Flowering Onset Detection: Traditional Learning vs. Deep Learning Performance in a Sparse Label Context

Mauricio Soroco[*]ⓡ **Joel Hempel**[*]ⓡ **Xinze Xiong**[*]ⓡ
Department of Computer Science
The University of British Columbia

Mathias Lécuyer[†]ⓡ
Department of Computer Science
The University of British Columbia

Joséphine Gantois[†]ⓡ
Institute for Resources, Environment and Sustainability
The University of British Columbia

Abstract

Detecting temporal shifts in plant flowering times is of increasing importance in a context of climate change, with applications in plant ecology, but also health, agriculture, and ecosystem management. However, scaling up plant-level monitoring is cost prohibitive, and flowering transitions are complex and difficult to model. We develop two sets of approaches to detect the onset of flowering at large-scale and high-resolution. Using fine grain temperature data with domain knowledge based features and traditional machine learning models provides the best performance. Using satellite data, with deep learning to deal with high dimensionality and transfer learning to overcome ground truth label sparsity, is a challenging but promising approach, as it reaches good performance with more systematically available data.

1 Introduction

In temperate regions, temporal shifts in the timing of plant flowering are a key signal of an ecosystem's response to environmental change. In particular, climate change is inducing long-term shifts in average weather, which tends to advance the onset of plant flowering [1]. This impacts a broad range of ecological processes: flower exposure to late frosts or heat waves, which is a major driver of plant reproductive success including crop yield [26]; the synchronization between plants and their pollinators, which respond to different environmental cues [27]; and the onset of pollen season, which impacts pollen allergy prevalence [24]. However, plant flowering is also one of the developmental stages that is most difficult to monitor, compared to more uniform stages like leaf opening, which is well-tracked by remote sensing [30]. In particular, ground observations on the timing of plant flowering are sparse, and difficult to exploit due to variation in flowering onset definition and monitoring frequency. In this paper, **we seek to construct a large-scale proxy measure of the onset of flowering** that is: (1) available at a relatively fine spatial resolution to capture local variations, (2) able to capture interannual variations in the timing of flowering, and (3) able to extrapolate across time and space, to map flowering across many years and large geographical areas. In particular, **we compare the performance of deep learning models that use rich and widely available satellite imagery, with lower dimensional models that use domain knowledge informed features, based on high-frequency weather data or annual vegetation indices.**

Implications Measuring interannual shifts in the onset of plant flowering at high resolution is critical to fully capture climate change impacts on plant yield. Flowering is the period of highest

[*] Students, equal contribution, ⓡandom order. [†] Senior authors, ⓡandom order.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

sensitivity to detrimental weather events [26, 18, 4], and accurate climate impact estimates require to precisely measure the incidence of weather extremes within this shifting window of sensitivity, rather than relying on time-invariant flowering calendars [23]. High resolution flowering data can also help resolve mixed evidence on climate change impacts on plant-pollinator synchronization [27], by increasing the amount of co-occurring data on plant flowering and pollinator flight. It can also inform important management decisions: timing pesticide application to avoid damages to beneficial insects during flowering, timing herbicide application to minimize seed production of invasive plant species, timing beehive placement to optimize pollen availability for bees, or planning for local tourism or spring allergy season. In this paper, we focus here on ex-post detection of flowering onset rather than on forecasting. The target application is having locally accurate data on flowering onset across large spatial and temporal scales, to evaluate more precisely the ex-post impact of climate change and management decisions on ecological outcomes, which will help guide future decisions.

Related work Current efforts to monitor the onset of spring flowering at regional to continental scales rest on numerous but sparse ground observations [19, 22, 29]; on weather-based proxies like the First Bloom Index [25] (included in our model comparison), or weather-based species-specific models of flowering trained on herbarium data [20] (mean absolute error (MAE) ranges from 7 to 79 days across models, and the median MAE is 19 days); and on remotely sensed vegetation metrics, often not validated using ground truth observations [7]. To overcome the difficulty of comparing approaches evaluated on different test datasets, which can vary widely in their underlying variability, we provide a comparison of a comprehensive set of models.

2 The Flowering Onset Prediction Task

At a high level we seek to predict the date of the onset of flowering at the species and landscape level, that is the first day of the year when a certain density of flower clusters are observed across plants of that species in this landscape. In practice, we target a spatial resolution of 500 m to match the resolution of publicly available input data. We instantiate this task for the Common Lilac, an early spring-flowering shrub, using 398 citizen science labels from a long-term monitoring program recording lilac phenology [22]. We focus on a region in the Northeast US, shown on Fig. 1, and on the period 2006-2022, to maximize monitoring density (App. A.1).

Input Data We consider two potential sources of input data, which vary in granularity, richness, and availability. First, we leverage *weather data* from PRISM, which provides daily temperature and precipitation data at a 4 km resolution across the continental US [10]. We construct daily time series of accumulated growing degree days (ADDs), which have long been used for plant phenology modeling [6] and capture how much temperatures accumulate over a predetermined threshold over time [11, 9] (App. A.2). For example, ADDs above 0°C starting from January 1 are the sum of growing degree days above above 0°C, starting from that date, where growing degree days capture how much temperatures exceed 0°C over a day. We also consider using an annual proxy for the onset of flowering, the First Bloom Index, produced by the National Phenology Network [8] using PRISM weather data, latitude, and day length [25, 2]. The main drawback of weather and phenology models is the difficulty to account for local adaptation to climate, or short-term acclimation to interannual weather fluctuations [6]. In addition, spatially and temporally fine grain temperature measurements are not equally available everywhere.

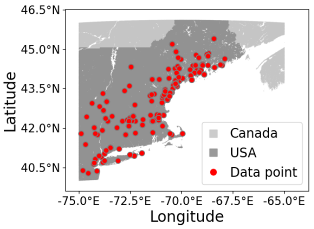


Figure 1: Study region (ground truth labels in red)

Second, we can leverage *satellite data*, such as multispectral reflectance measurements from the MODIS Terra satellite. In particular, we can use the well-calibrated 8-day surface reflectance product, available globally at a 500 m resolution and 8-day frequency but daily granularity [28] (App. A.3). We also consider using annual phenology metrics constructed from MODIS data, such as the onset of greenness, or greenup midpoint (mid-greenup), which capture different stages of seasonal vegetation development [15]. A priori, satellite data could capture the onset of flowering more precisely than weather data by relying on characteristic transitions in reflectance that could generalize more flexibly. The main challenge lies in extracting relevant features from multidimensional data, which we aim to do with deep learning models.

Evaluation Goals and Metrics To evaluate our models' performance in capturing interannual variations in the timing of flowering at any given location, we assign entire years to training, validation, and test sets. We test on the last years (2020-2022) and two randomly drawn years (2015, 2016). This way, we can evaluate the models' extrapolation to "unseen" years, without relying on spatial correlation in phenology. We measure model performance using the root mean squared error (RMSE). To further evaluate the generalizability of our models, we evaluate their performance "out-of-region" at all flowering label locations across the continental US, excluding our training region.

3 Modeling Approaches

Baselines We consider three baselines. Our null model uses the *average* date of flowering onset across the training set labels as a constant prediction. Our second baseline is guided by prior observations that plants flower when exposed to a specific quantity of heat [21]. We fit a simple *threshold* model by finding the degree day base, threshold of ADDs, and start date of accumulation, which best predict the onset of flowering. Fig. 2 illustrates the procedure of RMSE minimization for base 10°C and start date of January 1. Our strongest baseline uses the *First Bloom Index* to predict flowering onset [2], a well-documented reference in the literature [14]. As a caveat, the First Bloom Index is created using flowering labels from the same data source that we use. It relies on three other spring-flowering shrub species, and is not meant to strictly reflect Common Lilac phenology. Conversely, it is trained on data from all years and locations, which creates pseudo test set leakage in its evaluation, and could raise the bar for our other models, which do not see test years and full US.

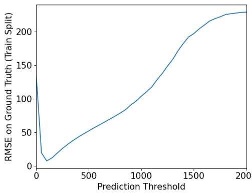


Figure 2: RMSE, base 10°C ADD thresholds.

Domain Expertise Models We develop multiple models based on domain expertise, which leverage the two types of input data available. First, we construct models based on *weather data*, using daily time series of ADDs (above 0°C, 1°C, 5°C, 10°C, and between 0°C and 5°C over the previous year to capture cold temperature accumulation), weekly precipitation, latitude, and longitude. We use elasticnet regression for feature selection, with cross-validation to pick hyperparameters, and we evaluate models with and without interaction terms. These models have the advantage of being interpretable. Fig. 3 shows coefficients from a model that uses as predictors ADDs above 5°C for each day of the calendar year: greater accumulation of warm temperatures in winter is associated with later flowering onset (positive coefficients), and greater accumulation of warm temperatures in early spring and early summer is associated with earlier flowering onset (negative coefficients). Second, we construct models based on *satellite data*, using pre-processed annual phenology metrics such as mid-greenup. We fit both linear models and traditional machine learning models such as random forests, elasticnet, and support vector regressions, with feature selection to improve generalization and cross-validation to pick hyperparameters.

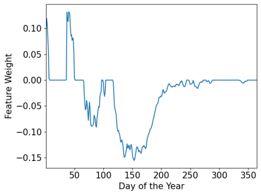


Figure 3: Elastic-net weights for ADD base $5°C$.

Deep Learning Models To leverage more complex but widely available *satellite data*, we develop two deep learning architectures. While other models directly predict the date of first flowering, deep learning models predict a time-series of binary "pre/post flowering onset" from which the onset date is inferred. We want the models to learn temporal patterns across the different spectral bands, which matter for predicting the timing of flowering. Our intuition is that there is a sharp temporal transition in what a landscape looks like around that event. We thus base our models on two popular time-series architectures: a one-dimensional (over time) ResNet [16], inspired by a "non-causal" Temporal Convolutional Network [3]; and a Temporal Fusion Transformer [17]. Appendix B details the architectures and the changes we made to the base models. We train both models from scratch. Since our training data is sparse, we explore transfer learning techniques. Specifically, we pre-train models to predict time series of ADDs and/or flowering predictions from an elasticnet model using ADDs (denoted T). We then fine-tune pre-trained models on ground truth labels (App. B.5 provides further details). After these two steps of training, model predictions only require satellite data as input, and can be extrapolated to regions that lack fine grain temperature data.

	Model	**RMSE (std-dev)** in-region	**RMSE (std-dev)** out-of-region
Baselines	Average	12.25 (N/A)	20.86 (N/A)
	ADD 10°C Threshold	15.54 (N/A)	16.47 (N/A)
	First Bloom Index	9.19 (N/A) $*$	12.63 (N/A) $*$
Domain Expertise Models	Elasticnet (ADD 5°C, precipitation)	**7.64** (0.01)	**12.68** (0.16)
	Random Forests (mid-greenup, lat)	8.15 (0.07)	15.68 (0.15)
	Elasticnet (maturity, lat, lon, elevation)	8.07 (0.00)	38.96 (0.29)
Deep Learning Models	ResNet (GT; includes lat, lon)	9.99 (0.42)	24.33 (0.22)
	ResNet (Transfer: ADD & T)	**8.82** (0.05)	**17.00** (0.09)
	TFT (GT)	**9.76** (0.40)	**19.82** (2.03)
	TFT (Transfer: ADD; includes lat, lon)	10.74 (0.07)	16.68 (0.36)

Table 1: Model performance. Best model overall and best deep learning models are highlighted. In-region uses test years in our training region, out-of-region uses all years on the whole U.S. except the training region. $*$ Test set leakage caveat.

4 Results

Quantitative analysis Table 1 summarizes the performance of our best models (see App. C for more models), and Fig. 4 shows error distributions in the region. We make five observations. First, the First Bloom Index is a good baseline (modulo pseudo test set leakage issues) with an RMSE of 9.19 days, easily beating the null model with a 25% reduction in RMSE. In comparison, the best ADD threshold model (10°C) is not more informative than the null. Second, traditional ML approaches combined with domain expertise can leverage weather data more effectively than the First Bloom Index does. By converting raw temperature to accumulated degree days, optimizing over temperature bases and predictors, and using a well regularized model (elasticnet), we achieve an RMSE of 7.64 days. This is our best model, a 17% improvement over the First Bloom Index and 38% improvement over the null, with a large increase in high precision predictions (Fig. 4). Given the interpretability of elasticnet, this model is promising for downstream applications, for geographies where fine grain temperature data is available. Third, using satellite data to predict flowering is challenging. Using annual expert-guided features calculated from satellite data (mid-greenup or maturity) yields an RMSE around 8 days, which slightly underperforms the best weather model. However, these models perform worse out-of-region, especially the elasticnet model, which could be due to the inclusion of longitude. Fourth, deep learning manages to leverage complex satellite data despite the sparsity of ground truth data. The TFT model trained on full satellite time series using only training ground truth labels—denoted TFT (GT)—yields an RMSE of 9.76 days. It only slightly underperforms the weather-based First Bloom Index, which is a positive surprise, although average-sized errors are more frequent (Fig. 4). Fifth, transfer learning can improve deep learning performance, by pre-training the model to predict temperature data, and fine tuning it on ground truth labels. This yields poor results in-region for the TFT but improves ResNet performance and out-of-region extrapolation, which achieves the best deep learning results. While it doesn't beat our elasticnet ADD model and also performs more variably, it perform 4% better than the First Bloom Index. Because MODIS is available globally, this model can be used to make predictions in any geography. A caveat to this analysis is that we test predictions against plant-level labels instead of averaging labels by exact location and year. In our data, the date of flowering onset can vary across individual plants in a given location and year, with onset dates spanning 4 days on average. This mechanically increases prediction errors.

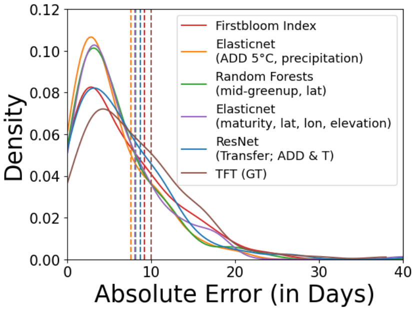
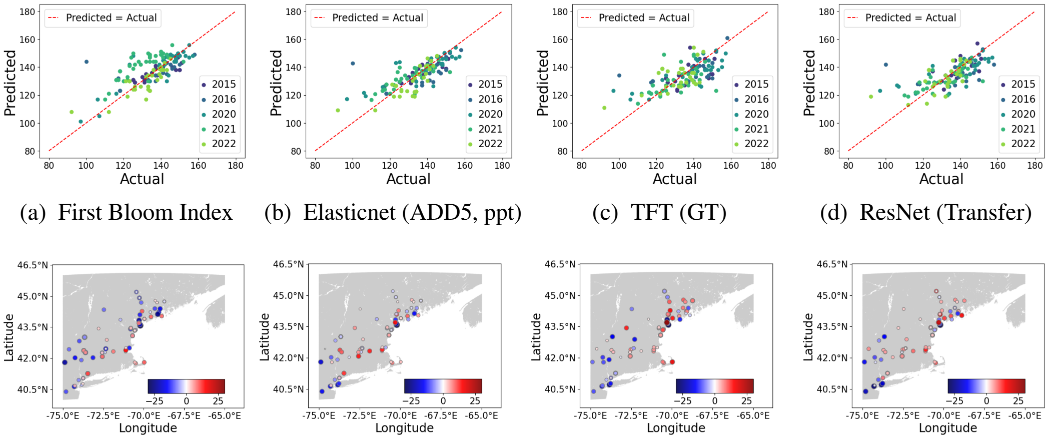


Figure 4: Errors density, RMSE (vert.).

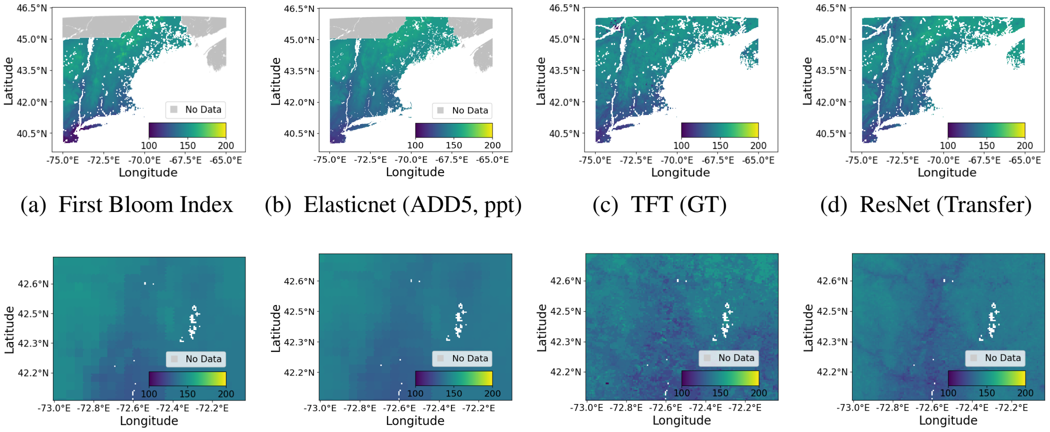
Qualitative interpretation Visualizing predictions and errors over time and space (Fig. 5, App. C) sheds further light on model behavior. We observe that the First Bloom Index errors are correlated



(a) First Bloom Index (b) Elasticnet (ADD5, ppt) (c) TFT (GT) (d) ResNet (Transfer)

(e) First Bloom Index (f) Elasticnet (ADD5, ppt) (g) TFT (GT) (h) ResNet (Transfer)

Figure 5: Prediction vs. ground truth (top row) and error over space (bottom row), across all test years.



(a) First Bloom Index (b) Elasticnet (ADD5, ppt) (c) TFT (GT) (d) ResNet (Transfer)

(e) First Bloom Index (f) Elasticnet (ADD5, ppt) (g) TFT (GT) (h) ResNet (Transfer)

Figure 6: Visualizing the predictions of our models in (test year) 2020. Top row: study region. Bottom row: zooming in around Amherst to show the fine grain spatial behavior of predictions.

within years, with systematic over-prediction in 2020. Other models do not display this pattern, but tend to predict closer to the mean over time (small values are over-predicted, large values under-predicted). This is more pronounced for deep learning models. The ResNet has the widest spatial cluster of good predictions (Fig. 5). All models predict regional variations consistent with the strongest baseline (Fig. 6). However, TFT (GT) predictions have high local variance, whereas ResNet with transfer captures fine grain patterns more credibly.

Next steps Traditional machine learning and to some extent deep learning models can improve our ability to detect the onset of flowering with high precision over large geographies. The flexibility of deep learning models offers exciting new directions, such as leveraging spatial features, and multi-task learning by simultaneously predicting multiple stages of plant development. In addition, this work sets the stage for forecasting models, where improving accuracy by a few days is particularly valuable.

Acknowledgments and Disclosure of Funding

We are grateful for the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [reference number RGPIN-2022-04469]. This research was enabled by computational support provided by the Digital Research Alliance of Canada (alliancecan.ca), and by the University of British Columbia's Advanced Research Computing (UBC ARC).

References

[1] Jill T Anderson, David W Inouye, Amy M McKinney, Robert I Colautti, and Tom Mitchell-Olds. Phenotypic plasticity and adaptive evolution contribute to advancing flowering phenology in response to climate change. *Proceedings of the Royal Society B: Biological Sciences*, 279 (1743):3843–3852, 2012.

[2] Toby R Ault, Raul Zurita-Milla, and Mark D Schwartz. A matlab© toolbox for calculating spring indices from daily meteorological data. *Computers & geosciences*, 83:46–53, 2015.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[4] Beáta Barnabás, Katalin Jäger, and Attila Fehér. The effect of drought and heat stress on reproductive processes in cereals. *Plant, cell & environment*, 31(1):11–38, 2008.

[5] Vijay V Barve, Laura Brenskelle, Daijiang Li, Brian J Stucky, Narayani V Barve, Maggie M Hantak, Bryan S McLean, Daniel J Paluh, Jessica A Oswald, Michael W Belitz, et al. Methods for broad-scale plant phenology assessments using citizen scientists? photographs. *Applications in plant sciences*, 8(1), 2020.

[6] Isabelle Chuine and Jacques Régnière. Process-based models of phenology for plants and animals. *Annual Review of Ecology, Evolution, and Systematics*, 48:159–182, 2017.

[7] Elsa E Cleland, Isabelle Chuine, Annette Menzel, Harold A Mooney, and Mark D Schwartz. Shifting plant phenology in response to global change. *Trends in ecology & evolution*, 22(7): 357–365, 2007.

[8] Theresa M Crimmins, R Lee Marsh, J Switzer, Michael A Crimmins, Katharine L Gerst, Alyssa H Rosemartin, Jake F Weltzin, and S Jewell. *USA National Phenology Network gridded products documentation*. US Department of the Interior, US Geological Survey, 2017.

[9] Anthony Louis D'Agostino and Wolfram Schlenker. Recent weather fluctuations and agricultural yields: implications for climate change. *Agricultural economics*, 47(S1):159–171, 2016.

[10] Christopher Daly, Michael Halbleib, Joseph I Smith, Wayne P Gibson, Matthew K Doggett, George H Taylor, Jan Curtis, and Phillip P Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International Journal of Climatology: a Journal of the Royal Meteorological Society*, 28(15):2031–2064, 2008.

[11] RAF De Réaumur. Observations du thermomere. *Memories Academie Royale Sciences Paris*, pages 545–576, 1735.

[12] Dan J Dixon, J Nikolaus Callow, John MA Duncan, Samantha A Setterfield, and Natasha Pauli. Satellite prediction of forest flowering phenology. *Remote Sensing of Environment*, 255:112197, 2021.

[13] Mark Friedl and Damien Sulla-Menashe. *MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061* [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center, 2022. Accessed June 2023 from `https://doi.org/10.5067/MODIS/MCD12Q1.061`.

[14] Katharine L Gerst, Theresa M Crimmins, Erin E Posthumus, Alyssa H Rosemartin, and Mark D Schwartz. How well do the spring indices predict phenological activity across plant species? *International journal of biometeorology*, 64(5):889–901, 2020.

[15] Josh Gray, Damien Sulla-Menashe, and Mark Friedl. *MODIS/Terra+Aqua Land Cover Dynamics Yearly L3 Global 500m SIN Grid V061* [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center, 2022. Accessed June 2023 from `https://doi.org/10.5067/MODIS/MCD12Q2.061`.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[17] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 2021.

[18] Ariel Ortiz-Bobea and Richard E Just. Modeling the structure of adaptation in climate change impact assessment. *American Journal of Agricultural Economics*, 95(2):244–251, 2013.

[19] Otso Ovaskainen, Evgeniy Meyke, Coong Lo, Gleb Tikhonov, Maria del Mar Delgado, Tomas Roslin, Eliezer Gurarie, Marina Abadonova, Ozodbek Abduraimov, Olga Adrianova, et al. Chronicles of nature calendar, a long-term and large-scale multitaxon database on phenology. *Scientific data*, 7(1):1–11, 2020.

[20] Isaac Park, Alex Jones, and Susan J Mazer. Phenoforecaster: A software package for the prediction of flowering phenology. *Applications in Plant Sciences*, 7(3):e01230, 2019.

[21] Adolphe Quetelet. *Sur le climat de la Belgique*, volume 1. Hayez, 1849.

[22] Alyssa H Rosemartin, Ellen G Denny, Jake F Weltzin, R Lee Marsh, Bruce E Wilson, Hamed Mehdipoor, Raul Zurita-Milla, and Mark D Schwartz. Lilac and honeysuckle phenology data 1956–2014. *Scientific data*, 2:150038, 2015.

[23] William J Sacks, Delphine Deryng, Jonathan A Foley, and Navin Ramankutty. Crop planting dates: an analysis of global patterns. *Global Ecology and Biogeography*, 19(5):607–620, 2010.

[24] Amir Sapkota, Raghu Murtugudde, Frank C Curriero, Crystal R Upperman, Lewis Ziska, and Chengsheng Jiang. Associations between alteration in plant phenology and hay fever prevalence among us adults: implication for changing climate. *Plos one*, 14(3):e0212010, 2019.

[25] Mark D Schwartz, Rein Ahas, and Anto Aasa. Onset of spring starting earlier across the northern hemisphere. *Global change biology*, 12(2):343–351, 2006.

[26] Matthew H Siebers, Rebecca A Slattery, Craig R Yendrek, Anna M Locke, David Drag, Elizabeth A Ainsworth, Carl J Bernacchi, and Donald R Ort. Simulated heat waves during maize reproductive stages alter reproductive growth but have no lasting effect when applied during vegetative stages. *Agriculture, ecosystems & environment*, 240:162–170, 2017.

[27] Michelle J Solga, Jason P Harmon, and Amy C Ganguli. Timing is everything: an overview of phenological changes to plants and their pollinators. *Natural areas journal*, 34(2):227–235, 2014.

[28] Eric Vermote. *MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V061* [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center, 2021. Accessed June 2023 from `https://doi.org/10.5067/MODIS/MOD09A1.061`.

[29] Elizabeth M Wolkovich, Benjamin I Cook, Jenica M Allen, Theresa M Crimmins, Julio L Betancourt, Steven E Travers, Stephanie Pau, James Regetz, T Jonathan Davies, Nathan JB Kraft, et al. Warming experiments underpredict plant phenological responses to climate change. *Nature*, 485(7399):494–497, 2012.

[30] Yelu Zeng, Dalei Hao, Alfredo Huete, Benjamin Dechant, Joe Berry, Jing M Chen, Joanna Joiner, Christian Frankenberg, Ben Bond-Lamberty, Youngryel Ryu, et al. Optical vegetation indices for monitoring terrestrial ecosystems globally. *Nature Reviews Earth & Environment*, 3 (7):477–493, 2022.

A Data Description

A.1 Ground Truth Flowering Labels

Rationale for ground truth data choice and region delineation We look for publicly available data sources on plant flowering that consistently define and record flowering onset, have large spatial and temporal coverage within species, and have coverage that overlaps with weather and satellite data availability. A long-term monitoring program of lilac and honeysuckle phenology provides consistent records of first flowering over several decades and the entire United States, for several species of lilac and honeysuckle [22], which makes it a good candidate. We exclude other data sources that have reasonable spatial coverage, based on inconsistencies in flowering onset definition, insufficient coverage within species, or insufficient temporal coverage. In particular, digitized herbarium records and citizen science datasets like iNaturalist provide abundant records of flower presence and absence; however, spatial coverage comes at the expense of revisit frequency in those datasets, so records are often not frequent enough in a given location and season to allow the derivation of a reliable onset date [5]. We focus on a region in the Northeast US to decrease computational time while maximizing sample size. In addition, different plant species are known to flower at different times of the year. We focus on a single species, and choose the Common Lilac to maximize sample size (Fig. 7). This reduces ground truth label availability to 2006-2023 in our study region.

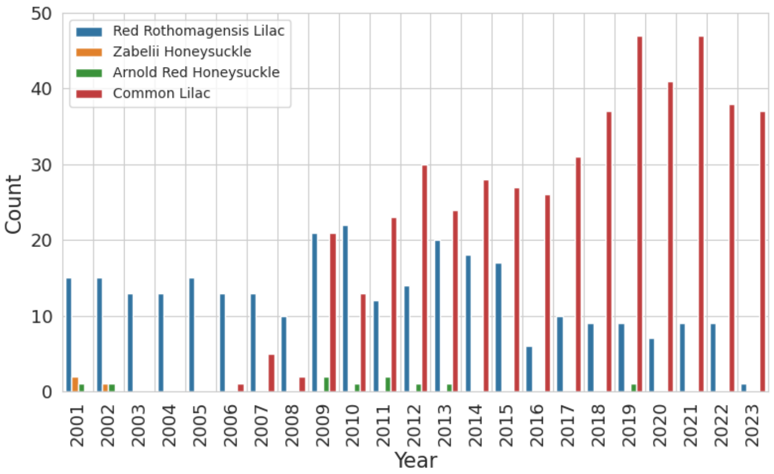


Figure 7: Number of ground truth labels over time, by species, in our study region.

Flowering onset definition Records of first flowering for lilac correspond to the date when at least 50% of lilac flower clusters have at least one open flower, on a given individual plant. Observers participating into the program followed specific protocols for monitoring, recording, and sharing their data, which have been stable over time.

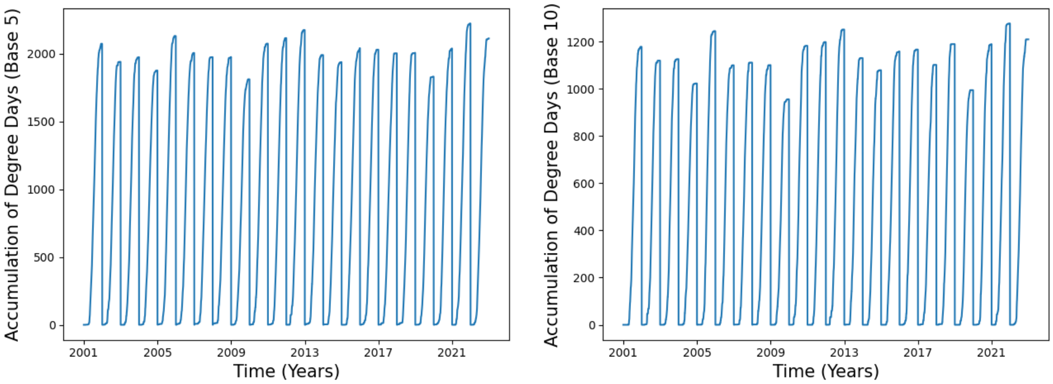
Pre-processing We filter outliers by dropping data points that are beyond 3 standard deviations of the whole-US average. If we follow the outlier analisys from [22] instead, and exclude records with onset dates exceeding 1.5 times the interquartile range for individual plants that have at least 10 observations, we obtain a near-identical dataset. After dropping outliers, if there are multiple records of flowering onset for a given plant and year, we take the earliest date recorded.

Dataset description After pre-processing and filtering pixels based on land cover (Appendix A.3), we have 398 ground truth labels in our study region. These are split between 230 training labels, and 168 test labels (years 2015, 2016, 2020, 2021, 2022).

A.2 Weather Data

Data source Daily weather data comes from the "Parameter- Elevation Regressions on Independent Slopes Model" (PRISM), which covers the contiguous United States at a 4 km resolution. These include daily total precipitation, and daily mean, minimum, and maximum temperature for 1981-2023. A key advantage of PRISM is that it takes into account complex factors driving spatial variation in climate regimes (rain shadows, temperature inversions, coastal effects), which makes it particularly adapted to capturing weather variations at high spatial resolution.

Pre-processing We extract PRISM data over our region and study period using Google Earth Engine, after reprojecting it to the MODIS 500 m resolution. From daily temperature data, we compute growing degree days, which capture how much temperatures exceed a threshold for a chosen period of time [11]. In order to capture important within-day temperature fluctuations, we follow the literature and first infer the within-day temperature distribution from minimum and maximum temperature, using a sinusoidal function; we then compute the fraction of each day spent above the degree day base, e.g. 0°C, which corresponds to the area under the temperature curve above the base threshold throughout the day ([9], see appendix for formulas). Calculating daily ADDs then simply requires to accumulate daily degree day values, starting from a predetermined date. Fig. 8 shows the time series of ADDs using base 5°C and base 10°C, at a sample location.



(a) Base 5°C

(b) Base 10°C

Figure 8: Time series of Accumulated Degree Days (ADDs), at a sample location: (y, x) = (400, 600).

A.3 Satellite Data

Rationale for satellite data choice We look for data sources on satellite imagery, which are publicly available and long-term, to test our models' ability to capture interannual shifts in flowering onset. We also try to maximize image frequency, spatial resolution, and spectral range, as infrared wavelengths in particular can be key for monitoring vegetation characteristics. We choose to use MODIS data, which has the advantages of longevity (it starts in 2001), high image frequency (daily), and large spectral range (spanning the visible to the short-wave infrared). Its main disadvantage is its spatial resolution of 500m, but higher resolution satellite data usually lacks either temporal depth or image frequency (see fig.1 in [30] and table 1 in [12]). It has the additional advantage of being a well calibrated satellite product, which makes it suited to making comparisons over both space and time. We primarily use the MODIS Terra Surface Reflectance 8-Day Global 500m product ([28]), which selects the least cloudy, aerosol-loaded, and zenith-angled images within each 8-day period. This effectively filters images for quality, which we have found helps model learning in our sparse label setting, relative to using lower quality but daily MODIS data. Despite the 8-day frequency of that data product, we can aim for a strictly smaller error in onset prediction. This is because, for each pixel and each 8-day period, we know the exact date at which the "best" MODIS image selected was taken. Our models can thus map spectral reflectance data to exact days, even if temporal coverage is discontinuous.

Data description From the 8-day MODIS product, we extract 7 reflectance bands, spanning the visible to the short-wave infrared; a data quality mask, from which we extract 7 quality variables; a state quality mask, from which we extract 4 variables capturing cloud state, land/water, aerosol quantity, and cirrus cloud presence; and day of year (image date within the 8-day period). We add land cover information from the MODIS Land Cover Type Yearly Global 500m product, which is derived using supervised classifications of MODIS Terra and Aqua reflectance data ([13]). Fig. 9 shows a representation of the satellite data: in each 8-day period, only one day has data (highlighted in pink), and this day can vary for each period.

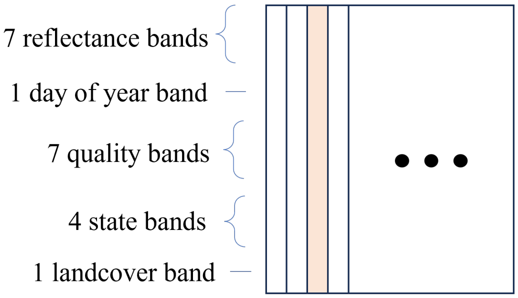
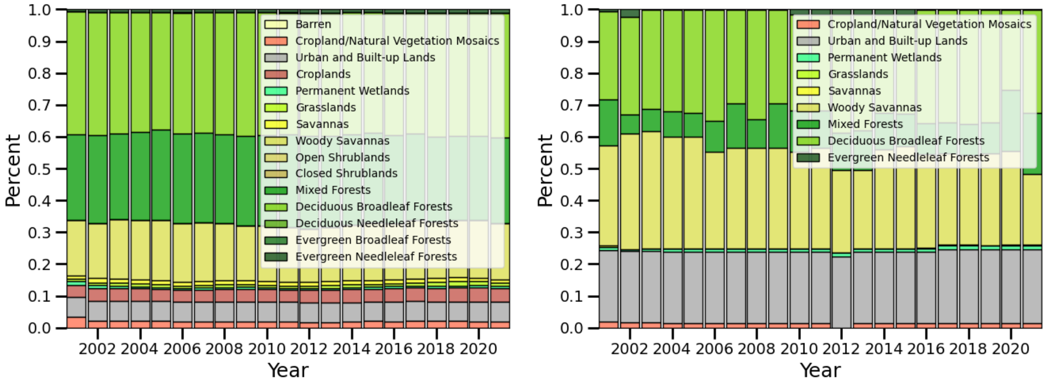


Figure 9: Satellite data [28] layout.

Pre-processing We download all satellite data using Google Earth Engine's python API. We mask out pixels classified as "water" or "snow", based on their dominant land use type throughout the study period. We also mask out pixels with missing satellite or weather data. Fig. 10 shows the distribution of land covers across our study region, and across pixels where there are ground truth labels. There is good representation across ground truth pixels of the main land cover types present in the region (aside from agriculture). This gives us hope that models can extrapolate from the limited set of labels to the entire region.



(a) All pixels

(b) Ground truth pixels

Figure 10: Land cover types across our study region, and across ground truth pixels (water and snow are excluded).

B Deep Learning Architecture Details

B.1 Input Transformation

Recall that the input consists in several reflectance measurements and other features for $N = 30$ observations, each summarizing an 8-day period between January and August for a given year, depicted in Fig. A.3. To preserve the exact date of observations and work with consistent input sizes, we first expand the data into a daily time series of T days (Fig. 11), in which observed data (pink) is copied to the corresponding day. The remaining data (white) is padded with zeros, while the landcover data and day of year data are applied to all days since the landcover value is annual and the day of year data is known.

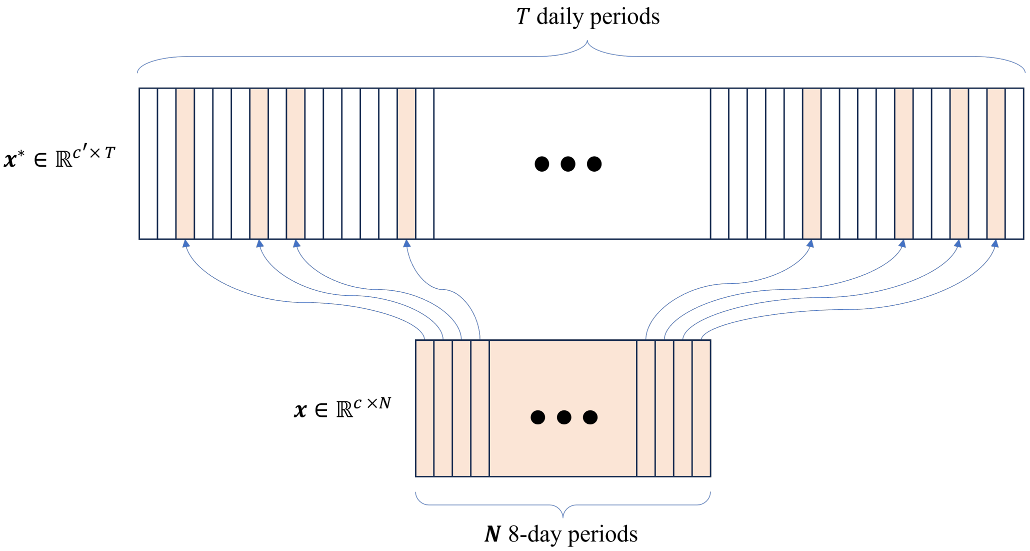

T daily periods
$x^* \in \mathbb{R}^{c' \times T}$
$x \in \mathbb{R}^{c \times N}$
N 8-day periods

Figure 11: Input expansion from 8-day periods to a daily time series. Data is in pink, 0-padding in white.

In the case of ResNet/TCN, each categorical variable is embedded into a vector of trainable values. Since the day of year band is ordinal, we include both the raw and embedded version. We also append a binary feature describing whether the day corresponds to data from product [28] or whether it is zero padding.

In the case of TFT models, each categorical variable is embedded into a vector of trainable values $x_{i,t} \in \mathbb{R}^{d_{model}}$, where d_{model} is a hyper-parameter of the model. The weights for all quality bands are shared across time since they carry the same meaning. The continuous variables are linearly transformed to vectors $x_{i,t} \in \mathbb{R}^{d_{model}}$. We also include both transformed and original value of the day of the year.

B.2 Prediction Encoding

Both our TFT and TCN/ResNet models predict a one dimensional time series at the same granularity as the input (daily for us). We then put each daily prediction in a sigmoid activation, which outputs a number in $[0, 1]$, which we interpret as a probability of being post-flowering for that day. We use a binary cross-entropy loss (summed over each day). The label is hence encoded as zeros until the day before flowering, and a one on the day of flowering and for the rest of the year.

B.3 TFT

Our TFT follows the original architecture [17], using 1 attention head, and d_{model} of 128. We considered landcover information as static metadata. We introduce four changes. First, we have two separate variable selection networks in the TFT (each with shared weights): one for days with actual data [28], and one for days for which we don't have data. This way, the model can make independent selection decision for real or extrapolated days. Second, we add an "embedding convolution" between the variable selection layer and the LSTM layer (in light blue on Fig. 12), consisting of a single depth-wise time convolution with a kernel size of 35. Third, we add a post-processing time convolution, which is a one-dimensional convolution producing a single value per time step, and is applied just before the sigmoid activation. Fourth, skip connections use a weighted sum (instead of a plain sum) with learning weights. Empirically, this helps alleviate instabilities during training.

B.4 ResNet and TCN

We start from a TCN [3], using 1D residual "causal" convolutional blocks, that is a time convolution in which the output depend only on past data, depicted in Fig. 13a. Since we predict flowering a

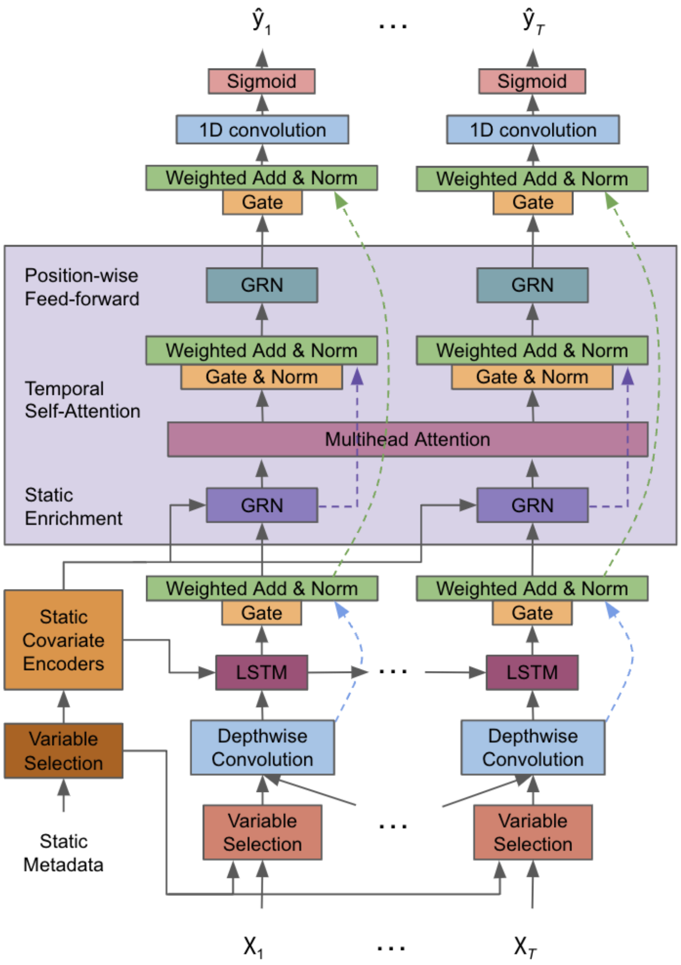
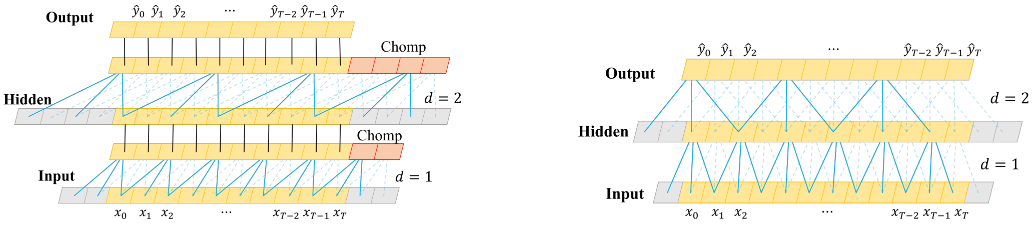


Figure 12: TFT architecture.


Output
$\hat{y}_0$ $\hat{y}_1$ $\hat{y}_2$ ... $\hat{y}_{T-2}$ $\hat{y}_{T-1}$ $\hat{y}_T$
Chomp
Hidden
$d = 2$
Chomp
Input
$d = 1$
x_0 x_1 x_2 ... x_{T-2} x_{T-1} x_T
$\hat{y}_0$ $\hat{y}_1$ $\hat{y}_2$ ... $\hat{y}_{T-2}$ $\hat{y}_{T-1}$ $\hat{y}_T$
Output
$d = 2$
Hidden
$d = 1$
Input
x_0 x_1 x_2 ... x_{T-2} x_{T-1} x_T

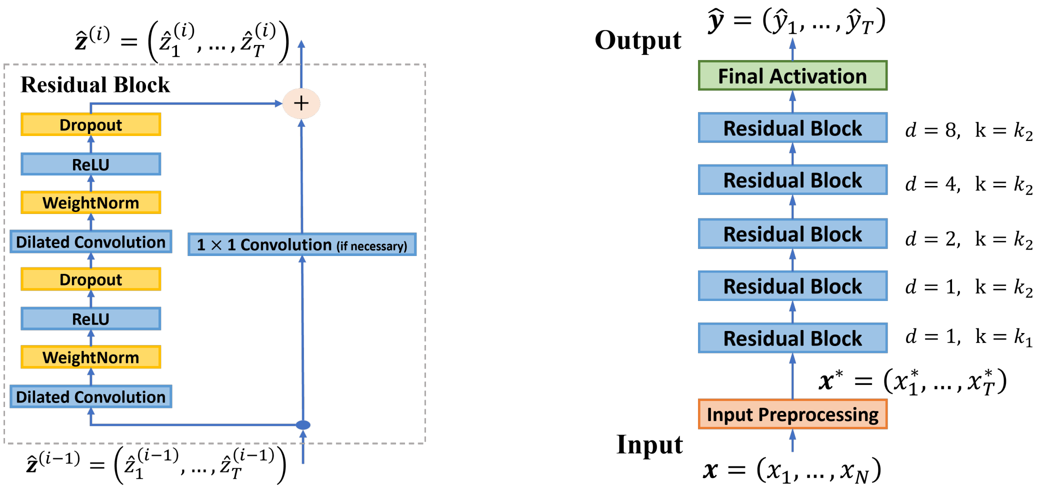
(a) TCN

(b) ResNet

Figure 13: Causal and non-causal convolution.

posteriori (detection), we do not need a causal convolution, and also try a classic convolution variant. We describe it as a time based version of the ResNet [16]. Fig. 13b shows the difference. The convolution layers are combined into a residual block (Fig. 14a), which are stacked to make the final architecture (Fig. 14b).

We find that on all our tasks, the ResNet version performs better than the TCN. This is consistent with the fact that we do flowering onset detection *a posteriori*, which can rely on data from the future, and not forecasting which could rely only on past data. Since the ResNet always outperforms the TCN, we only report ResNet results.


$\hat{\mathbf{z}}^{(i)} = \left(\hat{z}_1^{(i)}, \ldots, \hat{z}_T^{(i)}\right)$
Residual Block
+
Dropout
ReLU
WeightNorm
Dilated Convolution
1 × 1 Convolution (if necessary)
Dropout
ReLU
WeightNorm
Dilated Convolution
$\hat{\mathbf{z}}^{(i-1)} = \left(\hat{z}_1^{(i-1)}, \ldots, \hat{z}_T^{(i-1)}\right)$
Output
$\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_T)$
Final Activation
Residual Block
$d = 8,\ \mathrm{k} = k_2$
Residual Block
$d = 4,\ \mathrm{k} = k_2$
Residual Block
$d = 2,\ \mathrm{k} = k_2$
Residual Block
$d = 1,\ \mathrm{k} = k_2$
Residual Block
$d = 1,\ \mathrm{k} = k_1$
$\boldsymbol{x}^* = (x_1^*, \ldots, x_T^*)$
Input Preprocessing
Input
$\boldsymbol{x} = (x_1, \ldots, x_N)$

(a) Residual block

(b) ResNet

Figure 14: Convolution based architectures. In a residual block, the input has shape $\hat{\mathbf{z}}^{(i-1)} \in \mathbb{R}^{c_{i-1}\times T}$ and the output has shape $\hat{\mathbf{z}}^{(i)} \in \mathbb{R}^{c_i \times T}$. Should $c_{i-1} \neq c_i$, then a 1×1 convolution is necessary. Both the TCN and ResNet use increasing dilation sizes d, in addition to a larger first convolution filter size $k_1 = 33$ in the first layer which is then reduced to $k_2 = 7$ in the remaining layers.

B.5 Training

We study two training approaches: using only labeled data (ground truth), and transfer learning from intermediary, widely available outcomes.

Training on ground truth only (GT). The first training approach is to train our deep learning model only using spacial point for which we have labels for the onset of flowering. We train the TFT for 1000 epochs using the SGD optimizer with "ExponentialLR" learning rate scheduler starting at 0.1 with gamma 0.9989 and with early-stopping, and the ResNet for 400 epochs using SGD (weight decay 0.0003, Nesterov momentum of 0.5, momentum decay of 0.02) with the "ReduceLROnPlateau" learning rate scheduler (learning rate starts at 0.1, with at most four reductoins of 0.5), and a batch size of 256 with early-stopping and dropout of 0.2.

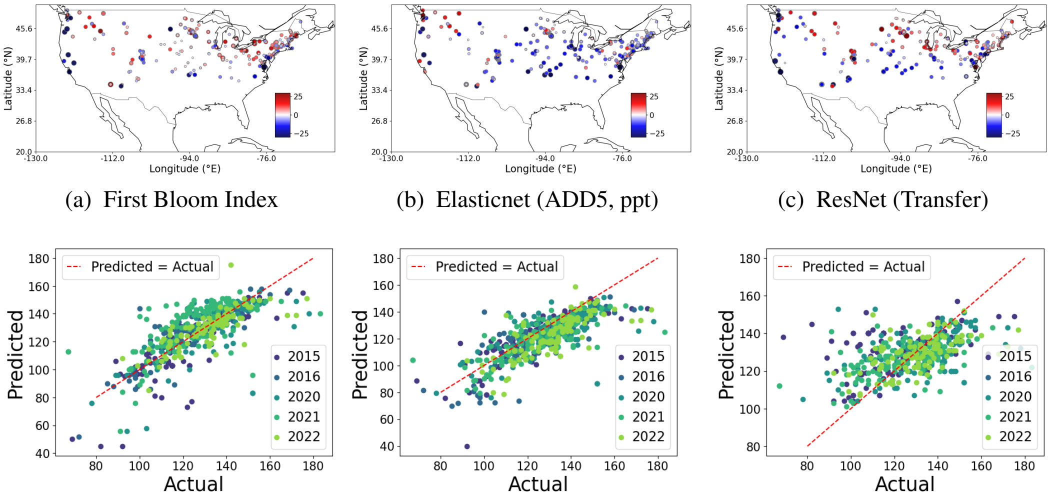
Transfer Learning (Transfer). We also consider transfer learning to alleviate the sparsity of labeled data. In this setting, we first train our model on one or more intermediary outcomes. The first intermediary prediction we consider is the ADD time-series, since it is the most informative feature for traditional ML (noted Transfer; ADD). We also consider the predictions of our ADD based model, which uses the base 5 ADD time-series to predict the onset of flowering (noted Transfer; T). Transfer learning proceeds in two steps. First, we train our deep learning model (TFT or ResNet) to predict the intermediary outcome (whole time series) from its input data. We use the same hyper-parameters as above. In this first step, we train the TFT for 5 epochs, and the ResNet for 4 epochs.

In the second step, we start from our first step model, and fine tune it on labeled data to predict the onset of flowering. In this second step, we train the TFT for 100 epochs using the SGD optimizer, and the ResNet for 400 epochs using the SGD optimizer.

C Additional Results

In this section, we document model performance across a larger set of models (Table 2), along with a visualization of predictions and errors over space (18) and time (Fig. 19). We also include predictions of our core baselines and models in each test year (Fig. 17).

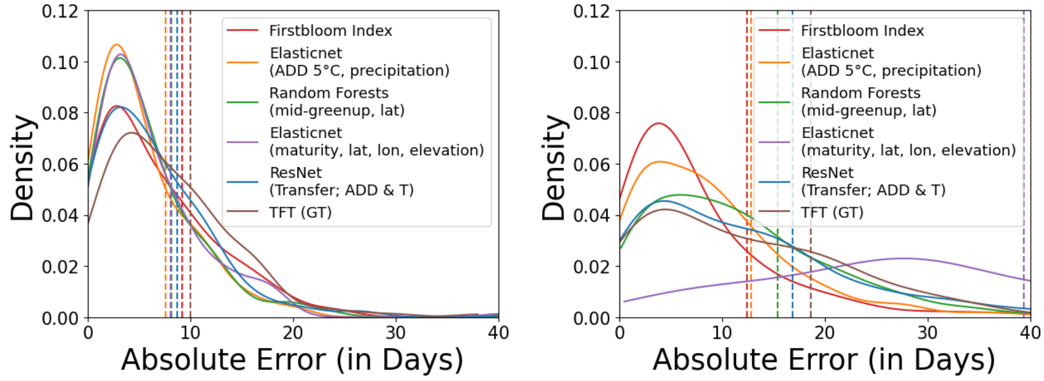
To understand the generalization of our models outside of their training region, we evaluate each model on ground truth data on the whole US (outside of the training region), for all years for which



(a) First Bloom Index (b) Elasticnet (ADD5, ppt) (c) ResNet (Transfer)

(d) First Bloom Index (e) Elasticnet (ADD5, ppt) (f) ResNet (Transfer)

Figure 15: Prediction vs. ground truth (top row) and error over space (bottom row), across all test years.



(a) Errors density and RMSE (vert.), in region for test years.

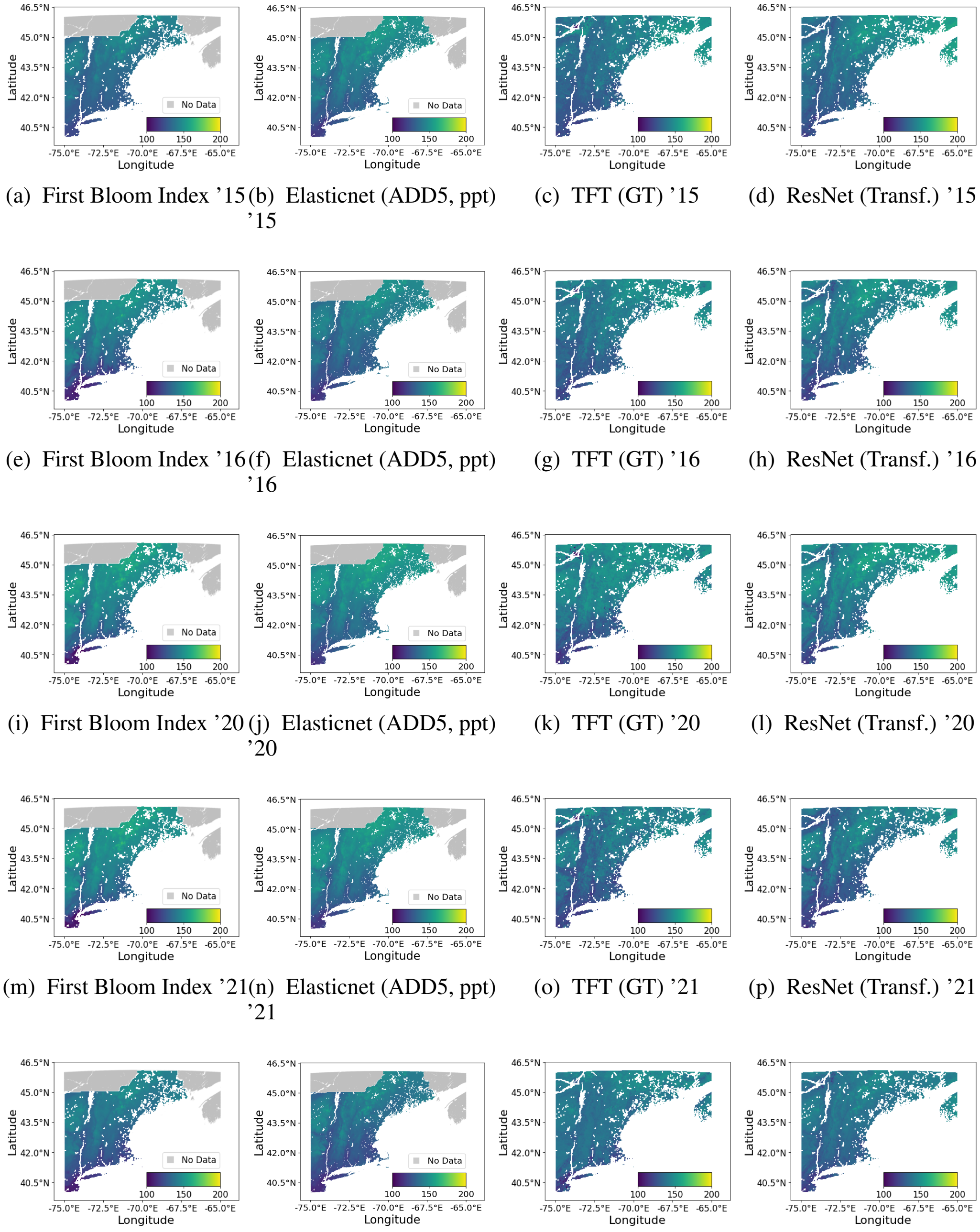
(b) Errors density and RMSE (vert.), out of training region for all years.

Figure 16: Prediction vs. ground truth (top row) and error over space (bottom row), across all test years.

we have data. This shows the performance of different models on geographies never seen during training. Table 2 shows the RMSE and standard deviation for each model. We observe that some domain expertise models generalize quite well. Most notably, the Elasticnet model based on ADD 5°C and precipitation performs as well as the First Bloom Index, a baseline from the literature that was trained on out-of-region data. Other models however clearly overfit their training region, and do not generalize well to other geographies. This is the case of the Elasticnet model trained on maturity, lat, lon, and elevation. Deep learning models trained solely on ground truth tend to overfit as well, but those using transfer learning do generalize quite well over different geographies. This is a promising sign for scaling fine grain flowering onset prediction using satellite data. Figure 15 shows the geographical distribution of errors for the models that generalize best, as well as their distribution compared to the ground truth. Finally, Figure 16 shows the density of prediction errors out-of-region, compared to that in-region. We can observe that out of region predictions are larger and more dispersed. Good models have a density concentrated over low errors, with a larger tail than in region. Geographically overfitted models are not able to predict well out of region and show a large tail of high errors.

	Model	**RMSE (std-dev)** in-region test years	**RMSE (std-dev)** out-of-region all years
Baselines	Average	12.25 (N/A)	20.86 (N/A)
	ADD 0°C Threshold (optimal threshold value=550)	16.73 (N/A)	15.65 (N/A)
	ADD 1°C Threshold (optimal threshold value=450)	16.89 (N/A)	17.23 (N/A)
	ADD 5°C Threshold (optimal threshold value=250)	15.77 (N/A)	16.47 (N/A)
	ADD 10°C Threshold (optimal threshold value=100)	15.54 (N/A)	16.47 (N/A)
	ADD 15°C Threshold (optimal threshold value=50)	18.72 (N/A)	86.18 (N/A)
	First Bloom Index	9.19 (N/A) ∗	12.63 (N/A) ∗
Domain Expertise Models (annual phenology features)	Random Forests (mid-greenup, lat)	8.15 (0.07)	15.68 (0.15)
	Random Forests (mid-greenup, lat, lon)	8.22 (0.07)	15.00 (0.14)
	Random Forests (greenup, mid-greenup, lagged dormancy)	9.40 (0.12)	16.21 (0.39)
	Elasticnet (maturity, lat, lon, elevation)	8.07 (0.00)	38.96 (0.29)
Domain Expertise Models (daily weather features)	Elasticnet w/ interactions (ADD 5°C, ADD 0-5°C, precipitation)	7.57 (0.02)	15.17 (0.87)
	Elasticnet (ADD 5°C, ADD 0-5°C, precipitation)	7.64 (0.02)	12.66 (0.25)
	Elasticnet (ADD 5°C, precipitation)	7.64 (0.01)	12.68 (0.16)
	Elasticnet (ADD 5°C, ADD 0-5°C, precipitation, lon)	7.66 (0.04)	13.64 (0.10)
Deep Learning Models	ResNet (GT)	11.60 (2.04)	20.12 (0.26)
	ResNet (GT; includes lat, lon)	9.99 (0.42)	24.33 (0.22)
	ResNet (Transfer; T)	10.28 (1.00)	20.75 (3.54)
	ResNet (Transfer; T; includes lat, lon)	10.99 (0.67)	25.40 (3.46)
	ResNet (Transfer; ADD)	8.86 (0.09)	17.15 (0.09)
	ResNet (Transfer; ADD; includes lat, lon)	9.42 (0.12)	17.80 (0.07)
	ResNet (Transfer; ADD & T)	8.82 (0.05)	17.00 (0.09)
	ResNet (Transfer; ADD & T; includes lat, lon)	9.74 (0.08)	16.17 (0.06)
	TFT (GT)	9.76 (0.40)	19.82 (2.03)
	TFT (GT; includes lat, lon)	10.22 (0.12)	16.88 (0.90)
	TFT (Transfer; ADD)	11.60 (0.80)	17.65 (0.18)
	TFT (Transfer; ADD; includes lat, lon)	10.74 (0.07)	16.68 (0.36)

Table 2: Model performance across a wider set of models. Models highlighted in gray are included in Table 1. ∗ Test set leakage caveat.



(a) First Bloom Index ’15 (b) Elasticnet (ADD5, ppt) ’15 (c) TFT (GT) ’15 (d) ResNet (Transf.) ’15

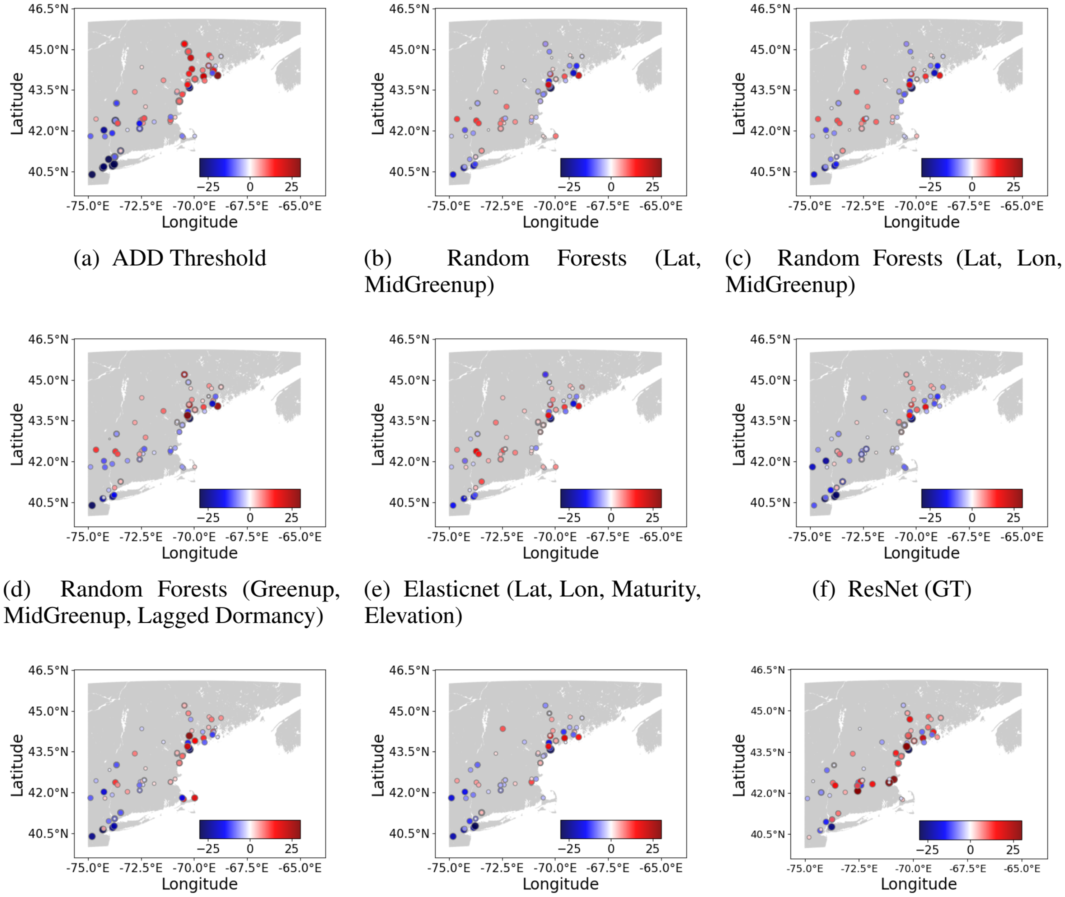
(e) First Bloom Index ’16 (f) Elasticnet (ADD5, ppt) ’16 (g) TFT (GT) ’16 (h) ResNet (Transf.) ’16

(i) First Bloom Index ’20 (j) Elasticnet (ADD5, ppt) ’20 (k) TFT (GT) ’20 (l) ResNet (Transf.) ’20

(m) First Bloom Index ’21 (n) Elasticnet (ADD5, ppt) ’21 (o) TFT (GT) ’21 (p) ResNet (Transf.) ’21

(q) First Bloom Index ’22 (r) Elasticnet (ADD5, ppt) ’22 (s) TFT (GT) ’22 (t) ResNet (Transf.) ’22

Figure 17: Visualizing the predictions of our models in test years 2015, 2016, 2020, 2021, 2022.



(a) ADD Threshold

(b) Random Forests (Lat, MidGreenup)

(c) Random Forests (Lat, Lon, MidGreenup)

(d) Random Forests (Greenup, MidGreenup, Lagged Dormancy)

(e) Elasticnet (Lat, Lon, Maturity, Elevation)

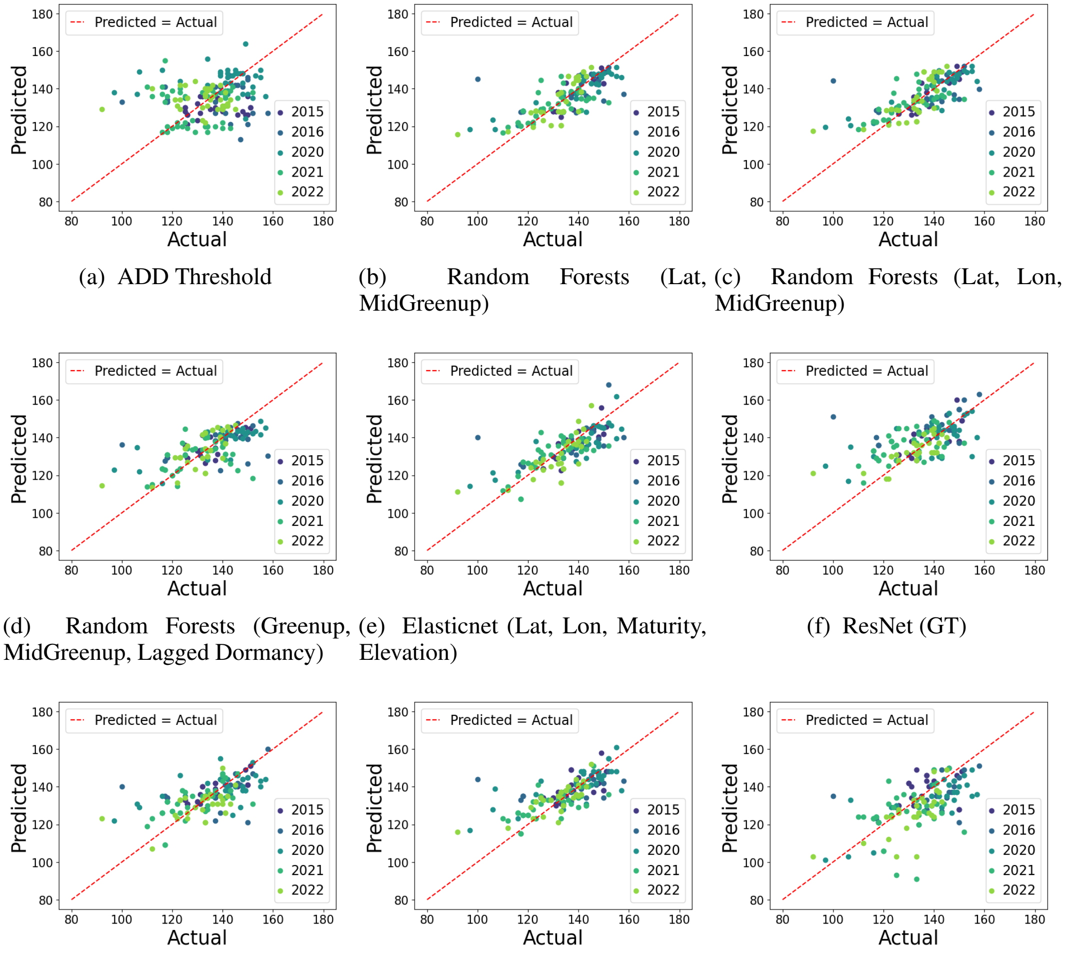
(f) ResNet (GT)

(g) ResNet (Transfer; T)

(h) ResNet (Transfer; ADD)

(i) TFT (Transfer; ADD)

Figure 18: Model error over space, combining all test years (additional models).



(a) ADD Threshold

(b) Random Forests (Lat, MidGreenup)

(c) Random Forests (Lat, Lon, MidGreenup)

(d) Random Forests (Greenup, MidGreenup, Lagged Dormancy)

(e) Elasticnet (Lat, Lon, Maturity, Elevation)

(f) ResNet (GT)

(g) ResNet (Transfer; T)

(h) ResNet (Transfer; ADD)

(i) TFT (Transfer; ADD)

Figure 19: Prediction vs. ground truth across all test years (additional models).