Hybrid Phenology Modeling for Predicting Temperature Effects on Tree Dormancy

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

Biophysical models offer valuable insights into climate-phenology relationships in both natural and agricultural settings. However, there are substantial structural discrepancies across models which require site-specific recalibration, often yielding inconsistent predictions under similar climate scenarios. Machine learning methods offer data-driven solutions, but often lack interpretability and alignment with existing knowledge. We present a phenology model describing dormancy in fruit trees, integrating conventional biophysical models with a neural network to address their structural disparities. We evaluate our hybrid model in an extensive case study predicting cherry tree phenology in Japan, South Korea and Switzerland. Our approach consistently outperforms both traditional biophysical and machine learning models in predicting blooming dates across years. Additionally, the neural network's adaptability facilitates parameter learning for specific tree varieties, enabling robust generalization to new sites without site-specific recalibration. This hybrid model leverages both biophysical constraints and data-driven flexibility, offering a promising avenue for accurate and interpretable phenology modeling.

Code — https://github.com/WUR-AI/HybridML-Phenology

Introduction

Modeling tree phenology is important to understand the effects of a changing climate on natural and agricultural ecosystems. Temperate trees require a sustained period of cold temperatures to overcome dormancy, protecting them from harsh winter conditions. Temperature changes can delay or accelerate this process, making this phenological process an important component in modeling secondary effects, such as impacts on orchard yields (Baldocchi and Wong 2008; Luedeling 2012), biological carbon sequestration (Cleland et al. 2007; Richardson et al. 2010) and the phenological stages of other organisms (Visser and Holleman 2001; Kudo 2014)). Cherry trees are of special interest in climate change research, as their flowering dates are strongly temperature dependent (Samish 1954) and their cultural significance in Japan has allowed the collection of a long record of flowering dates throughout the country (going back ∼1200 years in Kyoto (Aono and Kazui 2008)).

Trends in cherry tree blooming dates data have been used as an indicator of changes in climate (Aono and Omoto 1993; Primack, Higuchi, and Miller-Rushing 2009).

Biophysical phenology models generally assume the daily accumulation of temperature-dependent units towards some threshold at which a phenological transition occurs. Models typically distinguish between endodormancy, in which a period of cold temperatures needs to be endured, and a subsequent requirement for warmer temperatures during ecodormancy. Commonly used models, however, differ in how they relate daily temperature to their contribution towards endodormancy release and display differences in behaviour under the same climate scenarios (Luedeling and Brown 2011; Fernandez, Whitney, and Luedeling 2020). The models have typically been proposed as the simplest functions that are able to explain observations from controlled experiments on optimal chilling temperatures and effective temperature ranges, and weight their contribution accordingly. These models work well for predicting blooming dates but do not fully capture the true dynamics that are known to be more complex (Kaufmann and Blanke 2019). Given the uncertainty in our understanding of these biological processes, choosing a particular model often introduces a strong bias.

The increasing availability of data on climate conditions, along with large datasets on corresponding plant phenology, allows for a more data driven approach to model these dynamics through expressive machine learning methods (Masago and Lian 2022). Complex over-parameterized models, due to their tendency to overfit to the training conditions, have difficulty matching simple process-based models based on biological insights when tasked to predict dormancy release based solely on temperature data (Saxena et al. 2023). (Nagai, Morimoto, and Saitoh 2020) show self-organizing maps can be used to predict flowering dates, but show higher prediction error than process-based models. Moreover, the inherent interpretability provided by these process-based models is lost (Rudin 2019). Introducing the appropriate inductive bias to machine learning models by coupling them with mechanistic models (an approach often referred to as hybrid modelling) is widely recognized as a means to improve their data-efficiency and interpretability by constraining them to solutions that are consistent with domain knowledge and has been successfully applied in a wide variety of domains (Karpatne et al. 2017; Reichstein
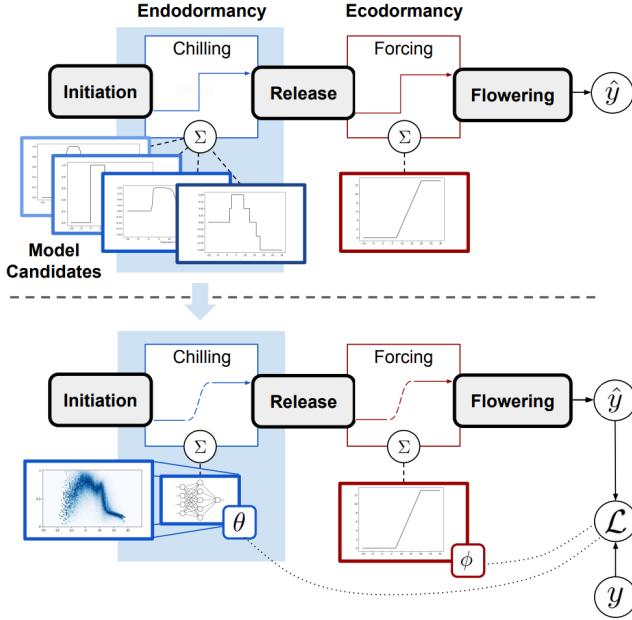
Figure 1: Schematic overview of two approaches modeling the phenological stages between dormancy initiation and flowering. Biophysical model candidates (Top) show structural discrepancies in how temperature contributes towards endodormancy progression. In this work (Bottom), we replace this module with an MLP and jointly optimize it with the biophysical model parameters, obtaining a new temperature response function. We then evaluate its ability to generalize to unobserved seasons and reflect on its biophysical plausibility.

et al. 2019; Shen et al. 2023).

In this work, we propose a hybrid model for predicting phenological transitions by providing a generalized piecewise differentiable approximation to commonly used biophysical models of tree dormancy and substitute the component showing large structural disagreement (i.e. the function relating daily temperatures to progression towards endodormancy release) with a multilayer perceptron (MLP), sketched in Figure 1. Gradients with respect to the MLP parameters θ and process-based model parameters ϕ can be computed by performing backpropagation through the constrained model. These parameters are then jointly optimized using an extensive dataset on cherry tree flowering dates in Japan, Switzerland and South Korea, paired to MERRA-2 hourly temperature data. The substitution of the biophysical model component allows us to obtain a learned temperature response function provided by the MLP. We refrain from using more complex architectures, to enable direct intercomparison of inputs/outputs with biophysical models that generally use stateless chill functions with a limited temporal context window (ruling out recurrent networks, for example). We compare the model to commonly used process-based models w.r.t. their ability to predict blooming dates. Finally, we reflect on the plausibility of the temperature response function learned by the MLP.

Methodology

We formalize the task of predicting blooming day-of-year (DoY) as a regression problem and compare models M : $\mathbb{R}^{S \times 24} \to \mathbb{R}$ that relate hourly temperature values for every day within the season (of length S) to a DoY value that is discretized during evaluation.

Dataset

The dataset used was based on data released by George Mason's Department of Statistics as part of a competition on cherry blossom DoY prediction (Auerbach, Kepplinger, and Wolkovich 2022). It is composed of blooming dates for numerous locations with diverse climates scattered throughout Japan, Switzerland and South Korea. We paired observed dates y in this dataset with MERRA-2 hourly surface temperature data $\mathbf{X} \in \mathbb{R}^{S \times 24}$ forming a dataset $\{(\mathbf{X}_n, y_n)\}_N$, where $\mathbf{X}_n = [\mathbf{x}_1, .., \mathbf{x}_S]$ contains the temperature data of the season matching flowering DoY y_n in some year at some location. Similar to related work on modeling cherry tree phenology we set the 1st of October as start of the season and based on observed bloom DoY the season length was set to $S = 274$ (ending on the 1st of July on non-leap years). MERRA-2 provides data starting from 1980, so any preceding blooming dates have been omitted. The locations were mapped to their respective tree variety as reported by the Japanese Meteorological Agency and Swiss Phenology Network for Japan and Switzerland, respectively. Not all tree varieties could be identified for the locations in Japan. These (7) locations were excluded. Trees in South Korea were assumed to be of the *Prunus×Yedoensis* variety based on previous work (Hur, Ahn, and Shim 2014; Chung et al. 2009) and its frequent occurrence. Figure 2 shows the locations and distribution of tree varieties included in the dataset. The process based models that are considered are not suitable for the subtropical regions (Okinawa, Amami) where the *Prunus Campanulata Maxim* trees grow. These were excluded to make a fair model comparison. Trees species that were represented in one only one location were removed as well. The Japanese dataset thus only contains data on the *Prunus ×Yedoensis* and *Prunus Sargentii* varieties. Since the data sources for each country adhered to their own flowering definition, models are trained separately for Japan (N=3317), Switzerland (N=4780) and South Korea (N=930).

Models

Three classes of models are evaluated in this work: mechanistic biophysical models, our proposed hybrid model, and purely data-driven neural networks. Since biophysical models show strong structural similarity, a generalized description is provided first, followed by an overview of how the different models fit within this framework. Subsequently, we describe our approach and finally the data-driven baseline.

Phenology Models Mechanistic phenology models of dormancy in deciduous trees generally consider two successive stages: endodormancy and ecodormancy. Before being able to transition between stages, a tree is required to spend a
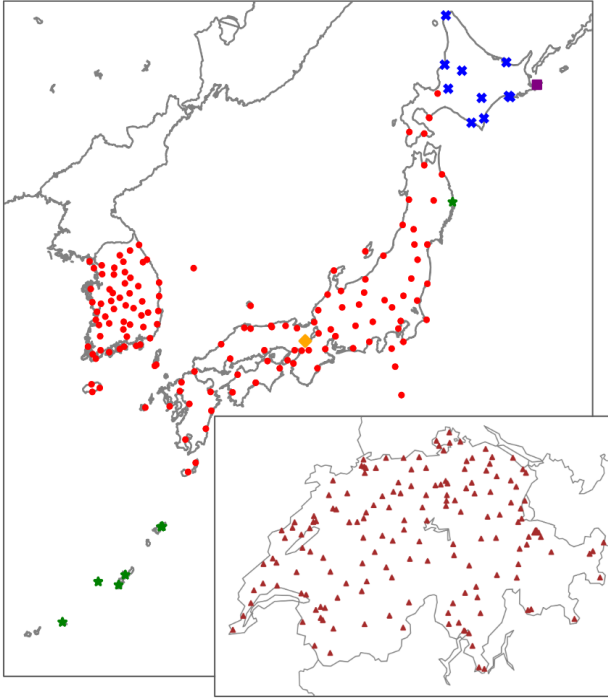
Figure 2: Spatial distribution of locations that were available in the dataset for each country (not to scale), colored by (partially assumed) tree variety occurrence. Varieties include: *Prunus ×Yedoensis* (Red circles), *Prunus Sargentii* (Blue crosses), *Prunus Campanulata Maxim* (Green stars), *Prunus Nipponica Matsum* (Purple squares), *Prunus Jamasakura* (Orange diamonds), *Prunus Avium* (Brown triangles).

period in cold temperatures, which is typically modeled using the daily accumulation of temperature-dependent units (referred to as chilling units) that capture the progression towards completing this requirement. Subsequently, before flowering, a tree is required to spend a period under warm temperatures that are modeled in a similar fashion using the accumulation of "forcing" units towards some required threshold, after which blooming occurs.

Let $c(\mathbf{x}_t)$, $f(\mathbf{x}_t)$ ($c, f : \mathbb{R}^{24} \to \mathbb{R}_{\geq 0}$) denote functions that define the respective daily contribution of chilling and forcing based on hourly temperature data. The exact formulation of how these functions relate temperature levels to contribution to the phenological stages is, for now, left unspecified, since this is where the considered biophysical models show discrepancies. Let C_t denote the cumulative sum of chilling units at day t and $r_t^{(c)} \in \{0, 1\}$ an indicator variable specifying whether a threshold $\beta^{(c)}$ has been met:

$$C_t = \sum_{\tau=1}^{t} c(\mathbf{x}_\tau) \qquad r_t^{(c)} = \mathbb{1}(C_t \geq \beta^{(c)}) \qquad (1)$$

The accumulation of forcing units is only considered after the chilling requirement is fulfilled. Its cumulative sum F_t

and forcing requirement $r_t^{(f)}$ are defined similarly:

$$F_t = \sum_{\tau=1}^{t} f(\mathbf{x}_\tau) r_\tau^{(c)} \qquad r_t^{(f)} = \mathbb{1}(F_t \geq \beta^{(f)}) \qquad (2)$$

where $r_t^{(f)}$ indicates that flowering has occurred and $\beta^{(c)}, \beta^{(f)}$ are parameters of the model.

Baseline Phenology Models The biophysical models that are considered all fit in the framework described in the previous section, but differ in their choice of c and f. A rough description of commonly used chilling models is provided below, followed by those used for forcing. For details on the exact weighting of temperature levels we refer to their original publication. The Chill Hours model (Bennett et al. 1949; Weinberger et al. 1950) simply counts the number of hours spent between 0°C and some threshold temperature (usually 7.2°C). The Utah model (Richardson, Seeley, and Walker 1974) weights the effectiveness of hourly chill unit accumulation based on the respective temperature values. The Chill Days model (Cesaraccio et al. 2004) shares a parameter T_B that controls the optimal temperature for both chilling and forcing accumulation to be suitable for a wider variety of climates. Among these chilling models, only the Chill Days model was introduced with a complementary forcing function f. For the other models one was assigned instead. There is less variation between forcing functions and most models adapt a Growing Degree Days approach to hourly timesteps (Richardson et al. 1975), where some apply an additional weight function (Anderson, Richardson, and Kesner 1985; Luedeling et al. 2021). In this work we follow this Growing Degree Hours (GDH) approach, without the additional weight function:

$$f_{\text{GDH}}(\mathbf{x}_t) = \sum_{i=1}^{24} \max(0, x_{t,i} - T_B) \qquad (3)$$

where T_B is a base temperature parameter for controlling the effectiveness of temperature levels to forcing (Richardson, Seeley, and Walker 1974). In total, all models have three parameters which we collectively define as $\phi = (\beta^{(c)}, \beta^{(f)}, T_B)$.

Proposed Model In this section we describe how the mechanistic phenology model framework is adapted to enable substitution of the chill response function. A (piecewise) differentiable approximation to the mechanistic models is built using generalized logistic functions to approximate the step functions used by these models:

$$\sigma(x; \alpha, \beta, \gamma) = \frac{\gamma}{1 + e^{-\alpha(x-\beta)}} \qquad (4)$$

where α controls the slope, β the inflection point and γ the right-hand side asymptote. Mechanistic models generally agree on the forcing function f that is used. Moreover, it meets the piece-wise differentiability requirements (Eq. 3) and thus remains unchanged. The chilling function c, however, shows stronger dissimilarities and is substituted by a learnable multilayer perceptron (MLP) consisting of two hidden layers with each containing 64 neurons and a ReLU

activation function. In the final layer, a sigmoid function is used instead, such that C_t is bounded between 0 and season length S.

Let $\tilde{c}(\cdot; \theta)$ denote the daily MLP chilling contribution parameterized by θ. The complete description of our model is then provided by

$$\tilde{C}_t = \sum_{\tau=1}^{t} \tilde{c}(\mathbf{x}_\tau; \theta) \qquad \tilde{r}_t^{(c)} = \sigma(\frac{\tilde{C}_t}{S}; \alpha^{(c)}, \beta^{(c)}, 1) \quad (5)$$

$$\tilde{F}_t = \sum_{\tau=1}^{t} f(\mathbf{x}_\tau)\tilde{r}_\tau^{(c)} \qquad \tilde{r}_t^{(f)} = \sigma(\frac{\tilde{F}_t}{S}; \alpha^{(f)}, \beta^{(f)}, 1) \quad (6)$$

Two additional parameters $\alpha^{(c)}, \alpha^{(f)}$ are introduced that control the slope of the threshold approximation. $\tilde{C}_t, \tilde{F}_t$ are divided by S to limit the size of the exponent in Eq. 4 for numerical stability. The original model structure, and thus parameters $\phi = (\beta^{(c)}, \beta^{(f)}, T_B)$, are preserved.

The soft indicator $\tilde{r}_t^{(f)}$ no longer provides a binary indication of whether flowering has occurred, and we assign it a probabilistic interpretation. That is, we consider $\tilde{r}_t^{(f)} = \sigma(\cdot; \alpha^{(f)}, \beta^{(f)}, 1)$ to be the cumulative distribution function of a logistic distribution over $\frac{F_t}{S}$ defining the probability that blooming has already occurred at a certain stage of development (as provided by $\frac{F_t}{S}$). Similar to (Terres et al. 2013) and (Allen et al. 2014), we then define the probability of flowering at day t and minimize the corresponding negative log-likelihood:

$$\mathcal{L}(t, \theta, \phi) = -\log(\tilde{r}_t^{(f)} - \tilde{r}_{t-1}^{(f)}). \quad (7)$$

During evaluation, DoY predictions are obtained by considering the day with maximum probability of blooming $\underset{t}{\operatorname{argmax}}[\tilde{r}_t^{(f)} - \tilde{r}_{t-1}^{(f)}]$.

Data-driven Baseline   A data-driven baseline consists of a 1D convolutional neural network (Fukushima 1980; Le-Cun et al. 1989) (CNN), as well as a two-layer long-short term memory (Hochreiter and Schmidhuber 1997) (LSTM) model with a hidden state size of 128. At every point in time, this hidden state is the input of a linear layer with a final sigmoid activation to obtain daily probability values. This results in a model size that is an order of magnitude larger than the hybrid model, but we found that this was required to generalize well in all settings. We follow the approach used by (Saxena et al. 2023) and use a binary cross-entropy loss on the daily predictions provided by the model. When compared to models with location-specific parameters ϕ, normalized coordinates are appended to the model input to be able to distinguish between locations as well. These coordinate features are omitted otherwise.

Model Optimization

Mechanistic Models   For each of the considered models, a parameter grid search was performed over three parameters: chill requirement $\beta^{(c)}$, forcing requirement $\beta^{(h)}$ and base temperature T_B. Parameters were optimized per location or

tree variety and selected based on the mean squared error obtained using the training dataset.

Hybrid Model   The differentiability of the model allows us to perform backpropagation to compute gradients w.r.t. both the biophysical model parameters ϕ and the MLP parameters θ. Gradients were computed using PyTorch (Paszke et al. 2019) Autograd over the complete training dataset. An Adam optimizer (Kingma and Ba 2014) was used to update the model parameters with an initial learning rate of $1e^{-3}$ and weight decay of $1e^{-4}$. Over a total of 20000 epochs this learning rate decayed by a factor of 0.9 for every 2000 iterations over the dataset. Chilling requirement slope $\alpha^{(c)}$ was set to 50 to closely approximate the step function used in process based models. Forcing requirement slope $\alpha^{(f)}$ was interpreted as the scale parameter of a logistic distribution. Its value was initialized to 1 and learned during the training procedure, while enforcing a lower bound of 0.01.

Data-driven   The settings are identical to that of the hybrid model, except that, due to memory limitations, gradients are computed in batches with a batch size set to 512.

Experiments

To investigate the generalization capabilities of each model, they are fit and evaluated in several settings, as described below. Every experiment was run 10 times using different random seeds. Table 1 reports the mean absolute error (MAE) of predicted DoY blooming dates, as well as its standard error, obtained for each setting. MAE was chosen over other metrics for its lower sensitivity to outliers, since we found that the mechanistic models were more prone to give end-of-season predictions if thresholds were not met.

1. Temporal Generalization per Location   The first setting focuses on the ability of the models to predict blooming dates in years that are not used to fit the model, and for the same locations used for training. The years that span the dataset (1980-2021) are randomly split in train (75%) and test years to evaluate the model in unobserved weather conditions. No restrictions are imposed on the parameters ϕ, allowing them to take on different values for each considered location. Table 1 shows the ability of the three biophysical models to generalize to new years based on the parameters obtained using grid search. These models show comparable results in predicting blooming dates within each of the countries. All models are best able to predict blooming dates in Japan, followed by South Korea and Switzerland. We suspect this is caused by the lower data availability in South Korea with fewer data points available per location and stronger influence of local temperature differences in Switzerland due to elevation/shading. The LSTM, on the other hand, performs comparably in South Korea and Switzerland in the temporal generalization setting, since the noisier nature of the Swiss dataset may be compensated by a higher number of training samples compared to South Korea. In this lower data regime, the process-based models tend to outperform the LSTM. On the other hand, the LSTM outperforms all process-based models in data-rich Japan.

From Table 1 it can be seen the proposed model shows a lower mean absolute error than all process-based models, as

	MAE ±SE		
	Setting 1 - Temporal		
Model	**Japan**	**Switzerland**	**South Korea**
Chill Hour	3.76 ±0.11	6.52 ±0.18	6.42 ±0.36
Utah Chill	3.07 ±0.06	7.32 ±0.28	6.12 ±0.27
Chill Days	3.71 ±0.09	8.47 ±0.29	7.13 ±0.44
Hybrid	**2.17 ±0.03**	**5.36 ±0.15**	**4.86 ±0.06**
CNN	3.80 ±0.09	11.28 ±0.17	6.21 ±0.17
LSTM	2.80 ±0.06	7.11 ±0.09	7.22 ±0.57
	Setting 2 - Temporal (Variety)		
Model	**Japan**	**Switzerland**	**South Korea**
Chill Hour	18.03 ±0.28	17.64 ±0.35	9.50 ±0.26
Utah Chill	15.65 ±0.31	18.14 ±0.34	6.97 ±0.17
Chill Days	16.68 ±0.30	18.54 ±0.43	10.76 ±0.54
Hybrid	3.28 ±0.04	**10.2 ±0.15**	**5.41 ±0.08**
CNN	3.98 ±0.10	12.21 ±0.31	6.21 ±0.13
LSTM	**3.24 ±0.05**	12.75 ±0.35	7.10 ±0.37
	Setting 3 - Spatiotemporal (Variety)		
Model	**Japan**	**Switzerland**	**South Korea**
Chill Hour	18.34 ±0.78	17.44 ±0.62	11.23 ±0.98
Utah Chill	15.29 ±0.73	18.13 ±0.59	7.31 ±0.55
Chill Days	16.80 ±0.96	18.10 ±0.69	13.62 ±2.15
Hybrid	**3.64 ±0.09**	**10.6 ±0.37**	**6.07 ±0.38**
CNN	4.09 ±0.12	12.66 ±0.34	7.00 ±0.43
LSTM	3.94 ±0.18	13.24 ±0.41	8.30 ±0.94

Table 1: Mean Absolute Error (in days) and Standard Error obtained from evaluating three process-based models (Chill Hour, Utah Chill and Chill Days models), an LSTM, CNN, as well as the proposed model (Hybrid) in three different countries on a held out test dataset. The models are evaluated for their generalization capabilities in multiple settings, as described in the results section.

well as the LSTM, in each of the considered countries when evaluated on data from held out test years. Interestingly, our hybrid model behaves akin to the process-based models in that it is able to fit better to the South Korea dataset than in Switzerland. At the same time, it is able to profit from the higher data availability in Japan, thus outperforming all baselines on all datasets.

The proposed model is constructed using two adaptations from mechanistic models, namely the piece-wise differentiable approximation of the model structure and the substitution of the chill temperature response function by an MLP. To investigate whether it is truly the learned chill function that causes the improvement in model performance, when compared to traditional models, and not the differentiable model approximation, we optimize this approximation when using the Utah chilling function instead. The resulting MAE of 3.17 ±0.07 days for Japan does not match the 2.17 ±0.03 that was obtained when using our hybrid model, indicating the learned chill temperature response function does indeed cause the model improvement.

2. Temporal Generalization per Variety A common use-case for phenology models is to be applied in locations without data availability, and model parameters are thus often identified per tree variety, rather than per location. To investigate our model's ability to learn such parameters, we use an evaluation setup identical to the first with the exception that models are calibrated per tree variety (Figure 2), rather than having separate parameters ϕ for each location. This results in all locations with the same tree variety sharing parameters and thus showing the same response to temperature. In this case, Table 1 and Figure 3 show that the biophysical models underfit, and their ability to predict blooming dates degrades. The assumption is especially restrictive in countries with numerous locations over different climatic zones, as can be seen by the stronger model degradation in Japan and Switzerland. The expressiveness of the LSTM prevents it from underfitting in this setting and, with sufficient data, such as in Japan, it shows the lowest MAE of all models. In Switzerland and South Korea, this benefit is not apparent. Our hybrid model matches the LSTM in its generalization capability in Japan and outperforms all three biophysical models and LSTM in Switzerland and South Korea. The imposed structure helps in settings with less available data or data of lower quality, while its data-driven nature allows generalization across locations.

3. Spatio-temporal Generalization A subsequent step is to evaluate whether the variety-specific model parameters actually generalize to unseen locations as well. To do this, we hold out 25% of locations to be used as test, while keeping a temporal train/test split as in the previous settings. Variety-specific parameter sets are learned using the training dataset in the same manner as in the previous setting. Based on the results of the previous evaluation setting, our hypothesis was that the biophysical model parameters do not generalize well to held out locations. Surprisingly, the over-parameterized LSTM manages to retain much of its generalization capability in Japan. The hybrid model, however, manages to improve over all models with a lower increase of MAE (roughly half a day) for each country.

We expect the temporal train/test split to mitigate the effects of spatial autocorrelation between samples in the two datasets that may bias the evaluation. However, the climates represented in both sets will still be relatively similar, and we expect the location generalization capabilities to decrease the further its climate deviates from the training distribution.

Visualizing the Chill Response Function An approximation of the learned chill function is depicted in Figure 4 by plotting the mean daily temperature and corresponding chill response function output density for all samples in the test set in Japan. The color intensity reflects the density of a response value within each column in temperature intervals of 0.5 °C. Within these temperature levels, some variation in the chill response can be seen, although usually clustered in a small interval. All learned functions show a large variation in output in the colder temperature levels (roughly between -20 to -5 °C), perhaps due to these being less represented in the training dataset. All functions show a strong response around the same temperature range (-5 to 15 °C) with peaks mostly occurring around 0°C. Chill contribution decreases strongly in warmer temperature levels, and all levels provide some non-zero contribution towards fulfilling the chill re-
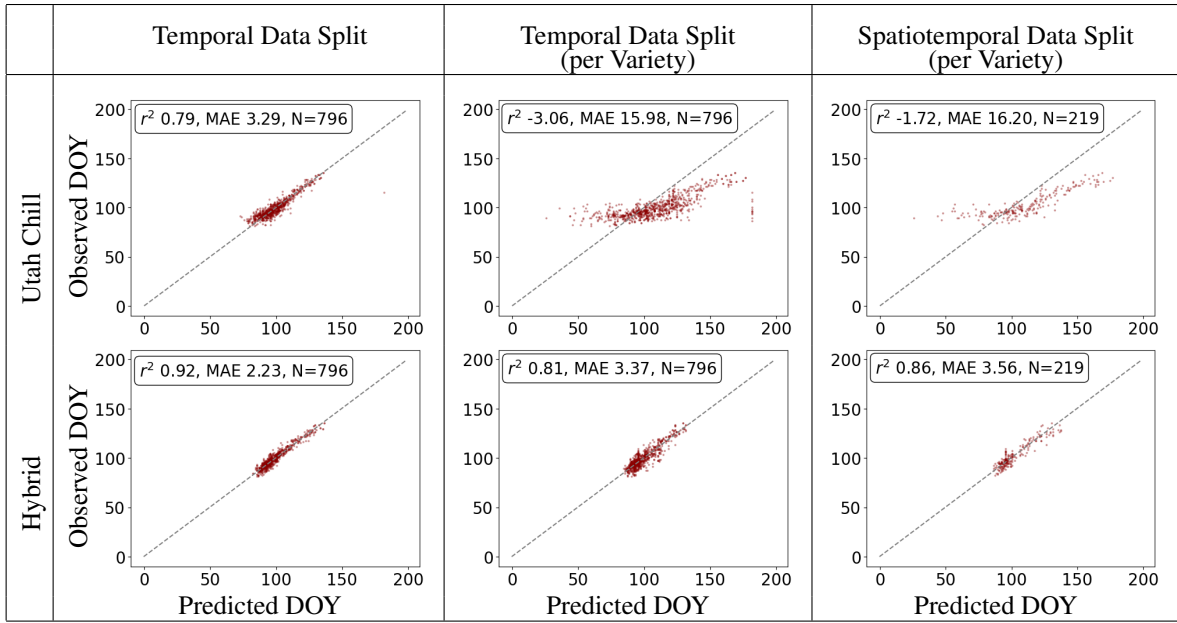
Figure 3: Scatter plots of blooming date predictions for all locations observing the *Prunus ×Yedoensis* tree variety in Japan. Plots are shown for the Utah model (Top) and the model with learned chilling function (Bottom) in the three evaluation settings. It can be seen that, once consistent tree variety parameters are enforced in all locations the mechanistic model underfits, whereas the temperature response function learned by the hybrid model transfers to many locations.

quirement. In Switzerland and South Korea, the model does not learn smooth functions weighting temperature contribution but shows a binary response only giving an output of 1 at temperatures below 15 °C, indicating the importance of the dataset quality.

Experiments were executed on a Lenovo p16 laptop (Intel Core i9-12950HX, 2300Mhz, 32GB RAM, RTX A5500 GPU) running Windows 10 Enterprise. All code used to obtain the publicly available datasets and run the experiments is available on GitHub.

Discussion

The proposed hybrid model shows improved generalization capabilities compared to both the mechanistic models and LSTM. When data is limited, such as in South Korea, LSTM models tend to overfit to the available data and have difficulty to match the performance of much simpler biophysical models. The approximations provided by these models, however, are not expressive enough and thus underfit in settings with high data availability, as is the case in Japan. The additional flexibility of the proposed hybrid approach with respect to the biophysical models learns an approximation that, within its structural constraints, obtains a closer fit and improves generalization within all the evaluated settings. Moreover, for the widely occurring *Prunus ×Yedoensis* variety, a parameter set was obtained that captures the tree behaviour well in a broad range of environments (Figure 2), enabling its usage in locations with low data availability if the variety is known. The mechanistic models, however, are not expressive enough to do the same and require re-calibration per location.

However, there is no guarantee that the hybrid model, within its biophysical constraints, learns a function that actually resembles the biological processes behind the phenological changes, since it is solely optimized for prediction under the provided training dataset. In fact, Figure 4 shows that, for different random weight initializations and dataset splits, different functions are obtained. Moreover, the hybrid model displays a wider range of temperatures that contribute to phenological development than all biophysical models. The Utah model (Richardson, Seeley, and Walker 1974), for example, originated from the following assumption: *"The chilling effect that advances rest completion has an optimum at 6°C and is lost at 0°C and 12.5°C."* The learned chill function does not show consistent peaks around 6°C but does show increased contribution between the 0°C - 10°C temperature range. Chill effect is not lost outside 0°C - 12.5°C. Subzero temperatures do show less contribution as temperatures become colder but with large variation in effect. A decreasing trend is also observed on the other extreme, where chilling effects become less significant in warmer days, but never reaching a level of zero effect. Here, the model conditionally shows implausible behavior, since none of the biophysical models show chill contribution above 12.5°C.

Despite obtaining different response functions for different weight initializations, Table 1 shows little variation in their ability to predict blooming dates, indicating the function is underconstrained. Subsequent research could focus on whether enforcing biophysical resemblance through regularization or including more observations can be used to obtain a temperature response function with more similarity
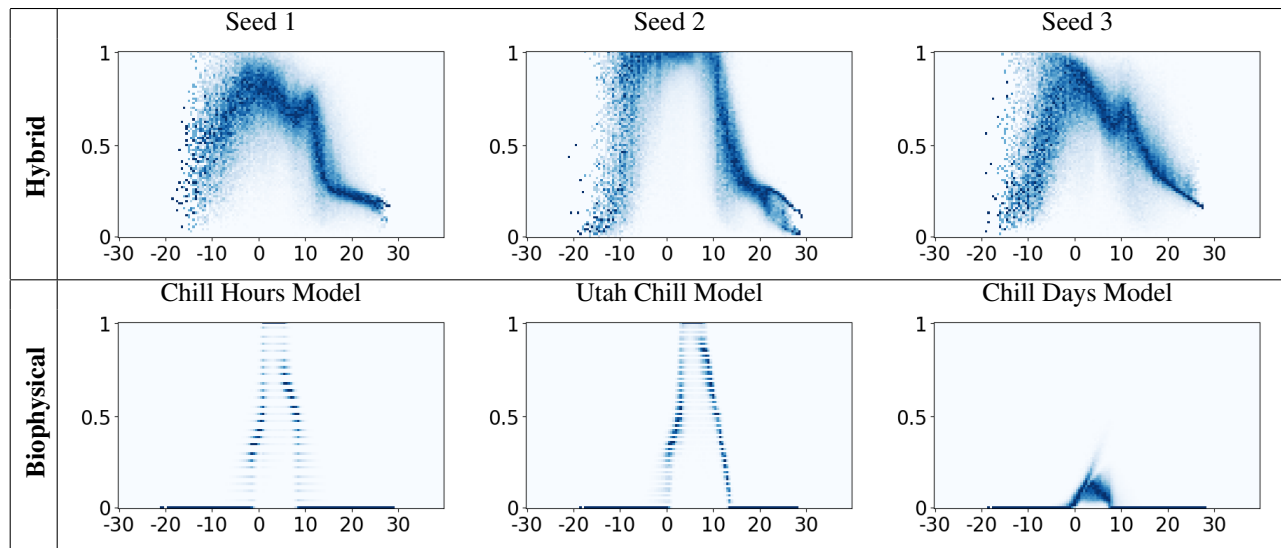
Figure 4: (Top) Visualization of the chill response function learned by the MLP in Japan for the first three random seeds used for evaluation, where the seed controls MLP weight initialization as well as the train/test split. Plots showing the effect of each source of randomness are included in the attached repository. The figure shows the intensity of the response plotted against the mean of the function input (i.e. mean temperature in °C of the 24 hourly measurements of the respective day) for all days occurring in the test dataset. Darker colors highlight the more frequently observed output at each temperature level by showing the density of each column (temperature level discretized to intervals of 0.5 °C). (Bottom) Same visualization when made using the three biophysical models used as baselines. The biophysical models show striped artifacts due to their weighting of discrete temperature intervals.

to existing models without sacrificing generalization ability. In particular, extreme weather events are underrepresented in the training data and no observations include the possibility of not flowering. In this way, the proposed model utilizes the predictive ability of machine learning methods while allowing direct comparison with mechanistic models through its biophysical constraints, therewith providing a promising avenue for improved phenology modeling.

Conclusion

In this work, we demonstrate that machine learning models can help biophysical models (and vice versa) for improved predictions on plant phenology by substituting a component modeling temperature contribution towards dormancy by a learnable function. In all considered environments, this hybrid model improves with respect to three commonly used mechanistic models that were fit using an exhaustive grid search. The structural approximation provided by the mechanistic models underfits on the data, which becomes more apparent when learning a single set of parameters for all locations with the same tree species/variety. Under conditions with sufficient data quality and quantity, the flexibility of the learnable component allows the hybrid model to fit these parameters, while retaining generalization capabilities. Preliminary inspection of the learned chill function shows resemblance to biophysical models, although it provides no guarantee for biophysical plausibility. Indeed, some variation between the temperature response can be seen when training the model with different random seeds, although the over-

all behaviour stays consistent. Despite this, each variation fits equally well to the data, indicating the model remains under-constrained. Further work could investigate whether scaling the model to additional environments results in less variation between the learned models. Alternatively, regularization of the chill function could enforce biophysical plausibility without affecting its generalization ability.

Acknowledgements

References

Allen, J. M.; Terres, M. A.; Katsuki, T.; Iwamoto, K.; Kobori, H.; Higuchi, H.; Primack, R. B.; Wilson, A. M.; Gelfand, A.; and Silander Jr, J. A. 2014. Modeling daily flowering probabilities: expected impact of climate change on Japanese cherry phenology. *Global change biology*, 20(4): 1251–1263.

Anderson, J.; Richardson, E.; and Kesner, C. 1985. Validation of chill unit and flower bud phenology models for'Montmorency'sour cherry. In *I International Symposium on Computer Modelling in Fruit Research and Orchard Management 184*, 71–78.

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.

Aono, Y.; and Omoto, Y. 1993. Variation in the March mean temperature deduced from cherry blossom in Kyoto since the 14th century. *Journal of Agricultural Meteorology*, 48(5): 635–638.

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.

Bennett, J.; et al. 1949. Temperature and bud rest period: effect of temperature and exposure on the rest period of deciduous plant leaf buds investigated. *California Agriculture*, 3(11): 9–12.

Cesaraccio, C.; Spano, D.; Snyder, R. L.; and Duce, P. 2004. Chilling and forcing model to predict bud-burst of crop and forest species. *Agricultural and Forest Meteorology*, 126(1-2): 1–13.

Chung, U.; Jung, J.-E.; Seo, H.-C.; and Yun, J. I. 2009. Using urban effect corrected temperature data and a tree phenology model to project geographical shift of cherry flowering date in South Korea. *Climatic Change*, 93(3-4): 447–463.

Cleland, E. E.; Chuine, I.; Menzel, A.; Mooney, H. A.; and Schwartz, M. D. 2007. Shifting plant phenology in response to global change. *Trends in ecology & evolution*, 22(7): 357–365.

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

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4): 193–202.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-term Memory. *Neural computation*, 9: 1735–80.

Hur, J.; Ahn, J.-B.; and Shim, K.-M. 2014. The change of cherry first-flowering date over South Korea projected from downscaled IPCC AR5 simulation. *International journal of climatology*, 34(7): 2308–2319.

Karpatne, A.; Atluri, G.; Faghmous, J. H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; and Kumar, V. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10): 2318–2331.

Kaufmann, H.; and Blanke, M. 2019. Substitution of winter chilling by spring forcing for flowering using sweet cherry as model crop. *Scientia Horticulturae*, 244: 75–81.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kudo, G. 2014. Vulnerability of phenological synchrony between plants and pollinators in an alpine ecosystem. *Ecological research*, 29: 571–581.

LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; and Jackel, L. 1989. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2.

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

Luedeling, E.; and Brown, P. H. 2011. A global analysis of the comparability of winter chill models for fruit and nut trees. *International Journal of Biometeorology*, 55: 411–421.

Luedeling, E.; Schiffers, K.; Fohrmann, T.; and Urbach, C. 2021. PhenoFlex-an integrated model to predict spring phenology in temperate fruit trees. *Agricultural and Forest Meteorology*, 307: 108491.

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.

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.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Primack, R. B.; Higuchi, H.; and Miller-Rushing, A. J. 2009. The impact of climate change on cherry trees and other species in Japan. *Biological Conservation*, 142(9): 1943–1949.

Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; and Prabhat, f. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195–204.

Richardson, A. D.; Andy Black, T.; Ciais, P.; Delbart, N.; Friedl, M. A.; Gobron, N.; Hollinger, D. Y.; Kutsch, W. L.; Longdoz, B.; Luyssaert, S.; et al. 2010. Influence of spring and autumn phenological transitions on forest ecosystem productivity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1555): 3227–3246.

Richardson, E. A.; Seeley, S. D.; and Walker, D. R. 1974. A model for estimating the completion of rest for 'redhaven'and 'elberta'peach trees1. *HortScience*, 9(4): 331–332.

Richardson, E. A.; Seeley, S. D.; Walker, D. R.; Anderson, J.; and Ashcroft, G. L. 1975. Pheno-climatography of spring peach bud development. *HortScience*, 10: 236–237.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Samish, R. M. 1954. Dormancy in woody plants. *Annual Review of Plant Physiology*, 5(1): 183–204.

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

Shen, C.; Appling, A. P.; Gentine, P.; Bandai, T.; Gupta, H.; Tartakovsky, A.; Baity-Jesi, M.; Fenicia, F.; Kifer, D.; Li, L.; et al. 2023. Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, 1–16.

Terres, M. A.; Gelfand, A. E.; Allen, J. M.; and Silander, J. A. 2013. Analyzing first flowering event data using survival models with space and time-varying covariates. *Environmetrics*, 24(5): 317–331.

Visser, M. E.; and Holleman, L. J. 2001. Warmer springs disrupt the synchrony of oak and winter moth phenology. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1464): 289–294.

Weinberger, J. H.; et al. 1950. Chilling requirements of peach varieties. In *Proceedings. American Society for Horticultural Science*, volume 56, 122–28.