 4: .