BloomBench: A Multi-Species Benchmark for Evaluating the Generalization of Fruit Tree Phenology Models

Ron van Bree [1], **Diego Marcos** [2], **Ioannis N. Athanasiadis** [1]

[1]Artificial Intelligence Group, Wageningen University & Research,
[2]Inria, University of Montpellier
ron.vanbree@wur.nl, diego.marcos@inria.fr, ioannis.athanasiadis@wur.nl

Abstract

The timing of phenological events in trees is incredibly important for understanding a wide range of secondary effects, such as the susceptibility of orchard yields to environmental stressors and the phenological timing of adjacent ecosystems. Tree phenology is strongly driven by temperature, and (agro-)ecologists typically use biophysical thermal-time models to relate changes in temperature to the timing of observed events. Mechanistic models, however, show large discrepancies since these dynamics are difficult to capture in simple equations. With the improved quality and quantity of data on plant phenology, this has popularized the use of machine learning methods for this purpose. Existing works, however, are evaluated for different species and specific regions, making inter-comparisons challenging. We provide the first benchmark covering different species, cultivars and climates for evaluating models that predict the timing of crop phenophases. We have compiled a consistent set of datasets linking climatic drivers with the timing of flowering in fruit trees. With this benchmark we (i) provide consistent model evaluation on datasets with different characteristics (e.g. size, cultivar information, observation trends, climate gradient) thus highlighting model strengths and weaknesses, (ii) provide a real multi-faceted use case for evaluating machine learning methods that focus on different types of domain shifts, (iii) accelerate ML research in this domain by facilitating a publicly available, ready-to-use dataset.

Introduction

Tree phenology, the study of seasonal life cycle events such as flowering, fruiting, and dormancy, is a critical area of research for understanding plant responses to environmental changes and optimizing agricultural practices. Deciduous trees face evolutionary pressure to optimize their dormancy release, balancing the benefits of an earlier release (and thus longer growing season) with the risks of frost damage (Vitasse et al. 2013). Understanding these dynamics is also of great importance to agriculture, where frost damage to deciduous fruit trees can cause significant yield losses or even failure (Baldocchi and Wong 2008; Luedeling 2012). A single frost event in western Europe in April 2021 was responsible for an estimated yield loss of up to 50% for various fruits (Vautard et al. 2022). The susceptibility of crops to environmental stressors is strongly dependent on their phenology which makes it an important component of models forecasting yields. Prospects of temperature increase and more frequent occurrence of extreme weather events makes the understanding of these processes even more pressing (Luedeling 2012; Lamichhane 2021). Despite many works estimating the impacts on a changing climate on fruit tree phenology and its implications on their productivity (Else and Atkinson 2010; Luedeling et al. 2011), there remains a large uncertainty and disagreement between predictions (Wang et al. 2020; Fernandez, Whitney, and Luedeling 2020). While works generally agree temperature is the main driver of these dynamics, the difficulty arises from quantifying temperature effects to phenological development and adjust these to individual locations, species or cultivars.

The rising availability of data on phenological events and climate has sparked an increased interest for using machine learning methods for these purposes. Temperature effects on phenological progression are known to be more complex than biophysical mechanistic models can capture and might thus be more suited for modeling using a data-driven approach. Works on machine learning for predicting the timing of phenological events, however, are fragmented in their case studies and the considered crops and location, highlighting the need for a common ground of evaluation for proper inter-comparison. Existing machine learning benchmarks on plant phenology focus on sematic segmentation of imagery or phenophase detection (Weyler et al. 2024; Seyednasrollah et al. 2019), as most machine learning research in plant phenology revolves around this task (Katal et al. 2022). The potential for monitoring of plant phenology at different scales is widely recognized (Piao et al. 2019), giving prospects for more data collection in the future. However, there is a lack of benchmarks for predicting the timing of phenological events based on climatic drivers.

For this purpose, we introduce BloomBench: a benchmark for machine learning models predicting the timing of phenological events of fruit trees. BloomBench is composed of a list of curated datasets with different characteristics and

Accepted at the First International Workshop on AI in Agriculture (Agri AI), co-located with AAAI 2026.

provides a way forward for inter-comparing machine learning methods on their strengths and weaknesses in various settings. Datasets differ in terms of size, spatial coverage, diversity in climate, species included, species subgroup/cultivar information, observation uncertainty and presence of observation trends, thus presenting a multi-faceted challenge. A set of commonly used ML methods for regression with time series features has been evaluated, forming a baseline for newly introduced models. All code to reproduce the benchmark and results in this paper can be found at https://github.com/WUR-AI/BloomBench.

The first reports of explaining the variation in phenological events based on temperature were made by (Reaumur 1735), introducing the concept of thermal time models. These models weight temperature levels and accumulate them until some threshold until the considered phenophase is complete. What followed was a long history of mechanistic phenology models that use the same general concept but apply different temperature weightings. The field is still active and (Chuine et al. 2025) provide an extensive overview of this period. In this section we report machine learning methods applied to this problem with a focus on fruit trees.

A popular region for phenology research is Japan, since the cultural significance of cherry tree flowering has enabled data collection in a wide variety of climates and a long time-span (Aono and Kazui 2008). (Masago and Lian 2022) evaluate different machine learning methods in their ability to predict cherry tree (Prunus ×*yedoensis*) flowering dates, with gradient boosted decision trees providing the closest estimates. (Oses et al. 2020) do a comparison of methods, but evaluate them using a case study on olive trees in Italy. (Nappa et al. 2024) use bayesian neural networks to provide uncertainty estimates while predicting phenological stages of olive trees. For a single location data availability is usually limited, thus hindering the use of machine learning models in these settings. (Saxena et al. 2023a) show that multi-task learning over multiple species can help overcome this when applying a recurrent neural network for budbreak prediction in grapevines and later also apply this to cold-hardiness (Saxena et al. 2023b). (Nagai, Morimoto, and Saitoh 2020) investigate this same issue but for cherry trees (Prunus ×*yedoensis*) in Japan and propose a solution using self-organizing maps. (van Bree, Marcos, and Athanasiadis 2025) constrain a neural network based on mechanistic models for cherry tree flowering and compare their dynamics and generalization capacity. (Shin et al. 2025) instead use a Bayesian state space model to predict different stages of flowering of cherry trees (Prunus *itosakura*) from 1924 to 2024 in Japan.

Several datasets exist for the development and validation of models related to plant phenology, although with different setups in mind. PhenoBench (Weyler et al. 2024) provides detailed pixel-wise annotations of crops and weeds for segmenting high-resolution images of agricultural fields. Similarly, VegAnn (Madec et al. 2023) provide annotated images for segmentation of agricultural fields but without a fixed camera perspective, resulting in a less controlled and more challenging benchmark. The PhenoCam (Seyednasrollah et al. 2019) network also connects computer vision and phenology by providing a collection of cameras monitoring vegetation globally. (Mori, Doi, and Iizumi 2023) use a biophysical phenology model to provide a global estimate of the timing of phenological events of various annual crop species. For this application, however, the choice of phenology model and parameterization introduces large uncertainty to these estimates (Fernandez, Whitney, and Luedeling 2020). Moreover, for annual crops the change in developed cultivars needs to be taken into account (Rezaei et al. 2018). To our knowledge, the annual cherry tree flowering competition hosted by the George Mason University (GMU) department of statistics (Auerbach, Kepplinger, and Wolkovich 2022) for predicting the upcoming season's flowering dates is the nearest equivalent to a benchmark for statistical methods for phenology modeling. We extend upon a subset of this data, that was originally collected by the meteorological agencies in Japan, Switzerland and South Korea, and integrate more data sources, tree species, and spatial coverage.

Benchmark

Task

BloomBench contains a collection of datasets for supervised learning of tree phenology, linking climate reanalysis models to large collections of phenological observations. A collection of datasets has been defined and the benchmark is set up to easily integrate more data sources. Each dataset is a set of N tuples $\{(\mathbf{X}_n, \mathbf{Y}_n)\}_N$, where $\mathbf{X}_n$ is a set of features and $\mathbf{Y}_n$ a corresponding set of targets (i.e. phenological observations). Each tuple corresponds to one observed tree in one season and is uniquely characterized by the following tuple (whose elements are contained in $\mathbf{X}_n$): (Data Source ID, Year, Location ID, Species ID, Subgroup ID). Each element of $\mathbf{Y}_n$ is a date corresponding to an observed phenological event defined by the widely used BBCH scale (Meier et al. 1997). A complete overview of a data sample and its included features/attributes is shown in Table 1.

Target Although multiple phenological events can be provided in $\mathbf{Y}_n$, we decided to initially focus the benchmark on predicting a single event, namely the onset of flowering, due to its data presence in the available data sources and great importance to agriculture. It is important to note that observations can refer to different stages within the event of flowering, and the target for each dataset can be found in Table 2. Other observations have not been omitted, as we would like to highlight the availability of this data and encourage research in this direction. The BloomBench has been set up as a regression task in which the date of flowering needs to be predicted for all samples in the contained datasets, as evaluated by the resulting mean absolute error. Models are evaluated per dataset, as they have different characteristics in terms of size, included species, information about cultivars, spatial extent, presence of trends, and variation in climate.

Features Standard climatic input variables associated with phenological timing are provided, namely daily mean temperature levels and total day length. These are provided from some start (SOS) to end (EOS) of season in the respective

Table 1: Overview of what is included in a single sample in the dataset. A sample consists of a `dict` containing different attributes. The table lists the keys (in the "Key" column) together with description and properties of the corresponding value. Some attributes (e.g. climate variables) are grouped in a `dict`, whose structure is presented as well. "In_Index" indicates whether an attribute is contained in the 5-tuple that uniquely identifies a data sample.

Attribute	Key	In_Index	Type
Data Source	`src`	Y	`string`
Season Year	`year`	Y	`int`
Location ID	`loc_id`	Y	`int`
Species ID	`species_id`	Y	`int`
Subgroup ID	`subgroup_id`	Y	`int`
Latitude (WGS84)	`lat`	N	`float`
Longitude (WGS84)	`lon`	N	`float`
Season Start Date	`season_start`	N	`datetime`
Season End Date	`season_end`	N	`datetime`
Climate Variables	`features`	N	`dict`
↪ *variable*	*variable name*		`ndarray[float32]`
Observations	`observations`	N	`dict`
↪ *observation*	*observation code*		`datetime`
Observations As Index	`observations_index`	N	`dict`
↪ *observation*	*observation code*		`int`

year, where EOS is included in the series. These dates can be configured as species-specific properties through a crop calendar but have all been set to October 1st (SOS) and a season length of 365 days (thus making EOS dependent on leap years). Meteorological data was obtained from the ERA5 climate reanalysis.

Evaluation To test the generalization capabilities of the models they are evaluated on a held-out dataset in unobserved years. More specifically, the first 75% of years in the complete time span of the dataset are considered training data and the remaining years are used for evaluation. Since predictions should be made for the future, the phenological trends (that are clearly present in the dataset) should be captured correctly.

Data Processing Pipeline

BloomBench is a framework for connecting sources of phenological observations, transforming them to a common format and subsequently pairing them with climatic variables. Phenophase observations were obtained from two data sources, namely the Pan Europen Phenology Project (PEP725) (Templ et al. 2018) and a dataset on cherry tree flowering compiled by the George Mason University (GMU) Department of Statistics (Auerbach, Kepplinger, and Wolkovich 2022). Data obtained from PEP725 (Templ et al. 2018) was sanitized, as we found location names including commas prevented csv files from being used directly. Outliers were removed by only considering the 1-99 percentile range of observations. Data obtained from the GMU dataset were paired with tree species information reported by their original data sources (i.e. the respective meteorological agency). All tree species are associated with a fixed time window that covers a single season. The homogenized phenophase observations were subsequently paired with ERA5 climate reanalysis data spanning this season (daily mean 2m temperature levels and daylength).

Datasets

Based on the available data a collection of datasets were defined. A complete list of compiled datasets included in BloomBench can be found in Table 2. Datasets differ in terms of size, spatial coverage, diversity in climate, species included, provided information about possible species subgroups/cultivars, observation uncertainty and presence of observation trends. Since some datasets are spatially quite dense in their observations it allowed us to estimate the inherent uncertainty of the observations by considering the variability of observed events in close proximity. That is, the variation that cannot be explained by the coarse gridded climate reanalysis data. Table 2 lists, for sufficiently dense datasets, the average standard deviation of observed events (in days) in some year within some (0.5°×0.5°) cell, considering the cell contained multiple observations.

Baseline models

Various machine learning models have been implemented and evaluated to form a baseline for newly introduced models. We put a focus on basic popular ML methods for regression with time series features and include some trivial models as well. All models were fit for five random seeds controlling the model initialization and fitting procedure. Resulting mean absolute error (MAE) with corresponding stardard deviations are reported in Table 3. Despite setting a solid, informative benchmark, we expect there to be ample room for improvement. We reflect on this in Section .

Baseline models include: **Mean** (i.e. simply take the mean of the observed dates, where means are aggregated per species), **Trend** - fit a linear trend on the mean observed dates per year, Random Forest (**RF**), Gradient Boosting Decision Tree (**GBDT**), **AdaBoost**, Convolutional Neural Network (**CNN**), Gated Recurrent Unit Network (**GRU**) and a

Table 2: This table summarizes all datasets currently included in BloomBench. The dataset name ("Key") is generally composed of the original source of data and the tree species common name. The "Observed Events" column lists the codes of the observed phenophases, with the dataset-specific target event underlined. Observed events include BBCH 60 (beginning of flowering), 65 (full flowering), 69 (end of flowering), 86 (first ripe fruits), 87 (fruits ripe for picking). Data obtained through the GMU repository did not adhere the the BBCH scale. We have labeled them as "gmu_x", where 0 refers to peak flowering, 1 to 25% of buds flowering, 2 to first flowering. "Size" refers to the total number of unique observations of the target variable included in the dataset, so other observed events are not considered. "SD" quantifies the uncertainty in the made observations by providing standard deviation estimates (as described in Section). *Prunus ×yedoensis, Prunus sargentii, Prunus campanulata, ** Likely Prunus ×yedoensis

#	Key	Species	Observed Events	Size	SD
1	PEP725_Apple	Malus ×domestica	BBCH 60, 69, 87	17791	2.99
2	PEP725_Pear	Pyrus communis	BBCH 60, 65, 69, 87	6914	3.14
3	PEP725_Peach	Prunus persica	BBCH 60	5332	4.07
4	PEP725_Almond	Prunus amygdalis	BBCH 60, 65, 69, 87	187	-
5	PEP725_Hazel	Corylus avellana	BBCH 60, 86	29361	7.91
6	PEP725_Apricot	Prunus armeniaca	BBCH 60, 87	283	-
7	PEP725_Plum	Prunus domestica	BBCH 60, 65, 69, 87	11563	4.11
8	PEP725_Blackthorn	Prunus spinosa	BBCH 60	26045	4.64
9	PEP725_Cherry	Prunus avium	BBCH 60, 65, 69, 87	17799	3.17
10	GMU_Cherry_JPN	*Multiple**	gmu_0	3033	-
11	GMU_Cherry_CHE	Prunus avium	gmu_1	2581	-
12	GMU_Cherry_KOR	*Unknown***	gmu_2	974	-
				121863	

Table 3: Mean absolute error (MAE) and standard deviation (SD) for all baseline models on all benchmark datasets (over five uniformly random selected seeds between 0 and 100). Lowest MAE scores are marked in a bold font.

	MAE±SD (Test)			
#	Mean	Trend	RF	GBDT
1	7.64 ± 0.00	7.14 ± 0.00	6.03 ± 0.04	5.93 ± 0.08
2	8.83 ± 0.00	10.11 ± 0.00	7.65 ± 0.11	6.59 ± 0.04
3	24.62 ± 0.00	19.25 ± 0.00	10.90 ± 0.29	10.25 ± 0.25
4	10.69 ± 0.00	11.35 ± 0.00	10.19 ± 0.09	11.00 ± 0.04
5	19.38 ± 0.00	20.08 ± 0.00	14.38 ± 0.06	14.95 ± 0.08
6	15.83 ± 0.00	45.62 ± 0.00	12.59 ± 0.62	12.62 ± 0.05
7	11.02 ± 0.00	8.04 ± 0.00	7.85 ± 0.07	7.73 ± 0.04
8	10.13 ± 0.00	11.98 ± 0.00	7.36 ± 0.08	6.95 ± 0.02
9	8.24 ± 0.00	7.18 ± 0.00	5.42 ± 0.05	4.38 ± 0.03
10	13.40 ± 0.00	12.18 ± 0.00	4.51 ± 0.04	4.29 ± 0.00
11	12.07 ± 0.00	10.31 ± 0.00	7.71 ± 0.06	7.21 ± 0.00
12	9.13 ± 0.00	$\mathbf{4.32 \pm 0.00}$	7.21 ± 0.06	6.89 ± 0.00
#	AdaBoost	CNN	GRU	LSTM
1	5.19 ± 0.10	4.74 ± 0.51	$\mathbf{4.30 \pm 0.39}$	4.42 ± 0.22
2	7.76 ± 0.12	6.04 ± 1.08	$\mathbf{5.99 \pm 0.15}$	6.50 ± 0.86
3	12.18 ± 0.71	9.70 ± 0.42	$\mathbf{9.25 \pm 2.35}$	11.60 ± 1.47
4	9.91 ± 0.28	$\mathbf{9.35 \pm 0.06}$	13.84 ± 2.47	14.64 ± 1.97
5	14.44 ± 0.15	12.39 ± 0.25	$\mathbf{12.11 \pm 0.41}$	12.42 ± 0.45
6	12.31 ± 0.28	45.68 ± 67.8	$\mathbf{11.28 \pm 1.58}$	12.75 ± 2.54
7	7.94 ± 0.18	7.29 ± 0.46	$\mathbf{6.23 \pm 0.24}$	6.55 ± 0.18
8	7.25 ± 0.09	$\mathbf{6.46 \pm 0.29}$	7.00 ± 0.55	7.09 ± 0.58
9	5.38 ± 0.18	4.58 ± 0.43	$\mathbf{4.14 \pm 0.18}$	4.37 ± 0.45
10	3.95 ± 0.04	4.01 ± 0.41	$\mathbf{3.16 \pm 0.52}$	3.80 ± 0.47
11	7.77 ± 0.08	7.39 ± 0.41	6.37 ± 0.24	$\mathbf{6.25 \pm 0.32}$
12	6.79 ± 0.17	8.83 ± 1.46	6.55 ± 0.76	5.99 ± 0.70

Long Short-Term Memory Network (**LSTM**). All models were fit on a species-level and do not take into account any differences between species subgroups/cultivars.

Hyperparameters for the RF, GBDT and AdaBoost models were randomly sampled and evaluated using 5-fold cross validation on the training set. Due to restrictions in computation we did not perform a hyperparameter optimization for the CNN, GRU and LSTM models and used early stopping to prevent overfitting. That is, over a 1000 training epochs, every 10 epochs the model was evaluated on a held out subset of the training data (split by years). If no significant increase (over 10^{-3}) in the validation loss was observed during 5 consecutive evaluations, the training procedure was terminated. Gradients were optimized using the Adam optimizer using a learning rate of 10^{-3} that decayed to around 10^{-4} over the full procedure.

Experiments were executed on a Lenovo p16 laptop (Intel Core i9-12950HX, 2300Mhz, 32GB RAM, RTX A5500 GPU) running Windows 11 Enterprise. Obtaining all results took around 4 days. All code used to obtain the publicly available datasets and run the experiments is available on GitHub.

Discussion

Model Generalization From looking at the MAE reported in Table 3, it is clear that the GRU shows the lowest generalization error in the majority of datasets. It is also clear, however, that all deep learning models show much larger variation between runs using different seeds for optimization and weight initialization and that the tree-based methods are more consistent. Even on datasets with a low number of samples (#4 Almond, #6 Apricot, with 187 and 283

samples, respectively) the models seem to do well. However, upon further inspection, the variation between runs is very large and on average the predictions are only $\sim$3 days off from the target mean. The CNN, unlike other deep learning models, clearly did not fit well on dataset #6. We suspect this difference between the CNN and other deep learning models originated from the modeling setup, where the recurrent models provided daily flowering probabilities based on which a binary cross-entropy loss was computed, whereas the CNN directly used a regression (MSE) loss.

Model Improvement It is likely that some reconfiguration of the training procedure can result in a closer fit. We suspect, however, that the most promising avenue for model improvement is to account for the differences in local and cultivar effects. The current baseline models operate on a species level and account for no such distinction. Most notably Japan where indeed multiple species are present in the dataset with unique flowering characteristics. It should also be noted however, that Japan has a wide variety of climates which also cause differences in timing. In the other single-species datasets this difference could be explained by cultivars, where some data is labeled to correspond to early or late blooming cultivars. Although not strictly following this benchmark, the inclusion of additional input features is another interesting direction of research. Temperature is considered the main driver of plant phenology, but other effects might not be accounted for (Fu et al. 2015; Laube et al. 2014; Jochner et al. 2013).

Benchmark Improvement Due to the data sources that are used, different parts of the globe are not well represented in BloomBench. In its current state, the dataset contains observations made in Western and Central Europe, as well as Japan and South Korea. More data sources should be integrated to improve the global representation of the dataset (e.g. the USA National Phenology Network and Chinese Phenology Observation Network), and increase the diversity of included climates. Yearly updates of the benchmark can include observations of the latest season and thus provide new validation data. In this work we focus only fruit trees to emphasize the relevance to both ecology as agriculture. However, many more plant species are being observed and the code base allows for easy extension. No lead time is included in the current task formulation. In operational use, the included models could be paired with weather forecasts to provide predictions in the future.

Conclusion

Accurate modeling of the timing of phenological events has many implications for both ecology and agriculture, ranging from understanding ecosystem response to climate change to predicting yield and optimizing crop management. Machine learning research applied to this problem has thus far been focused on separate case studies and is difficult to inter-compare. This work introduces BloomBench, a benchmark dataset for fruit tree phenology prediction using machine learning. The benchmark facilitates standardized model evaluation for datasets compiled from various observation networks containing multiple species in a diverse set of climates. BloomBench can reduce fragmentation and accelerate research in the domain. Moreover, it provides a real use-case for evaluating machine learning methods in their ability to generalize in various domain shifts, and provide reliable estimates given a change in climate.

References

Aono, Y.; and Kazui, K. 2008. Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 28(7): 905–914.

Auerbach, J.; Kepplinger, D.; and Wolkovich, E. 2022. George Mason University International Cherry Blossom Prediction Competition. https://competition.statistics.gmu.edu/. Accessed: 2024-03-29.

Baldocchi, D.; and Wong, S. 2008. Accumulated winter chill is decreasing in the fruit growing regions of California. *Climatic Change*, 87: 153–166.

Chuine, I.; de Cortázar-Atauri, I. G.; Kramer, K.; and Hänninen, H. 2025. Plant Phenology Models. In *Phenology: An Integrative Environmental Science*, 315–337. Springer.

Else, M.; and Atkinson, C. 2010. Climate change impacts on UK top and soft fruit production. *Outlook on Agriculture*, 39(4): 257–262.

Fernandez, E.; Whitney, C.; and Luedeling, E. 2020. The importance of chill model selection—A multi-site analysis. *European Journal of Agronomy*, 119: 126103.

Fu, Y. H.; Piao, S.; Vitasse, Y.; Zhao, H.; De Boeck, H. J.; Liu, Q.; Yang, H.; Weber, U.; Hänninen, H.; and Janssens, I. A. 2015. Increased heat requirement for leaf flushing in temperate woody species over 1980–2012: effects of chilling, precipitation and insolation. *Global change biology*, 21(7): 2687–2697.

Jochner, S.; Höfler, J.; Beck, I.; Göttlein, A.; Ankerst, D. P.; Traidl-Hoffmann, C.; and Menzel, A. 2013. Nutrient status: a missing factor in phenological and pollen research? *Journal of Experimental Botany*, 64(7): 2081–2092.

Katal, N.; Rzanny, M.; Mäder, P.; and Wäldchen, J. 2022. Deep learning in plant phenological research: A systematic literature review. *Frontiers in Plant Science*, 13: 805738.

Lamichhane, J. 2021. Rising risks of late-spring frosts in a changing climate. Nat Clim Chang 11: 554–555.

Laube, J.; Sparks, T. H.; Estrella, N.; and Menzel, A. 2014. Does humidity trigger tree phenology? Proposal for an air humidity based framework for bud development in spring. *New Phytologist*, 202(2): 350–355.

Luedeling, E. 2012. Climate change impacts on winter chill for temperate fruit and nut production: a review. *Scientia Horticulturae*, 144: 218–229.

Luedeling, E.; Girvetz, E. H.; Semenov, M. A.; and Brown, P. H. 2011. Climate change affects winter chill for temperate fruit and nut trees. *PloS one*, 6(5): e20155.

Madec, S.; Irfan, K.; Velumani, K.; Baret, F.; David, E.; Daubige, G.; Samatan, L. B.; Serouart, M.; Smith, D.; James, C.; et al. 2023. VegAnn, vegetation annotation of

multi-crop RGB images acquired under diverse conditions for segmentation. *Scientific Data*, 10(1): 302.

Masago, Y.; and Lian, M. 2022. Estimating the first flowering and full blossom dates of Yoshino cherry (Cerasus× yedoensis 'Somei-yoshino') in Japan using machine learning algorithms. *Ecological Informatics*, 71: 101835.

Meier, U.; et al. 1997. Growth stages of mono-and dicotyledonous plants= Entwicklungsstadien mono-und dikotyler Pflanzen= Estadios de las plantas mono-y dicotiledóneas= Stades phénologiques des mono-et dicotylédones cultivées. *Berlin [etc.]: Blackwell*.

Mori, A.; Doi, Y.; and Iizumi, T. 2023. GCPE: The global dataset of crop phenological events for agricultural and earth system modeling. *Journal of Agricultural Meteorology*, 79(3): 120–129.

Nagai, S.; Morimoto, H.; and Saitoh, T. M. 2020. A simpler way to predict flowering and full bloom dates of cherry blossoms by self-organizing maps. *Ecological Informatics*, 56: 101040.

Nappa, A.; Quartulli, M.; Azpiroz, I.; Marchi, S.; Guidotti, D.; Staiano, M.; and Siciliano, R. 2024. Probabilistic Bayesian Neural Networks for olive phenology prediction in precision agriculture. *Ecological Informatics*, 82: 102723.

Oses, N.; Azpiroz, I.; Quartulli, M.; Olaizola, I.; Marchi, S.; and Guidotti, D. 2020. Machine Learning for olive phenology prediction and base temperature optimisation. In *2020 Global Internet of Things Summit (GIoTS)*, 1–6. IEEE.

Piao, S.; Liu, Q.; Chen, A.; Janssens, I. A.; Fu, Y.; Dai, J.; Liu, L.; Lian, X.; Shen, M.; and Zhu, X. 2019. Plant phenology and global climate change: Current progresses and challenges. *Global change biology*, 25(6): 1922–1940.

Reaumur, R. d. 1735. Observations du thermomètre faites à Paris pendant l'année 1735, comparées avec celles qui ont été faites sous la ligne, à l'Isle de France, à Alger et quelques unes de nos iles de l'Amérique. *Mémoires l'Académie R des Sci*, 545–576.

Rezaei, E. E.; Siebert, S.; Hüging, H.; and Ewert, F. 2018. Climate change effect on wheat phenology depends on cultivar change. *Scientific reports*, 8(1): 4891.

Saxena, A.; Pesantez-Cabrera, P.; Ballapragada, R.; Keller, M.; and Fern, A. 2023a. Multi-Task Learning for Budbreak Prediction. *arXiv preprint arXiv:2301.01815*.

Saxena, A.; Pesantez-Cabrera, P.; Ballapragada, R.; Lam, K.-H.; Keller, M.; and Fern, A. 2023b. Grape cold hardiness prediction via multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15717–15723.

Seyednasrollah, B.; Young, A. M.; Hufkens, K.; Milliman, T.; Friedl, M. A.; Frolking, S.; and Richardson, A. D. 2019. Tracking vegetation phenology across diverse biomes using Version 2.0 of the PhenoCam Dataset. *Scientific data*, 6(1): 222.

Shin, N.; Fujiwara, H.; Sugiyama, S.; Morimoto, H.; and Saitoh, T. M. 2025. Estimation of true dates of various flowering stages at a centennial scale by applying a Bayesian statistical state space model. *PloS one*, 20(2): e0317708.

Templ, B.; Koch, E.; Bolmgren, K.; Ungersböck, M.; Paul, A.; Scheifinger, H.; Rutishauser, T.; Busto, M.; Chmielewski, F.-M.; Hájková, L.; et al. 2018. Pan European Phenological database (PEP725): a single point of access for European data. *International journal of biometeorology*, 62: 1109–1113.

van Bree, R.; Marcos, D.; and Athanasiadis, I. N. 2025. Hybrid Phenology Modeling for Predicting Temperature Effects on Tree Dormancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28458–28466.

Vautard, R.; van Oldenborgh, G. J.; Bonnet, R.; Li, S.; Robin, Y.; Kew, S.; Philip, S.; Soubeyroux, J.-M.; Dubuisson, B.; Viovy, N.; et al. 2022. Human influence on growing-period frosts like the early April 2021 in Central France. *Natural Hazards and Earth System Sciences Discussions*, 2022: 1–25.

Vitasse, Y.; Hoch, G.; Randin, C. F.; Lenz, A.; Kollas, C.; Scheepens, J.; and Körner, C. 2013. Elevational adaptation and plasticity in seedling phenology of temperate deciduous tree species. *Oecologia*, 171: 663–678.

Wang, H.; Wu, C.; Ciais, P.; Peñuelas, J.; Dai, J.; Fu, Y.; and Ge, Q. 2020. Overestimation of the effect of climatic warming on spring phenology due to misrepresentation of chilling. *Nature Communications*, 11(1): 4945.

Weyler, J.; Magistri, F.; Marks, E.; Chong, Y. L.; Sodano, M.; Roggiolani, G.; Chebrolu, N.; Stachniss, C.; and Behley, J. 2024. PhenoBench: A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.