Creating a workflow in Python to predict leaf emergence

A case study of the Eastern US

MSc. GIMA Master's Thesis

Final Thesis Report

Lars Kagie (0649562)

l.kagie@students.uu.nl

MSc. Geographical Information Management and Applications

University of Utrecht

20-05-2024

Supervisor: Mahdi Khodadadzadeh (University of Twente)

Responsible professor: Raul Zurita Milla (University of Twente)

Contents

Acknowledgments

First of all, I would like to thank my supervisor Mahdi Khodadadzadeh for helping me through this process. He was always ready to help and had suggestions and feedback for me to write my Thesis. Additionally, I would like to thank two developers of Springtime who helped me through writing the code: Peter Kalverla and Francesco Nattino. Springtime is a relatively new package and I was the first to use it properly so I ran into some problems. They helped me in that aspect. Without them, this definitely would not have been possible.

Abstract

Aim: The ability to predict the phenological stages of different species can have implications for ecosystems around the world. Many process-based prediction models (PB) have been developed, attempting to mimic the biological mechanism happening in different plant and tree species. However, no model has been proven superior to others, due to the complexity of these mechanisms. Recent studies have shown that machine learning (ML) can be used to predict phenological stages. Both approaches have advantages and disadvantages. The main challenge is the lack of a unified framework so that approaches can be implemented and compared fairly. In this study, a workflow is developed, at the hand of a case study, that can be repeated for different data, different phenological stages, and different models.

Innovation: To develop said workflow a Python package called Springtime is used, which is a project initiated by the University of Twente. Springtime will help streamline workflows and facilitate ML methods regarding phenological datasets. It allows for easier feature extraction and usage of different datasets and phenology prediction models. This is the first time Springtime is used to build such a workflow. A case study for the Eastern US with observation data from 2016-2022 is used to predict the leaf emergence date of deciduous trees for three PB models and three ML models. The thermal time (TT) model, sequential (S) model, unichill (UC) model, gradient boosted classifier model (GB), random forest model (RF), and support vector machine model (SVM).

Main conclusions: The development of a workflow was successful. For the ML methods, meteorological data and observational data were merged, and binary classified and features were extracted. The RMSE was calculated for all six models and the PB models performed better than the ML models by approximately 3 days. Among the PB models, the TT model performed best, while among the ML models, the SVM model showed the best performance. Overall, the TT model outperformed all the models. The developed workflow can be used in future research for different data and different PB and ML models.

Introduction

Lieth (1974) defined phenology as: "*The study of the timing of recurrent biological events, the causes of their timing with regard to biotic and abiotic forces, and the interrelation among phases of the same or different species (p. 295)*". The timing of recurrent biological events in vegetation, such as spring vegetation growth onset, or leaf emergence, has been significantly advanced by global warming (Liu et al. 2018). This change already has had serious implications for global carbon uptake, forest productivity, species distribution, plant-animal interactions, and the structure and function of ecosystems (Dai et al. 2019). Additionally, it can influence carbon, water, and energy exchanges between terrestrial ecosystems and the atmosphere, including increased evapotranspiration and carbon storage capacity (Fu et al. 2020). The timing of leaf emergence also plays a critical role in global carbon assimilation and heterotrophic respiration dynamics (Delpierre et al. 2009). Year-per-year variations in vegetation phenology result in significant differences in forest productivity and carbon sequestration, with early leaf emergence potentially leading to frost damage and reduced productivity (Caffarra et al. 2011). Therefore, accurate predictions of vegetation phenology, especially leaf emergence, under climate change are crucial for understanding land surface processes, global carbon and water cycles, and mitigating the impacts of rising temperatures (Caffarra et al. 2011). Phenology is mostly explained through biology, but Visser et al. (2010) explain that ecologists, physiologists, and epidemiologists also can be involved in the subject. Due to previously explained impacts however, it can be stated that phenology as a concept involves many more disciplines such as climate science, human health, and urban planning, again highlighting the importance of gaining knowledge on the subject of leaf emergence and our ability to predict it.

Different environmental drivers can affect leaf emergence in spring. Fu et al. (2020) explain these as (i) temperature, (ii) photoperiod, (iii) precipitation, (iv) air humidity, and (v) nutrients in the soil. Usually, the timing of leaf emergence is attributed to spring temperatures. Many phenology prediction models, taking different environmental factors into account, have been developed. Examples are (i) the Thermal time model (Robertson, 1968), (ii) the Spring warming model (Sarvas, 1974), (iii) the sequential model (Sarvas, 1974), (iv) the Parallel model (Landsberg, 1974), (v) the Alternating model (Cannel and Smith, 1989), (vi) the Four phase model (Vegis, 1964), and (vii) the Deepening rest model (Kobayashi et al., 1982). These are older models and do not guarantee a sufficient prediction. Many newly developed models are derived from or based on these older models, such as the Unified model (Chuine, 2000), the Unichill model (Chuine, 2000), and the DORMPHOT model (Caffarra et al., 2011). Recent developments include ML methods to predict different vegetation phenology stages, such as the backpropagation neural network, the support vector machine, and the random forest model (Guo et al. 2021). All these models can be divided into two categories:
PB models and ML models. PB models, such as the DORMPHOT model, can reflect nonlinear responses of phenology to temperature or other environmental factors. These get increasingly complex with more factors and parameterization becomes more difficult. ML methods are statistical and predominantly data-driven, which means that they depend on the quality and amount of data available.

Not one single approach has proven unequivocally superior, and prediction under broad spatial domains and different climate regimes remains challenging (Keenan and Richardson, 2015). The lack of standardization makes it difficult to compare and test models, which has slowed down literature on this subject (Chuine, 2000). Different models can have different outcomes in the same study area, due to the many different processes that can affect leaf emergence. To date, it has not yet been possible to perfectly simulate these processes and develop a model that can predict leaf emergence without a margin of error. Therefore, there is a necessity for a standardized workflow. Delpierre et al. (2009) describe two reasons as to why there is no consensus on a superior model. First, it is seen as difficult to measure parameters directly. Most case studies are focused on a particular species or area, which is why they often cannot be applied to a larger scale. Furthermore, there is often a lack of observational data which impedes the validation process. In the past, data acquisition in this context relied primarily on in-situ observations, making it difficult to accumulate large amounts of data. While remote sensing has gained considerable attention, it struggles with challenges in precisely determining the timing of leaf emergence (Macbean et al. 2015). The primary hurdle lies in the resolution limitations of remote sensing data. This hinders the acquisition of sufficiently detailed information, which is crucial for accurate assessments of the timing of leaf emergence.

Recognizing the challenges associated with predicting leaf emergence on a broader scale, efforts are underway to develop innovative solutions that can improve the accuracy and efficiency of such predictions. As a consequence, the University of Twente is developing a project that will be able to deal with some of these challenges. Springtime (Kalverla et al., 2024) is a Python package that will help streamline workflows and facilitate ML methods regarding phenological datasets. It allows for easier feature extraction and usage of different datasets and phenology prediction models. As mentioned above, the prediction of phenological events is a complex endeavor due to a variety of factors. Springtime was developed to ease this process. The main aim of this Thesis will be to use Springtime in combination with different phenological datasets and models to develop a workflow, through a case study, that makes predicting leaf emergence less complex and will allow researchers to get a better understanding of what factors influence this phenological event. The main packages that will be used for modeling are the pyPhenology (PB models) and scikitlearn (ML models), which are incorporated into Springtime itself. This paper will be the first step towards this research by stating research objectives and corresponding research questions, then laying out a theoretical framework that forms the foundation of the thesis, explaining the methodology, and lastly showing and discussing the results.

Research objectives

There is a lack of standardization in phenology prediction models, making comparisons between different models difficult on a broader scale. The main goal of this study is to establish a workflow using the Springtime package, facilitating comparisons across feature extraction methods, PB models, and ML models. This will be done via a case study. This leads to the main research question and sub-questions below.

Main research question

How to tailor a machine learning workflow for phenological modeling using observations from phenology networks and open-access geospatial datasets?

Sub-questions

1. Selecting a Case Study:
    i. What criteria should be used to identify and select a suitable geographic location, period, and species for conducting the case study?
    ii. What specific advantages does using Springtime offer for the development of a case study in phenological modeling?

2. Process Based Models:
    i. What are the key factors of PB phenology models and how do they differ?
    ii. How can Springtime features be used to predict outcomes of PB models at the hand of the pyPhenology package?

3. Machine Learning Models:
    iii. Which specific ML models are most promising for predictive phenological modeling, and how do they differ?
    iv. How can Springtime features be used to predict outcomes of the ML models at the hand of the scikitlearn package?

3. Feature extraction:
    i. What features can be extracted from geospatial datasets for phenological modeling?
    ii. How can features be extracted from a geospatial dataset for further analysis in phenological modeling?

Theoretical Framework

Effect of environmental drivers on vegetation leaf emergence

Many different factors can influence vegetation leaf emergence. These can be divided into environmental, such as temperature or photoperiod, and non-environmental, such as plant genetics and pollination. Important to note is that the focus of this study will be on environmental factors because these lie at the basis of phenology prediction models and more easily can be made measurable.

Currently, many different theories exist on what environmental driver has what effect on leaf emergence in different plant and tree species. Phenology research has a long history, with de Reaumur (1735) already making the connection between temperature and phenology 300 years ago. After this, Merriam (1894) showed that the distribution of animal and plant species is governed by the air temperature of the period of reproductive activity. Even though this statement relates to the distribution of species, it is still among the first research that explains a possible connection between plant growth and temperature. Subsequent research built on this by considering the relationship between plant distribution to periodical activities, temperature, altitude, and latitude (Hopkins, 1920). Furthermore, Livingston and Livingston (1913) soon realized that physiological phenomena in plants cannot be influenced by a single condition or factor. They state that biological science will eventually have to acknowledge that no single condition, or any combination of conditions that includes fewer than all of the successful ones, can be regarded as the identification of natural processes. The analysis of multiple complex environmental factors is necessary.

Even though there was an early realization that there are other environmental drivers that might influence plant phenology, early research still focusses on temperature as the only factor that contributes to different plant phenology stages (Lindsey & Newman, 1956; Voigt & Mohlenbrock, 1964; Holway & Ward, 1965; Evans et al. 1970). This focus can mainly be attributed to the fact that there was a lack of data and information available on other possible environmental drivers of phenology stages. Since 1980 however, the amount of scientific articles related to plant phenology increased by about 10-fold (figure 1) (Tang et al. 2016).

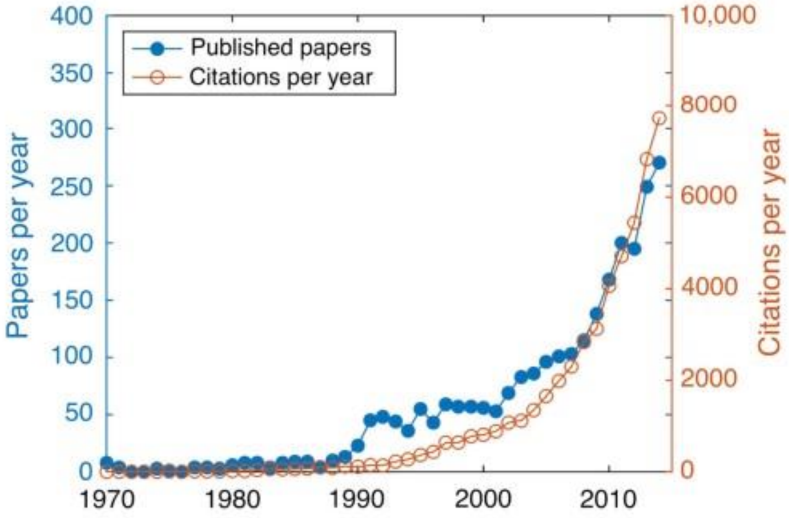


Figure 1. Trends of papers and citations of vegetation phenology each year between 1970 and 2014 (Tang et al. 2016)

Tang et al. (2016) give a reason for this sudden increase, which they state as technological developments in the phenology discipline and therefore a surge in the amount of data and information available. Due to this increase, the amount of articles on specific phenological stages also increases. As mentioned previously, the focus of this thesis will be on leaf emergence. Many articles now include different environmental drivers behind leaf emergence among different species. These include irradiance (Hill et al. 2001; Hill et al. 2002; Saleska et al. 2007), precipitation (Grossiord et al. 2017; Lutcher et al. 2019), and photoperiod (Warrington & Kanemasu, 1983; Ellis et al. 1992; Adams & Langton, 2005).

However, despite this wealth of information, a need persists for a deeper mechanistic knowledge of phenology stages, including its causes, variability, and intricate relationships with other physiological processes at different scales (Tang et al. 2016; Dai et al. 2019). Leaf emergence is still dependent on latitude and variations among different species, hindering the capacity to create prediction models (Basler, 2016). Solving these problems is not only an academic endeavor but also has real-world applications in terms of predicting when leaves will emerge, which is a crucial component of comprehending and potentially lessening the effects of climate change. As we explore the environmental factors that might influence leaf emergence, it becomes clear that understanding the complexity involved in this phenological event requires a nuanced approach.

Process-based Models

PB models regarding phenology aim to simulate underlying biological, physical, and chemical processes and are designed to capture the mechanisms that control the timing of different events such as plant flowering or leaf emergence (Asse et al. 2020), the latter being the focus of this study.

As mentioned before, the concept of degree-day sum was first established by Reaumur (1735), marking the beginning of phenology modeling's lengthy history. Instead of basing plant development on the temperature at the time of the phenological event, he suggested basing it on the total accumulated temperature over time. In modern-day vegetation phenology modeling, this assumption remains crucial. Accompanying the revolution in phenological research regarding environmental drivers around the 1980s, a multitude of PB phenology prediction models was developed. There are too many to explain, but below the most influential and representative are set out.

Thermal time model

The TT model is one of the first models that was developed and assumes that leaf emergence occurs on the day that the accumulated rate of daily heating temperatures exceeds a certain threshold. The article is based on earlier findings that the active period of forest trees almost exclusively depends on the time/temperature factor, as established by Reaumur (1735). The model only has three parameters: d_1 is the start of the accumulation, T_b is the threshold temperature and F is the number of days required to initiate leaf emergence. The time of the leaf emergence, y, in a given year is such that

$$\sum_{d=d1}^{y} Rf(td) = F$$

where,

$$Rf(td) = \begin{cases} 0 & if\ td < Tb \\ td - Tb & if\ td > Tb \end{cases}$$

and where t_d is the temperature on day *d* and R_f is an accumulation function. One of the criticisms however, as explained by Sharratt et al. (1989), lies in the assumption that T_b is constant, which might not be the case since at different development stages, plants respond differently to the same environmental drivers.

Sequential model

Sarvas (1974) built on the growing-degree day model and developed the sequential model, wherein rest and quiescence are considered as two strictly separate phases. Rest is defined as: "*the period in which buds remain dormant due to growth-arresting physiological conditions in the bud itself (p. 172)*" (Kramer, 1994), and quiescence is defined as: "*the period in which the buds remain dormant due to unfavorable environmental conditions (p. 173)*" (Kramer, 1994). A critical threshold of chilling needs to be attained before there is a transition from rest to quiescence, after which the model works the same as the growing-degree day model, meaning that there will be no transition from quiescence to the active phase unless the critical state of forcing is reached. As stated before, the S model builds on the growing-degree day model where the forcing function is given by

$$Rf(td) = \begin{cases} 0 \; if \; td < Tb \\ \dfrac{28.4}{1 + e^{-0.185(td - 18.4)}} \; if \; td > Tb \end{cases}$$

where the value for d_1 is set so that

$$\sum_{d=d0}^{d1} Rc(td) = C$$

where *C* stands for the number of chilling units required and R_c is a chilling function given by

$$Rc(td) = \begin{cases} 0 \; if \; td < Tmin \; or \; td > Tmax \\ \dfrac{td - Tmin}{Topt - Tmin} \; if \; Tmin < td < Topt \\ \dfrac{Tmax - td}{Tmax - Topt} \; if \; Topt < td < Tmax \end{cases}$$

Unified model

Chuine (2000) argues that most established phenology prediction models up until that point are specific cases of a more general model. Consequently, the author developed a unified model that aims to encompass many different models. A unified model with nine parameters (a_c, b_c, c_c, b_f, c_f, *w, k, C*, and* t_c) can be defined. The response to chilling temperature is defined by CF_c *(*a_c, b_c, c_c, x_t*)*, and the response to forcing temperature by CF$_f$ (0, b_f, c_f, x_t). Additionally, forcing units are accumulated once a critical state of chilling (*C**) is reached, and the critical state of forcing needed to burst (*F**) is a function of the total state of chilling reached at t_c ($F^* = we^{kCtot}$). Three main advantages of the unified model are set out. In the first place, it makes it possible to (i) estimate directly how buds expand in response to forcing or chilling temperatures and (ii) when these temperatures have an impact on bud growth. Second, standard statistical tests for any given species can be used to simplify it. Thirdly, it offers a consistent framework for phenological models, which is necessary for reliable model identification and comparison analysis.

DORMPHOT model

The DORMPHOT model (for DORMancy and PHOToperiod) was developed by Caffarra et al. (2011) as somewhat of an extension of the unified model. The main difference between the two models is that the DORMPHOT model counts for the process of dormancy induction by photoperiods in addition to low temperatures. Additionally, the parameter of C_{crit} (C^* in the unified model is not a simple threshold that determines a break in dormancy, but also establishes the form of response function to photoperiod during the ecodormancy phase. Vice versa, the parameter F_{crit} (F^* in the unified model) is considered a constant in the DORMPHOT model rather than being dynamic as in the unified model. Therefore, in contrast to the unified model, which characterizes the reaction as constant, the DORMPHOT model describes the response of cell growth to temperature in the eco-dormancy phase as dynamic and it depends on the state of chilling and on the length of the photoperiod. To account for this increase in complexity the DORMPHOT model uses two additional parameters, adding up to 11 parameters: DL_{crit}, D_{crit}, C_{crit}, F_{crit}, *aD, bD, aC, cC, dF, gT,* and *hDL.* Figure 2 shows a schematic representation of the model (Caffarra et al. 2011).

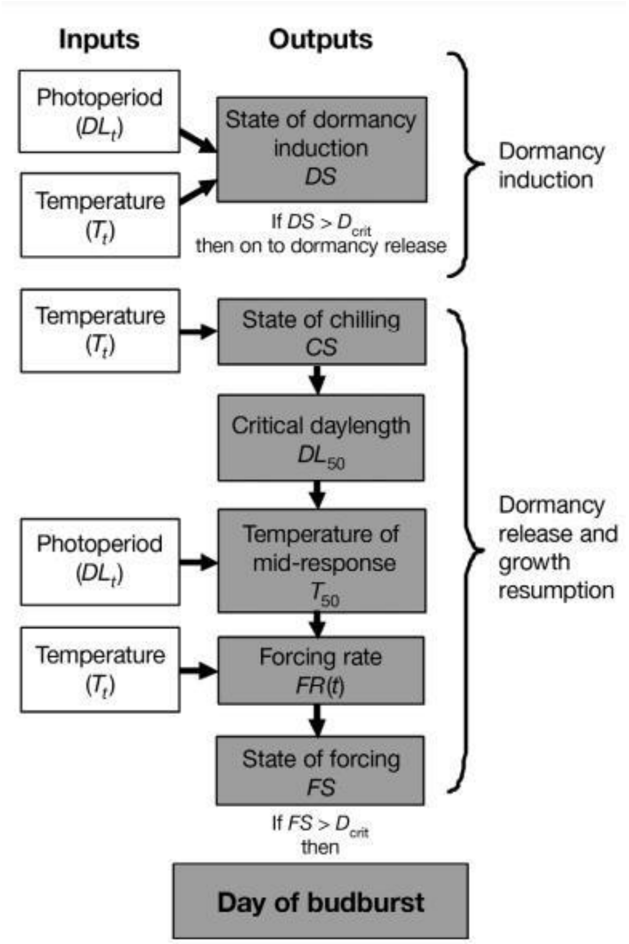


Figure 2. Conceptual schema of the DORMPHOT model (Caffarra et al. 2011)

Machine Learning in Phenological Modeling

The concept of ML is experiencing increasing prominence daily. The introduction of programs such as ChatGPT and AI developments of Microsoft signifies the rapid advancement in the field, pushing the boundaries of current possibilities. Mitchell (1997) was among the first to define what learning means for a computer program: "*A computer program is said to learn from* *experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E (p. 2)".*

To be able to comprehend and predict the complex relationships between biotic and abiotic systems in natural sciences, ML is superior to conventional statistical techniques (Dai et al. 2019). The ensemble approach – a strategy that integrates many models to increase a system's overall prediction performance and general robustness (Polikar, 2012) – is largely responsible for the superiority of ML (Dai et al. 2019). The theory behind this is that you may get better results by combining the predictions of several models using non-linear ML approaches. The study of Dai et al. (2023) used ensembling to improve phenological predictions of different ML models. ML can automatically extract rules from data and apply those rules to predict unknown data. The quality of science could be considerably increased through a broader use of ML techniques in ecology and earth science (Thessen, 2016).

In recent years, ML has become an important component of phenology prediction modeling, revolutionizing our ability to forecast leaf emergence events with higher accuracy. Dai et al. (2019) chose three widely used ML models to predict leaf emergence from phenology observations in Harbin, China. Their focus was on the RF, Neural Networks, and Gradient Boosting Decision Tree models. Wang et al. (2021) used the Boosted Regression Tree model to assess the contribution of land cover composition and topography on the spatial variation in land surface phenology throughout a bushfire in 2002 in New Mexico. Furthermore, Guo et al. (2021) conducted a study in which they used Multiple Linear Regression, Back-Propagation Neural Networks, SVM, and RF models to predict rice yields. Additionally, a very recent study by Dai et al. (2023) managed to compare 18 different ML models to test for the accuracy of leaf emergence prediction. To train and evaluate a model, however, long-term observations of a species are needed, which are unavailable for many species with only a few years or decades of observations (Van de Pol & Bailey, 2019). Moreover, overfitting is a common result when fitting a brief observation series with hundreds or even thousands of input variables (Janosi et al. 2020). Because ML is still relatively new in the discipline of phenological modeling, it is difficult to differentiate between "good" and "bad" methods. The abundance of ML models, coupled with the recent nature of numerous studies, prevents the emergence of any superior ML models. Consequently, the phenological prediction of many species with ML still faces significant obstacles.

Generally, in ML, two types of problems can be identified. These are explained by Hoffman et al. (2019). Firstly, in regression problems, the goal is to predict a continuous numeric value or quantity. The output is a real number rather than a category or label. In classification problems, the goal is to assign a label or category to a given input based on its features. The output is a discrete class or category. Predicting leaf emergence at the hand of ML methods can typically be considered a regression problem. A continuous numeric value – the timing of leaf emergence – is predicted and the output of the model will be a numeric representation that indicates the day of the year when the leaves are expected to emerge. However, through a method that is explained in the next section, it can also be made into a classification problem. Data is binary classified in this study because this results in an increased sample size and therefore is predicted to have an improved predictive performance (Dai et al., 2023).

Geospatial Data and Feature Extraction

Josefowicz et al. (2020) define geospatial data as: "*the digital version of location in threedimensional space (from below the earth to in space) of anything that can be assigned an identifier (e.g. people, vehicles, roads, geographical features, boundaries, pipes and wires, transmitters) and over time (the fourth dimension) (p. 27)*". Many information groups, such as forestry, marine studies, disaster management, natural resources, and health, can benefit from the use of geospatial data (Groot and McLaughlin, 2000). Geographical features are a part of geospatial data and generally distinctive physical or natural characteristics of the earth's surface (e.g. water bodies, vegetation, buildings, cities, infrastructure, political boundaries, etc.) (O'Brien & Cheshire, 2016).

Many phenology papers that are either comparative studies (Zhou et al. 2021; Dai et al. 2018; Kramer, 1994; Liu et al. 2017), phenological development studies (Caffarra et al. 2011; Chuine, 2000), or phenological model testing studies (Keenan and Richardson, 2015; Melaas et al. 2016) are studies that use phenological PB models that mostly require daily temperature or photoperiod as input. However, ML methods require certain features as input, which need to be extracted from geospatial datasets. Czernecki et al. (2018) extracted three types of features: spatial, meteorological, and remote sensing products, resulting in 19 different features that can be given as input for ML methods to predict leaf emergence. Specific features can be seen in Table 1. In studies by Dai et al. (2018) and Dai et al. (2023), the focus was on single monthly temperatures (beginning of May of the preceding year to April of the leaf emergence year) and daily mean temperature respectively. Capinha (2019) conducted a study on the emergence date of a certain species of mushroom in which they used a diverse range of preceding time windows to capture short-term detailed meteorological variations as well as more general variations in the long term. Within these time windows, they used the features mean of temperature and sums of precipitation.

Spatial	Meteorological	Remote sensing product
Altitude	Monthly mean air temp. (Jan-Dec)	Presence of snow cover (0-1)
Geographical coordinates	Monthly mean air temp. of the previous year December	Consecutive number of days with and without snow cover
Distance to coast	Seasonal mean air temp. of the previous year	Number of days with snow cover in a month
	Mean air temp. of winter and spring seasons	Day of the year with the last snow cover
	Total monthly precipitation (Jan-Dec)	Based only on the highest pixel reliability
	Total monthly precipitation of the previous year December	Based only on the highest and average pixel reliability
	Cumulative growing degree days	1-week rolling mean group by pixel reliability
	Cumulative growing precipitation days	Rate of change grouped by pixel reliability

Table 1. Different types of features according to Czernecki et al. (2018)

This Thesis aims to establish a workflow encompassing the extraction and selection of different features from meteorological datasets. Therefore, it becomes imperative to explain the processes behind the extraction of these features and why these are included in the workflow. A study by Dai et al. (2023) will form the foundation for the ML part of this research and is explained below. In this paper, features are extracted and selected in a way that aims to tackle the problem of the unavailability of data for many species, as explained by Van de Pol & Bailey (2019).

In the binary classification phenological model, for each ground observation record (DOY – Day Of Year), the researchers create two types of labels: positive labels (S) and negative labels (S'), and the sample size is constructed from the observations of each year. Any date before DOY is labeled as “0” (negative label), which indicates that the phenological event has not occurred in that specific year. Any date after the DOY is labeled as “1” (positive label), which indicates that the phenological event has occurred in that specific year. Then, a dynamic sliding time window is employed as a method to identify different weather variables that impact the phenology, with a length of N days before the corresponding date of the label. Within this time window, meteorological data is selected as features for each corresponding label. Consequently, the researchers choose the daily mean temperature as the meteorological variable for feature selection, but they note that other environmental drivers can be used too. The combination of these features and the corresponding label constitutes a sample (Figure 2).

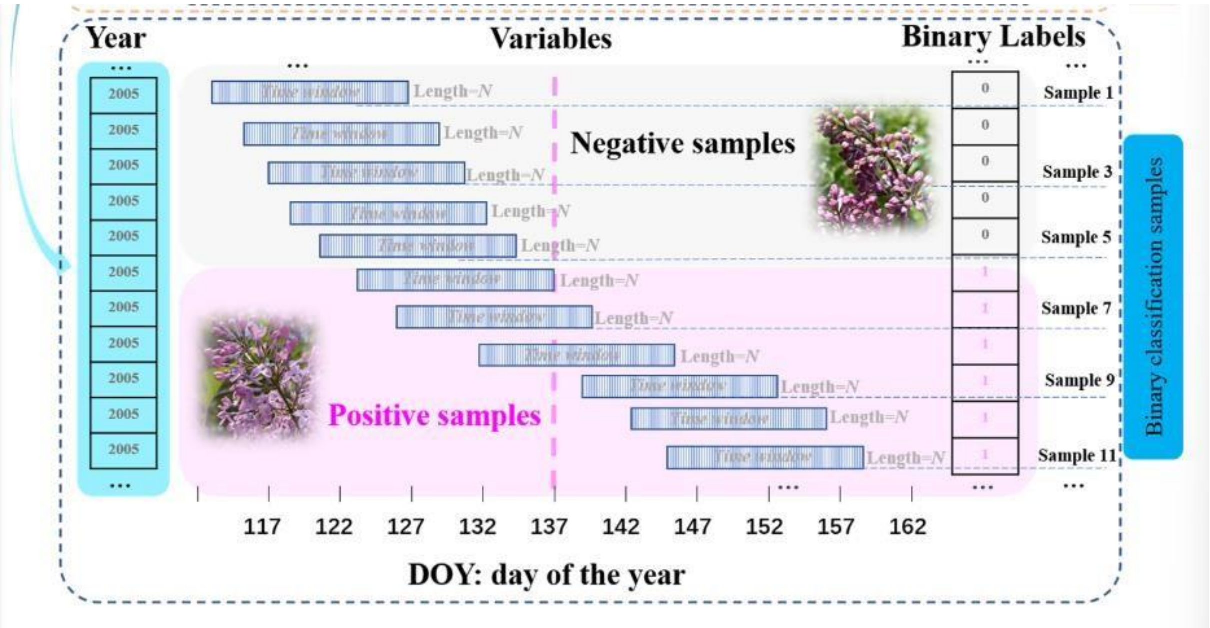


Figure 3. Sample composition of the binary classification phenological model (Dai et al. 2023)

Study materials

Study region

The study region encompasses a bounding box in the Eastern United States with the coordinates 35∘N−47∘N, 67∘W−90∘W (Figure 4). At these boreal and temperate latitudes, the climate is generally warming relatively rapidly (Caffarra et al. 2011). The wide range of leaf emergence dates at different latitudes within this box also creates significant variation in the dataset (Figure 6). The climate in this region is considered as humid cold temperate, humid warm continental, or humid warm temperate. Average temperatures in this box can range from -30 degrees to 30 degrees Celsius.

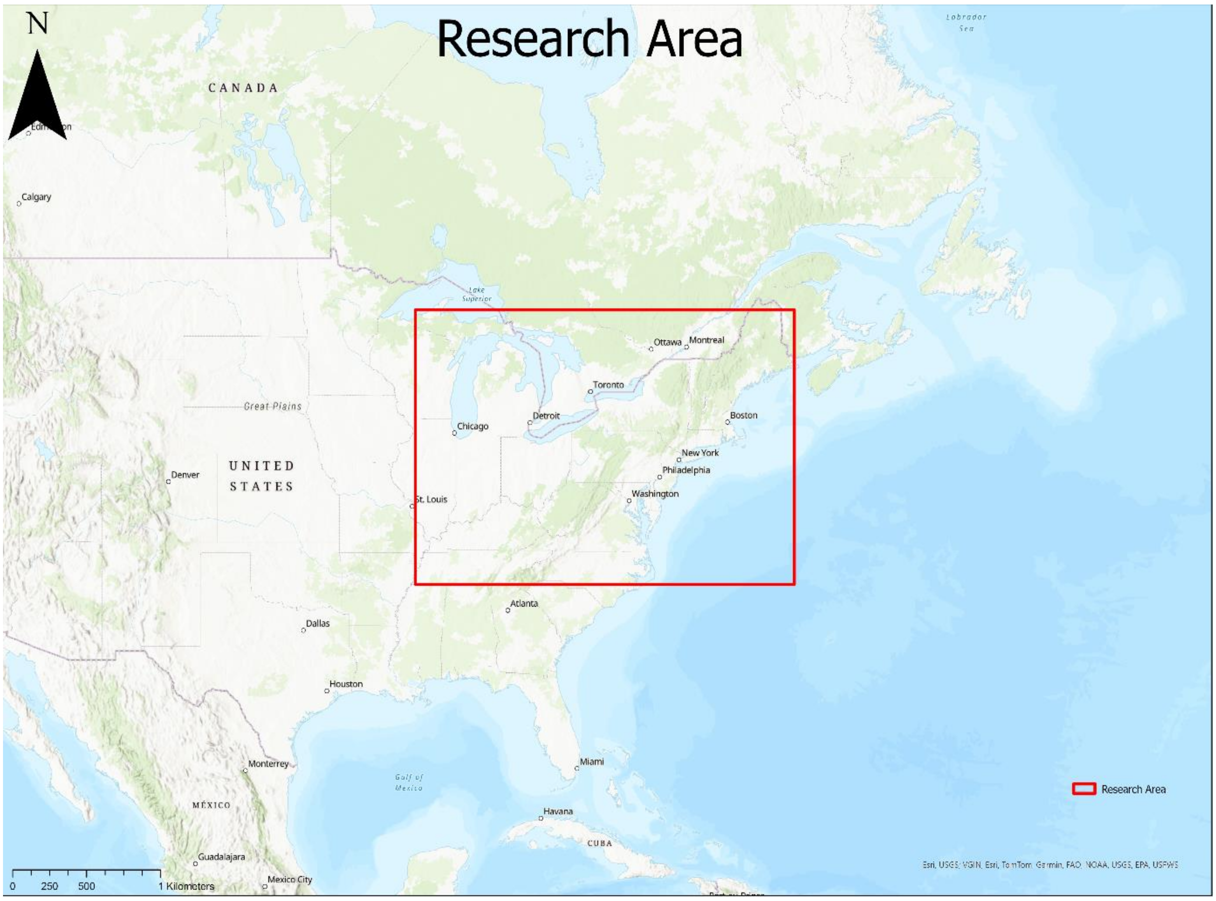


Figure 4. Research area

Ground observations

Phenological observations of leaf emergence dates are taken from the United States of America National Phenology Network (USA-NPN). The goal of this network is to, as stated by Rosemartin et al. (2018), advance science, inform decisions, and communicate and connect with people about phenology and responses to environmental change and climate variation. Observation phenology data is collected and submitted by scientists (professional and citizen). These observations include information on the timing of various phenological events, including leaf emergence, for many different species. To extract the annual spring leaf emergence dates for individual observations in this study, the first step involved identifying all available USA-NPN observations gathered within the previously stated research area. This allows for the determination of the first observation of each year. These observations correspond to the phenophase attribute recorded as ‘Breaking leaf buds’, indicating leaf emergence. This study is conducted on all deciduous broadleaf observations within the study region. Deciduous broadleaf is often chosen as a species to be researched within phenology prediction papers (Dai et al. 2019; Czernicki et al. 2018; Dai et al. 2023; Delpierre et al. 2009) due to its (i) distinct phenological phases and (ii) responsiveness to environmental drivers. The final dataset includes 4003 observations on leaf emergence for deciduous species collected from 2016-2022 within the bounding box. Furthermore, research by Gerst et al. (2015) showed that at larger scales, where sample sizes are big, the quality of the NPN dataset is adequate to support statistical modeling.

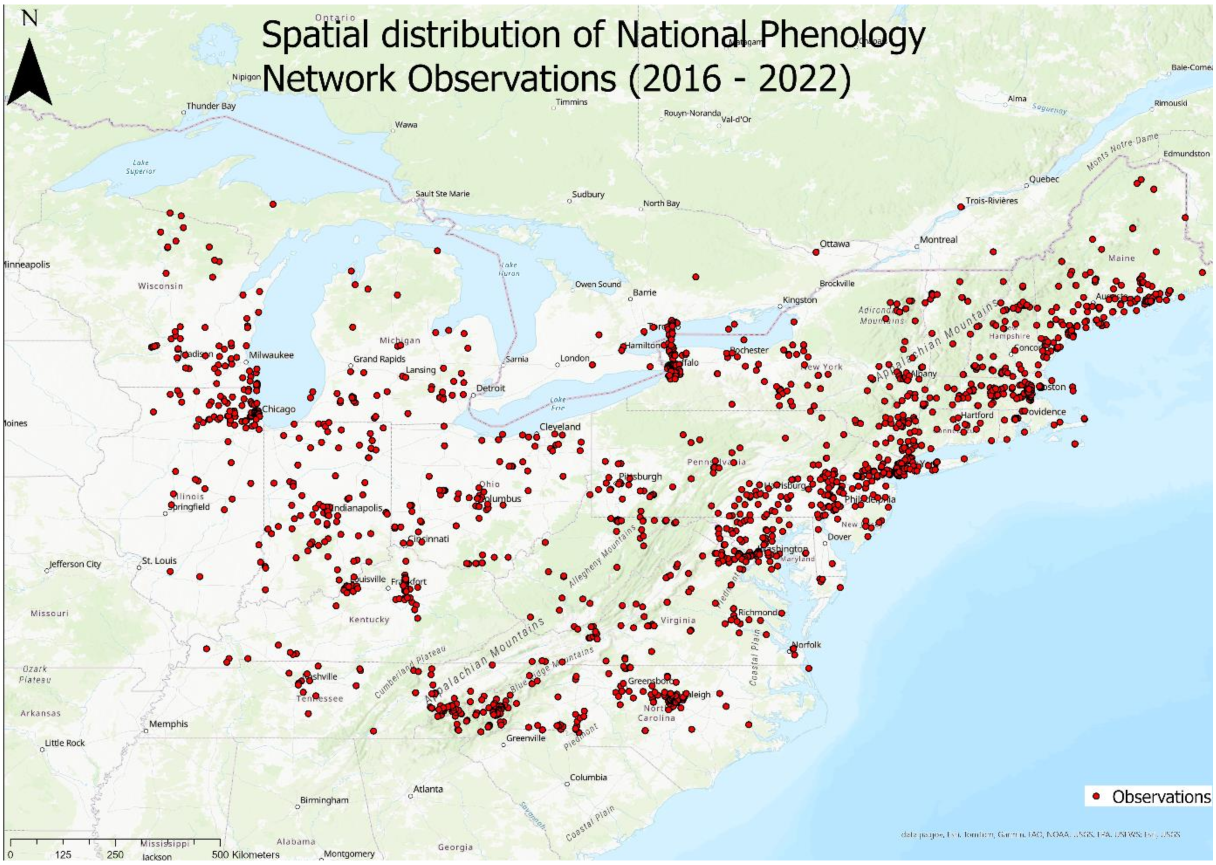


Figure 5. Spatial distribution of observations

Figure 5 shows how the observations are distributed throughout the research area. It can be seen that most observations have been taken by the northern part of the coast, but overall, predictions are well spread out in the bounding box. Figure 6 shows boxplots of leaf emergence per year. There is not a determinable trend in the boxplots, although earlier leaf emergence seems to be followed by a year of later leaf emergence. There is however an extreme amount of outliers, which could cause skewed performances by the models, which is why these are removed, as explained in the methodology. Outliers below day 100 and over day 175 are removed from the dataset. After removing the outliers a dataset with 2664 observations is left. Figure 7 shows the spatial distribution of the outliers, for which it can be stated that these are evenly spatially distributed over the research area. Figure 8 shows the boxplots of leaf emergence dates after outliers are removed. It can be seen that a year of later leaf emergence is often followed by a year of earlier leaf emergence. Lastly, Figure 9 shows how observations are distributed over the years.

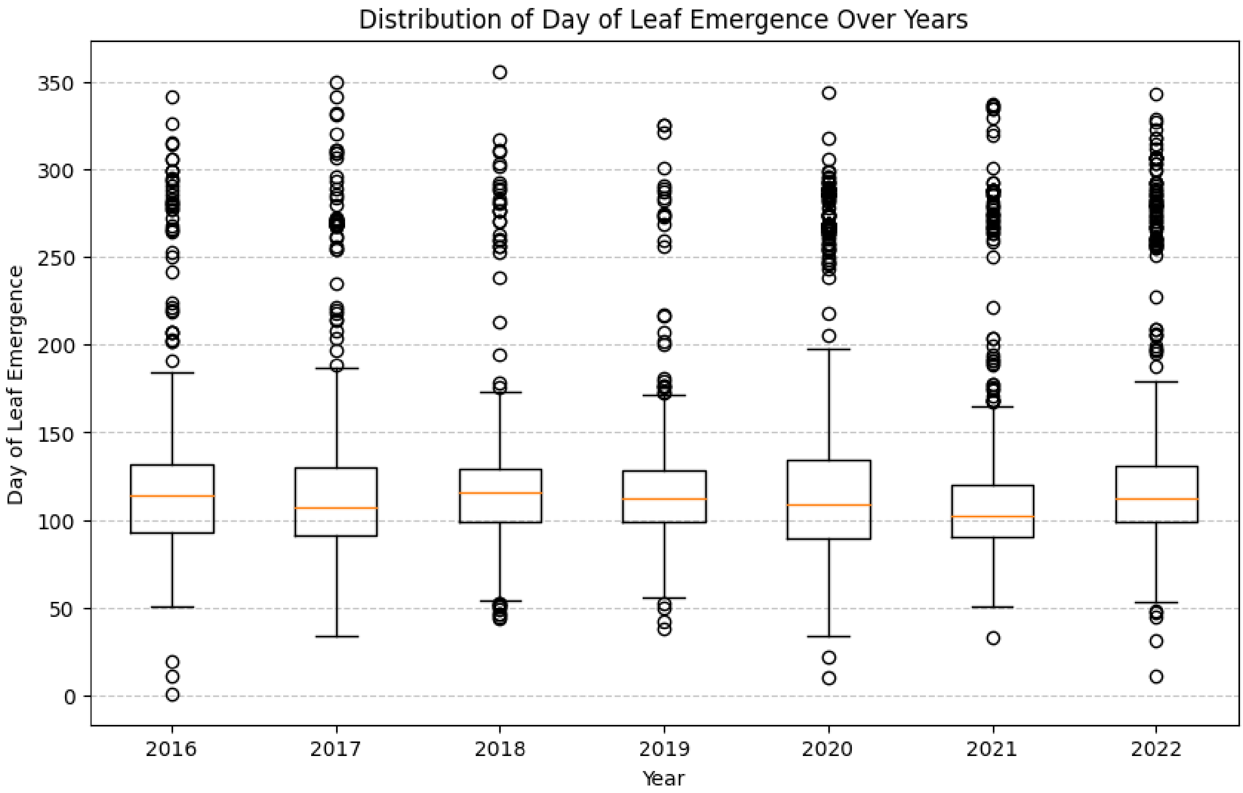


Figure 6. Boxplot of distribution of day of leaf emergence per year (DOY)

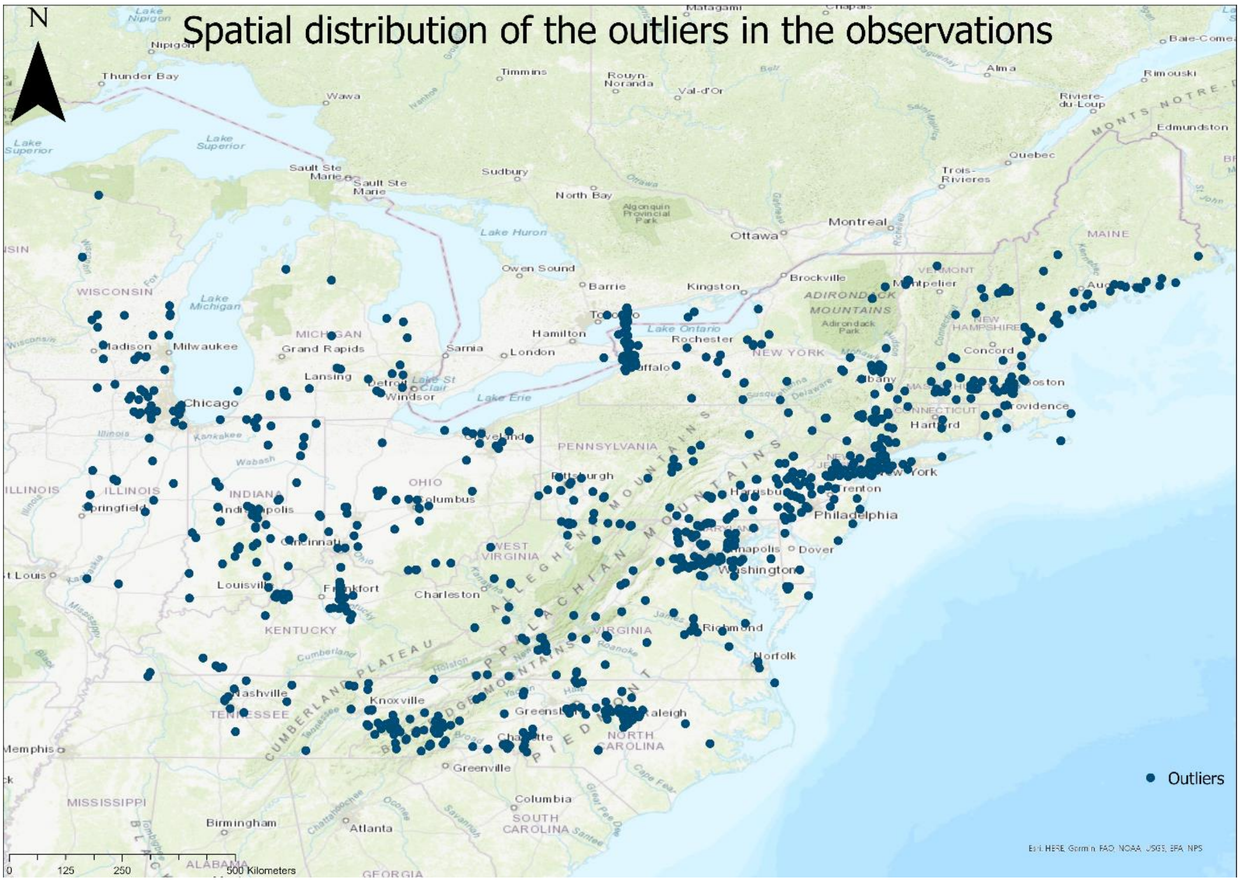


Figure 7. Spatial distribution of outliers in the data

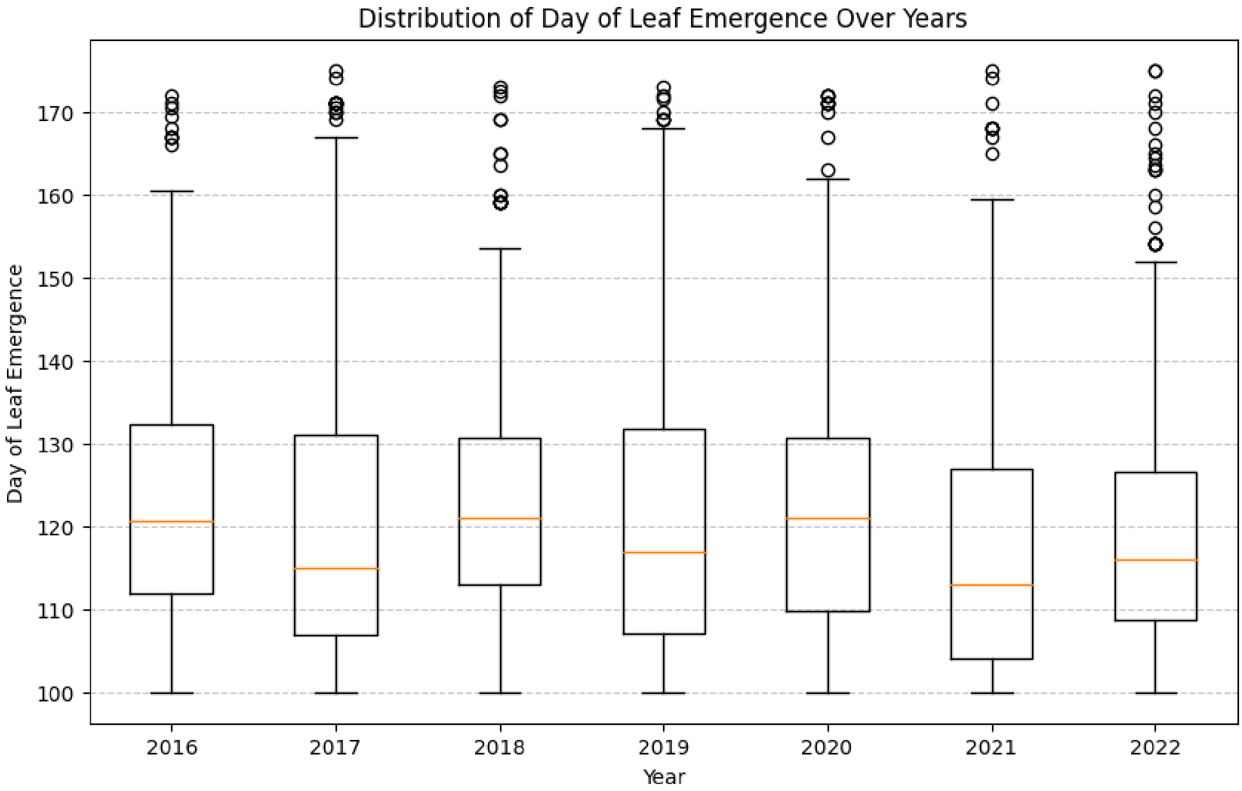


Figure 8. Boxplots of distribution of day of leaf emergence (after outliers are removed)

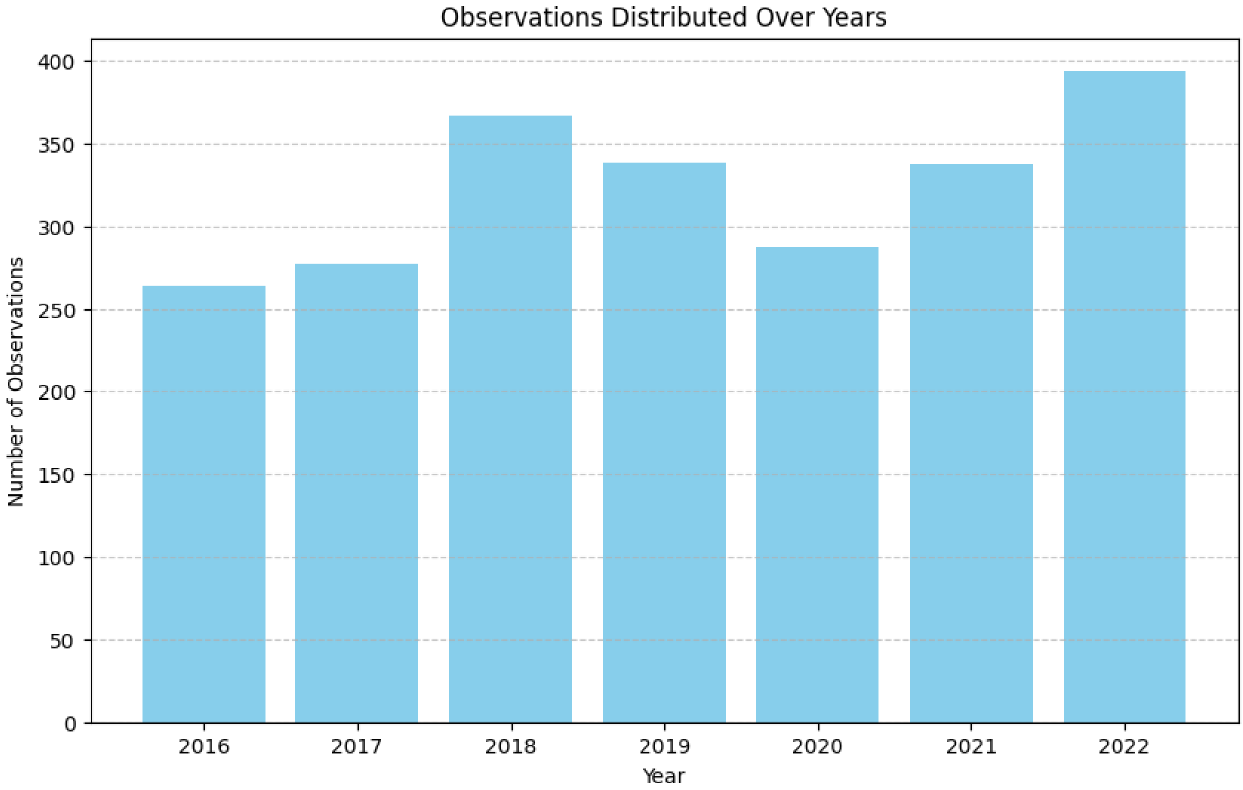


Figure 9. Distribution of observations per year

Meteorological data

For this study, weather data is used as a form of open-access geospatial datasets, but Springtime allows for the integration of other datasets into the workflow too. Meteorological data is downloaded from the Daymet dataset (through Springtime), which stands for "Daily Surface Weather Data for North America". The dataset includes various meteorological climate variables and environmental drivers, such as temperature, precipitation, humidity, and radiation, at a spatial resolution of 1km x 1km. Daymet is widely used in phenology research (Izquierdo-Verdiguier et al. 2018; St. Peter, 2015; Mehdipoor et al. 2019; Melaas et al. 2016) and is often used to analyze climate trends, assess the impact of the climate on various sectors, and make informed decisions in the area of natural research management. Daymet is derived from multiple sources, including but not limited to meteorological station data, satellite data, and weather models (Thornton et al. 2018). It is freely available to the public and often accessed through online platforms, or in this case downloaded through a script written in Springtime. This allows for the development of subsets in the form of single points, multiple points, or a bounding box. In this research, meteorological data in the form of average daily temperature is taken from 2016 to 2022 in the research area. Figure 10 shows the average temperature distribution per year. The average daily temperature is very similar throughout the research years, but there is a significant amount of outliers in the lower part of the graph.

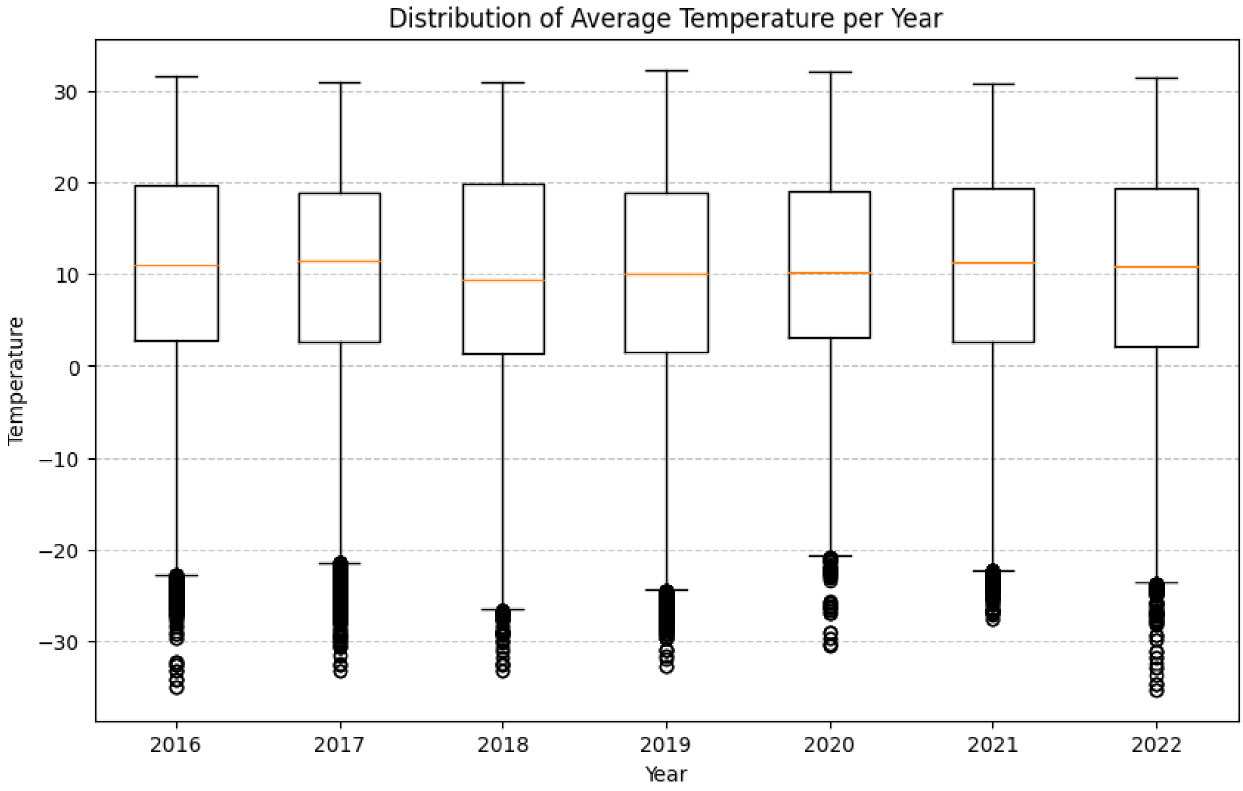


Figure 10. Average temperature distribution by year

Process-based models

The models explained in the theoretical framework are the same models that will be tested in Springtime, except for the DORMPHOT model. This model is used as an example to show the trend in process-based models for phenological modeling, but will not be tested because it is not available within the pyPhenology package. As these are already extensively defined, they are only briefly explained again. The TT model performed well in the research of Linkosalo et al. (2008) and is generally the most usable model in different situations. The S model also performed well in the research of Linkosalo et al. (2008). Additionally, this model includes chilling and is most representative of other models that use chilling, such as the alternating or parallel model (Melaas et al. 2013). Lastly, the UC model might be the best fit due to the high number of parameters. The UC model is essentially the same as the Unified model, but slightly easier to use. It also is the most comprehensive model available in the pyPhenology package, which is incorporated in Springtime. All these models are temperature-dependent, as this remains the most important driver of leaf emergence in deciduous trees (Dai et al. 2019).

Machine learning models

Gradient Boosted Decision Tree Model

The first ML model that will be used is the GB model, which is an ML ensemble technique that combines the predictive power of multiple decision trees. It iteratively improves the performance of a weak learner by focusing on the mistakes made in the previous iterations (Ke et al. 2017). It adds trees to the model, where each newer tree corrects the errors made by the previous ones. These models are highly effective for both types of ML problems: regression and classification, providing overall robust performance and handling complex data patterns well. It is usually a popular choice for various applications due to its ability to handle missing data and feature importance. Seeing as data is binary classified and features are an input of the model, this model is chosen to predict leaf emergence for this study. In regards to phenology, this method is the best-performing model according to comparative studies by Dai et al. (2018) and Czernicki et al. (2018).

Random Forest Model

The RF model is a technique that constructs multiple decision trees during the training of the model and merges these to get a more accurate and stable prediction. The difference with the GB model is that RF builds trees independently and uses a process called bootstrap aggregating, where each tree is trained on a random subset of the data with a replacement (Pellagatti et al., 2021). The final prediction of this model is often made by either averaging the predictions (in regression) or taking a majority vote (in classification) of all the developed individual trees. These types of models are mostly good at handling larger datasets with relatively high dimensionality and they are generally good at mitigating overfitting, making them applicable for a wide variety of tasks. They are often used for classification problems and proven as a reliable option for phenological modeling (Dai et al., 2018).

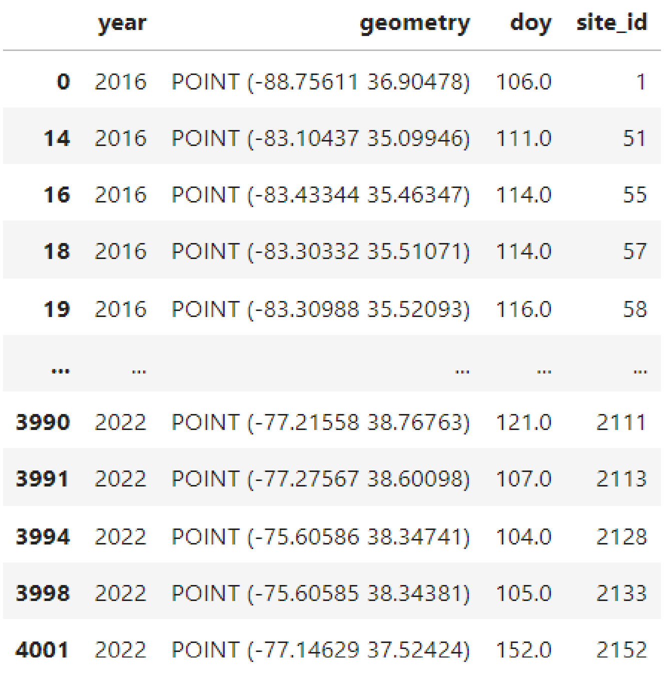
Support Vector Machine Model

The SVM model is a learning algorithm primarily used for classification tasks (Chandra & Bedi, 2021), so suitable for this study. It works by finding the hyperplane that best separates the data into different classes. It does this by maximizing the margin, which is the distance between the hyperplane and the nearest data points from every possible class, known as support vectors (Chandra & Bedi, 2021). With the use of kernel functions, SVM models can handle non-linear classification problems, making it a highly suitable model for this research.

Methodology

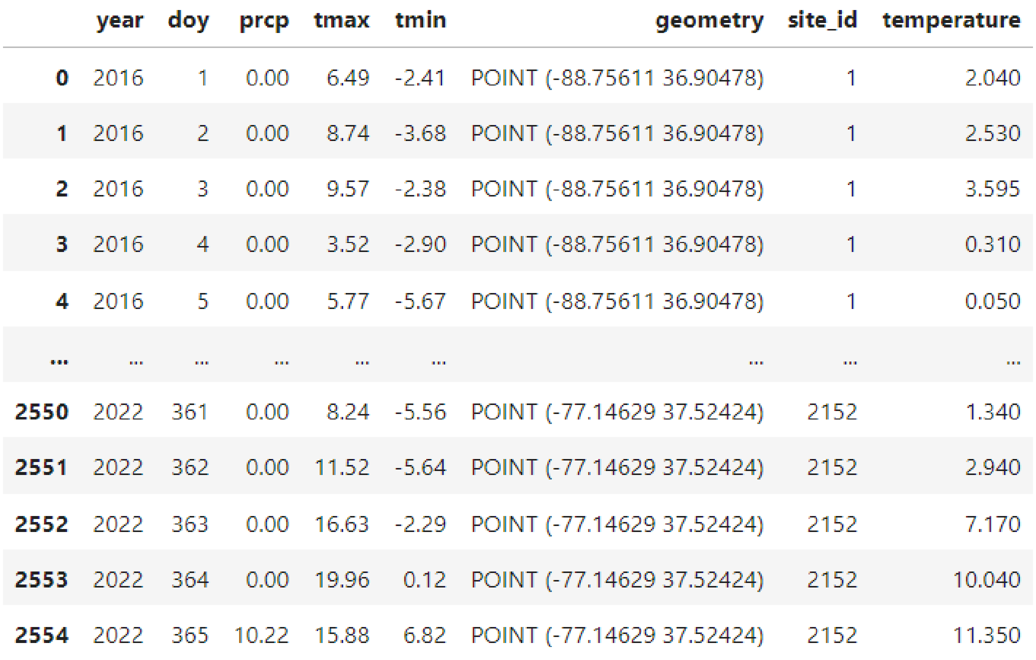
Data pre-processing

The first step in creating the workflow is deciding what phenophases and species are to be selected for the case study. ‘Breaking leaf buds’ and ‘deciduous broadleaf’ are selected. Next, a subset is created of data for the years 2016 to 2022, focusing on the selected species and phenophase within the specified research area (generally the Eastern US). The data is downloaded and loaded into a data frame. To ensure spatial accuracy, points outside of the coverage area of Daymet data are filtered out. Then, a new column of ‘site id’ is added to identify each site based on its geometry (data needs to be structured as such for the pyPhenology package). The resulting dataset consists of the columns ‘year’, ‘geometry’, ‘DOY’, and ‘site id’ (Figure 11). Lastly, outliers are removed according to the box plot of Figure 7. Observations with leaf emergence before day 100 and after day 175 are removed from the dataset. Unique coordinates are then extracted from the geometry field, which is necessary for the downloading of Daymet data. Meteorological data needs to be downloaded for every day of the year from 2016 to 2022, so Daymet data is downloaded in batches. Climate data is loaded into a DataFrame and unnecessary columns are dropped. Consequently, the ‘site id’ column is also added, next to a mean temperature column by adding the maximum temperature and minimum temperature and dividing by 2. This finalized data frame consists of the columns ‘year’, ‘day of the year’, ‘precipitation’, ‘geometry’, ‘site id’, and ‘mean temperature’ (Figure 12). As a result, there are now two DataFrames, one with observations and one with corresponding meteorological data.



	year	geometry	doy	site_id
0	2016	POINT (-88.75611 36.90478)	106.0	1
14	2016	POINT (-83.10437 35.09946)	111.0	51
16	2016	POINT (-83.43344 35.46347)	114.0	55
18	2016	POINT (-83.30332 35.51071)	114.0	57
19	2016	POINT (-83.30988 35.52093)	116.0	58
...	...	...	...	...
3990	2022	POINT (-77.21558 38.76763)	121.0	2111
3991	2022	POINT (-77.27567 38.60098)	107.0	2113
3994	2022	POINT (-75.60586 38.34741)	104.0	2128
3998	2022	POINT (-75.60585 38.34381)	105.0	2133
4001	2022	POINT (-77.14629 37.52424)	152.0	2152

Figure 11. Observational dataset



	year	doy	prcp	tmax	tmin	geometry	site_id	temperature
0	2016	1	0.00	6.49	-2.41	POINT (-88.75611 36.90478)	1	2.040
1	2016	2	0.00	8.74	-3.68	POINT (-88.75611 36.90478)	1	2.530
2	2016	3	0.00	9.57	-2.38	POINT (-88.75611 36.90478)	1	3.595
3	2016	4	0.00	3.52	-2.90	POINT (-88.75611 36.90478)	1	0.310
4	2016	5	0.00	5.77	-5.67	POINT (-88.75611 36.90478)	1	0.050
...	...	...	...	...	...	...	...	...
2550	2022	361	0.00	8.24	-5.56	POINT (-77.14629 37.52424)	2152	1.340
2551	2022	362	0.00	11.52	-5.64	POINT (-77.14629 37.52424)	2152	2.940
2552	2022	363	0.00	16.63	-2.29	POINT (-77.14629 37.52424)	2152	7.170
2553	2022	364	0.00	19.96	0.12	POINT (-77.14629 37.52424)	2152	10.040
2554	2022	365	10.22	15.88	6.82	POINT (-77.14629 37.52424)	2152	11.350

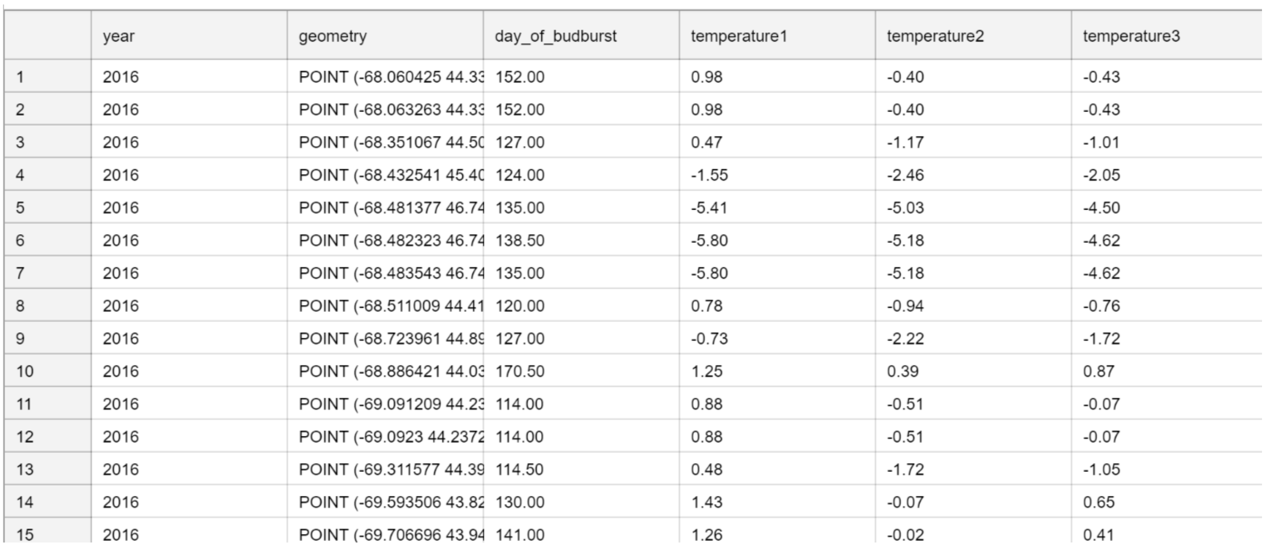
Figure 12. Meteorological dataset

Data splitting

For all observations, the dataset was then split into a training dataset (observations from 2016-2020, about 70% of the original dataset) and a test set (observations from 2021 and 2022, the remaining 30% of the original dataset, functioning as a stand-in for unseen data). Splitting data in this manner will provide a framework for the evaluation of both the PB models and ML models. Additionally, many studies show that using between 70% and 80% of data for training will allow the models to capture underlying patterns and relationships in the data (Q. Wei et al., 2013; Y.W.Chang et al., 2010; M.K. Ucar et al., 2020). Moreover, by assessing their performance on the same testing data, you can discern which approach better captures the impact of meteorological features on plant leaf emergence date. These four datasets are now ready to be used in the PB models.

Creating recipes

To prepare the datasets for ML analysis, the four created datasets are loaded in. These are the training and testing datasets for both the observations and the meteorological data. They are merged based on the common columns year and geometry to create two datasets, one for training and one for testing. The merging process ensures that each phenological observation is aligned with the corresponding weather data based on year and location. Following this merging process a detailed training dataset and testing dataset, which includes temperature data for each day of the year, is created, which is the recipe. This is necessary so that features can be extracted more easily. This involves iterating through the merged data and grouping the observations by year and geometry. For each group, a new record is created that captures the year, geometry, and DOY. Subsequently, for each day of the year, the corresponding temperature value is extracted from the group. Figure 13 shows an example of such a recipe. The recipe includes the observational data and temperature data for that whole year.



	year	geometry	day_of_budburst	temperature1	temperature2	temperature3
1	2016	POINT (-68.060425 44.33	152.00	0.98	-0.40	-0.43
2	2016	POINT (-68.063263 44.33	152.00	0.98	-0.40	-0.43
3	2016	POINT (-68.351067 44.50	127.00	0.47	-1.17	-1.01
4	2016	POINT (-68.432541 45.40	124.00	-1.55	-2.46	-2.05
5	2016	POINT (-68.481377 46.74	135.00	-5.41	-5.03	-4.50
6	2016	POINT (-68.482323 46.74	138.50	-5.80	-5.18	-4.62
7	2016	POINT (-68.483543 46.74	135.00	-5.80	-5.18	-4.62
8	2016	POINT (-68.511009 44.41	120.00	0.78	-0.94	-0.76
9	2016	POINT (-68.723961 44.89	127.00	-0.73	-2.22	-1.72
10	2016	POINT (-68.886421 44.03	170.50	1.25	0.39	0.87
11	2016	POINT (-69.091209 44.23	114.00	0.88	-0.51	-0.07
12	2016	POINT (-69.0923 44.2372	114.00	0.88	-0.51	-0.07
13	2016	POINT (-69.311577 44.39	114.50	0.48	-1.72	-1.05
14	2016	POINT (-69.593506 43.82	130.00	1.43	-0.07	0.65
15	2016	POINT (-69.706696 43.94	141.00	1.26	-0.02	0.41

Figure 13. Recipe of training data

Feature extraction and binary classification

For the feature extraction and binary classification process, the dataset is first read from the corresponding created recipe. From this dataset, the relevant columns of 'year', 'DOY', and the daily temperatures for a year (per observation) are selected. Lists to store features and labels are initialized, and two parameters are defined: S (sample size) and N (size of sliding time window). These are set to 100 and 50 respectively as done by the research of Dai et al. (2023). For each observation in the dataset, the DOY is extracted. Then, the temperature data for the year is retrieved, and the start indices for the sliding time window are calculated. Within this window, for each day, the corresponding temperature data is extracted. Days before DOY are labeled as 0 (negative sample), while days on or after the DOY are labeled as 1 (positive sample). These processed windows, along with their corresponding labels, are then appended to the features and label lists. Finally, the lists of features and labels are converted to numpy arrays for further processing and input for the ML models. This methodology ensures that features and labels are extracted and processed for binary classification based on temperature patterns around the DOY. Temperature is the only feature that will be used as an input feature of the ML models in this study. The PB models also only consider temperature, so this decision ensures justifiable comparisons between the different types of models. This method results in approximately 153.000 features for training and 73.000 features for testing.

Model comparison

Previous research has shown that the most effective metric for comparing various phenological models is the root mean square error (RMSE), which is computed from the observed records and the matching phenological predictions (Chuine et al., 1998). RMSE is thus used as the assessment metric for parameter optimization of PB models and its comparison with binary classification models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(obsi - prei)^2}{n}}$$

Where pre$_i$ and obs$_i$ are the model-predicted and observed leaf emergence dates (DOY) of year *i*, and *n* is the number of years recorded.

Additionally however, to possibly show difference between days better and give complementary information to the RMSE-values, the Mean Absolute Error is also calculated for each prediction. This will also act as somewhat of a control value to the RSME. This value is calculated by dividing the sum of the absolute errors by the sample size.

For the ML models predictions are given in an array of 0s and 1s and compared to the labels of the binary-classified observational data. However, for these binary arrays, no RMSE value can be calculated. Therefore, the predicted array needs to be recalculated as an array of actual predicted leaf emergence dates. To do this, first, a window size of 100 is set, which defines the size of the segments the array will be divided into. Then, a loop iterates over the predictions array in increments of the window size. In every window, the position is determined where the 0 switches to a 1 (which is once every window, resulting in the same amount of prediction days), and written to a new array. Lastly, the predicted days are calculated by adding the new array to the actual leaf emergence dates and subtracting 50. This last step is being done because the data was binary classified and 0 switches to a 1 in position 50 for all of the labels that are used for testing. Essentially, by doing this, the day that the label 1 appears first is considered as day of leaf emergence. Figure 14 shows the flowchart of the full methodology.

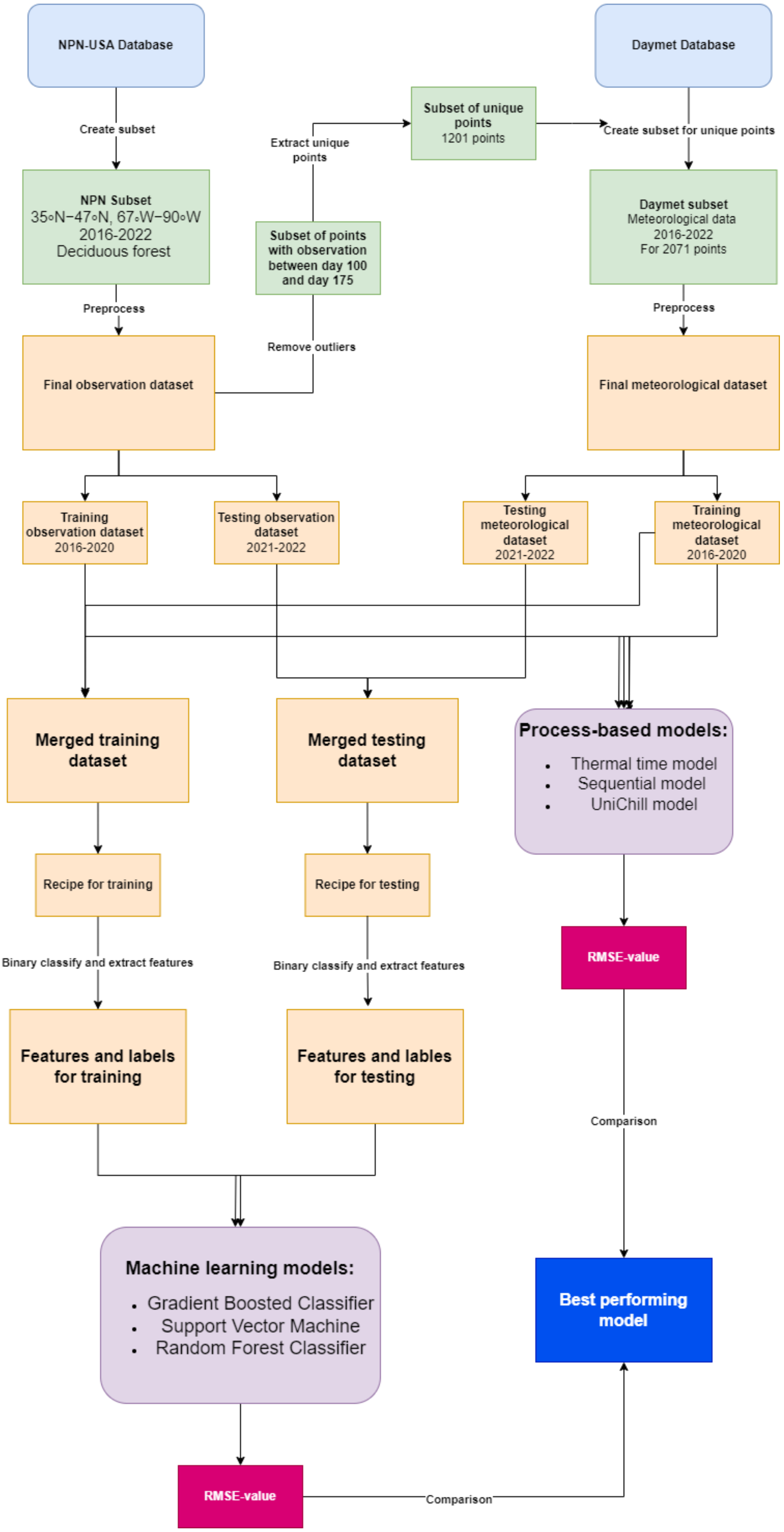


Figure 14. Flowchart of the methodology

Results

Prediction accuracy of process-based models

First, the training and testing datasets are fed into the three different process models and for every model, the parameters are set by using the training data. Then, using the testing data the root mean squared error (RMSE) between the predicted leaf emergence dates and actual leaf emergence dates is calculated. The test RMSE of the TT model is 12.957, the S model 13.307, and the UC model 13.019. This indicates that all models performed relatively evenly, where the TT model performed slightly better and the S model slightly worse. Scatterplots of the actual DOY vs. the predicted DOY for the PB models can be seen in Figures 15, 16, and 17. These scatterplots show a high concentration of leaf emergence dates in the earlier part of the year. Furthermore, in the earlier part of the year, most dates are overpredicted, and later in the year underpredicted with larger outliers. For the TT model and the UC model, in the earlier part of the year, the predicted dates are concentrated between day 110 and day 120. The predicted dates for the S model in this part of the year are more spread out. However, the S model has, consequently, higher outliers, especially in the later part of the year.

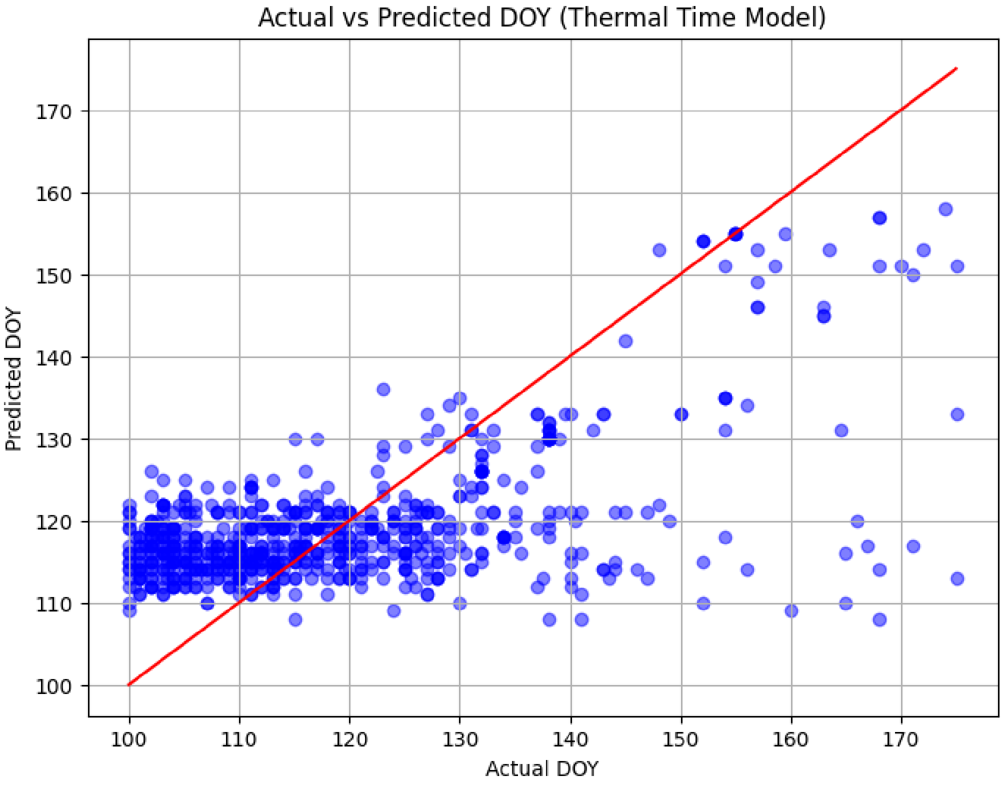


Figure 15: Scatterplot of the predicted DOY vs. the actual DOY for the TT model

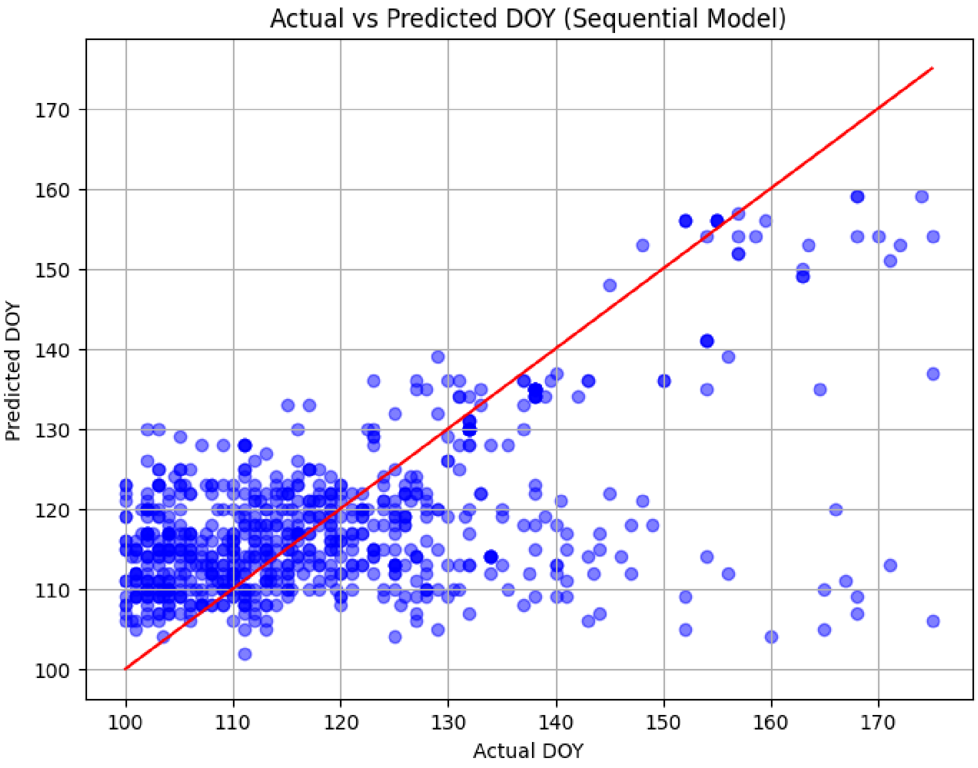


Figure 16: Scatterplot of the predicted DOY vs. the actual DOY for the S model

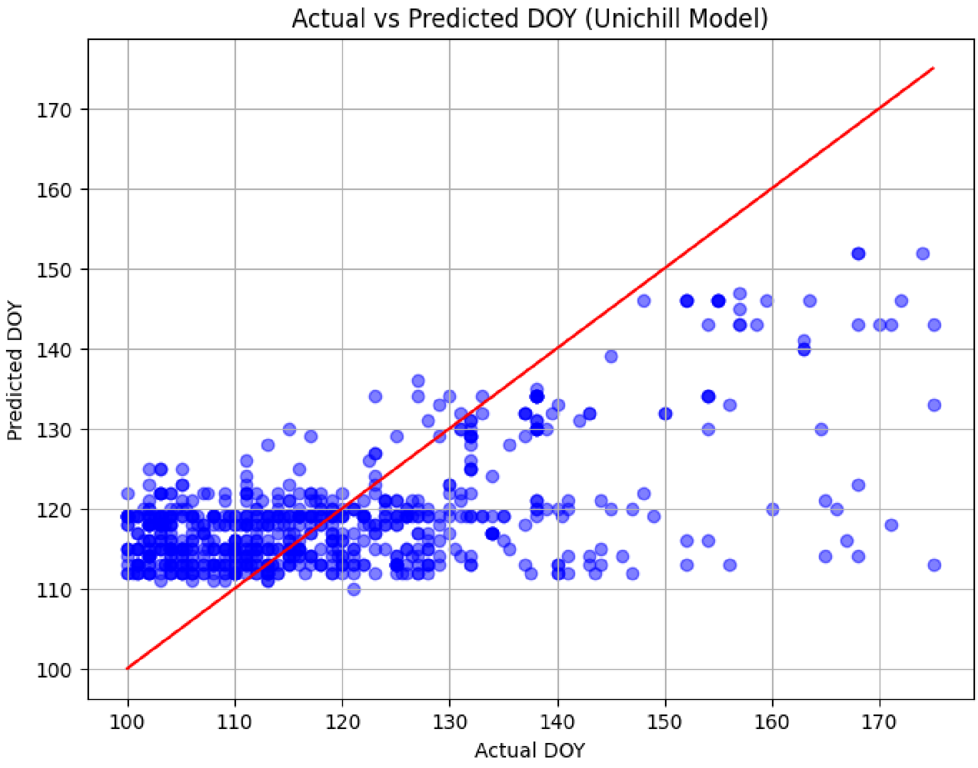


Figure 17: Scatterplot of the predicted DOY vs. the actual DOY for the UC model

Prediction accuracy of machine learning models

The binary classified datasets are fed into the ML models and accuracy is calculated based on the predicted labels vs. the actual labels. The GB model has a test accuracy of 0.882, the RF classifier model 0.883, and the SVM 0.879. Therefore, it can be stated that the GB model performs slightly better and the SVM slightly worse. However, margins are small, so conclusions based on these results are again difficult. Figures 18, 19, and 20 show the actual vs. predicted labels per model.

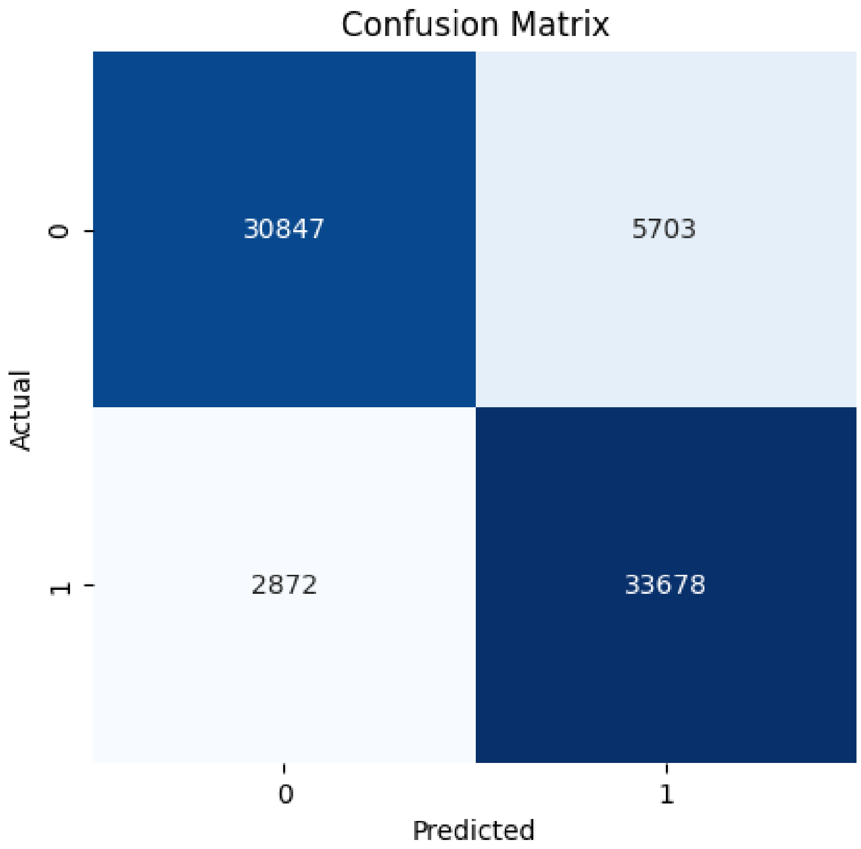


Figure 18: Confusion matrix of GB model

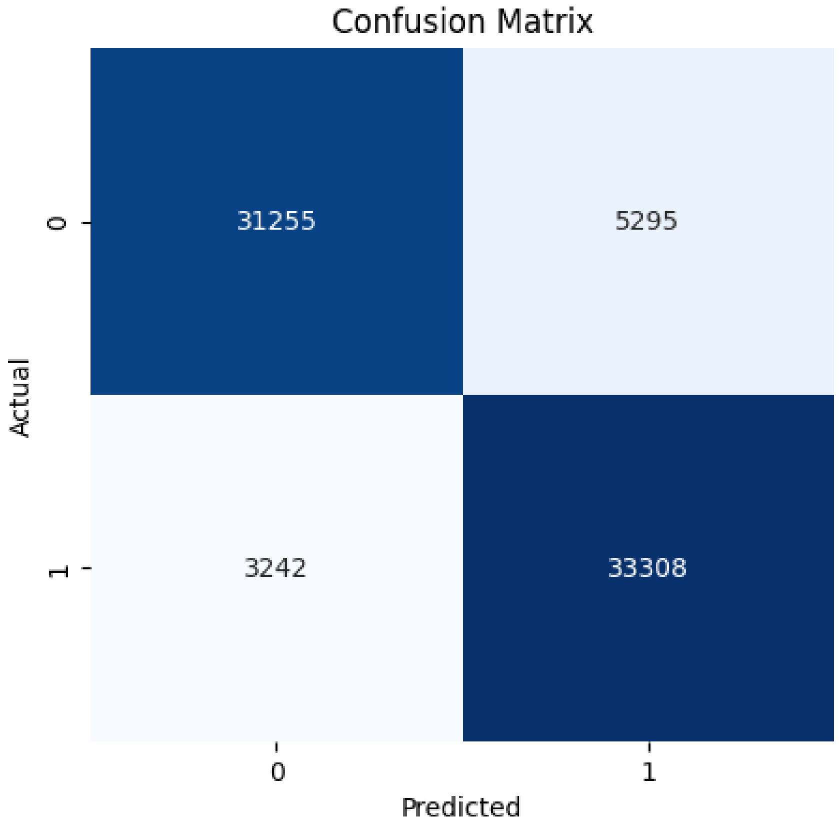


Figure 19: Confusion matrix of the RF model

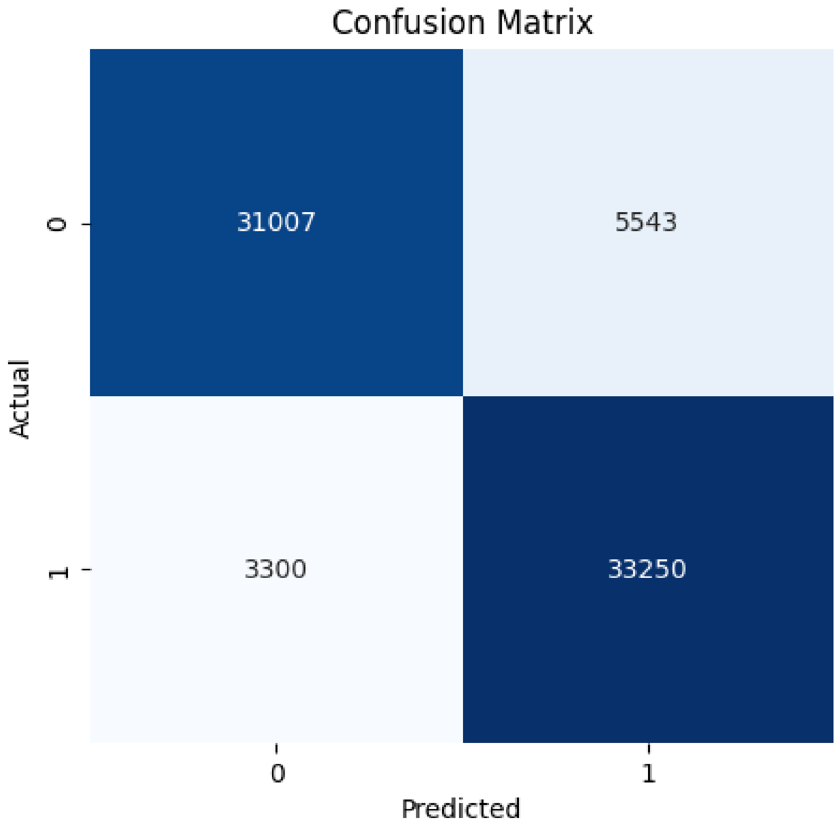


Figure 20: Confusion matrix of the SVM model

As mentioned before, the outcomes from the ML models can be recalculated and expressed as RMSE. For the GB this is 16.024, for the RF model 16.130 and the SVM model 15.888 days. When taking the RSME into account, the SVM model performs slightly better and the RF model slightly worse. Scatterplots of the actual DOY vs. the predicted DOY for the ML models can be seen in Figures 21, 22, and 23. The scatterplots are scattered around the line well, which is for a ML model. Visually, it is difficult to determine significant differences between the models, although it might be stated that the SVM model has slightly more extreme outliers in the later part of the year.

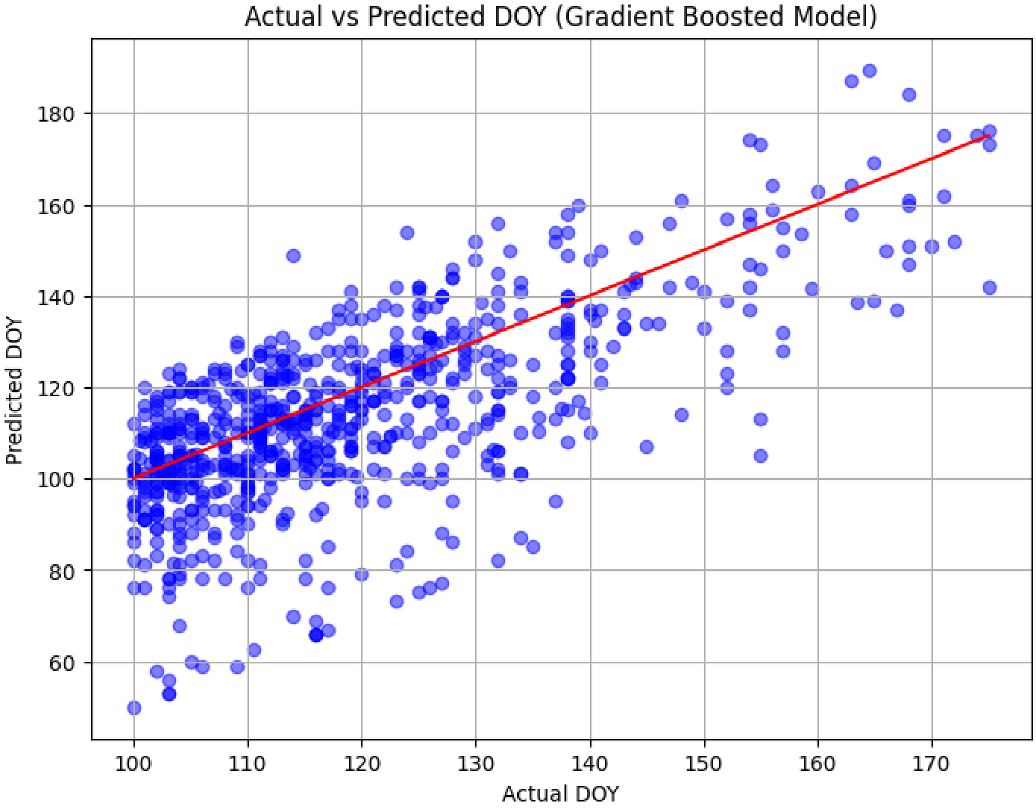


Figure 21: Scatterplot of the predicted DOY vs. the actual DOY for the GB model

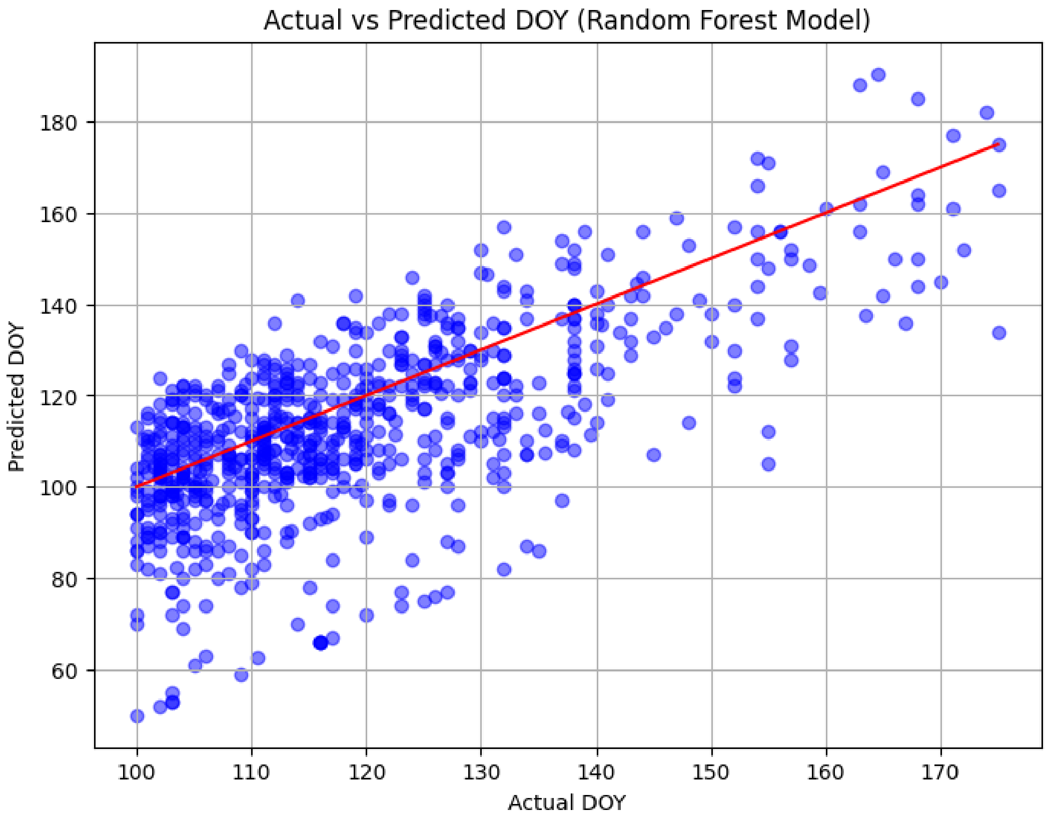


Figure 22: Scatterplot of the predicted DOY vs. the actual DOY for the RF model

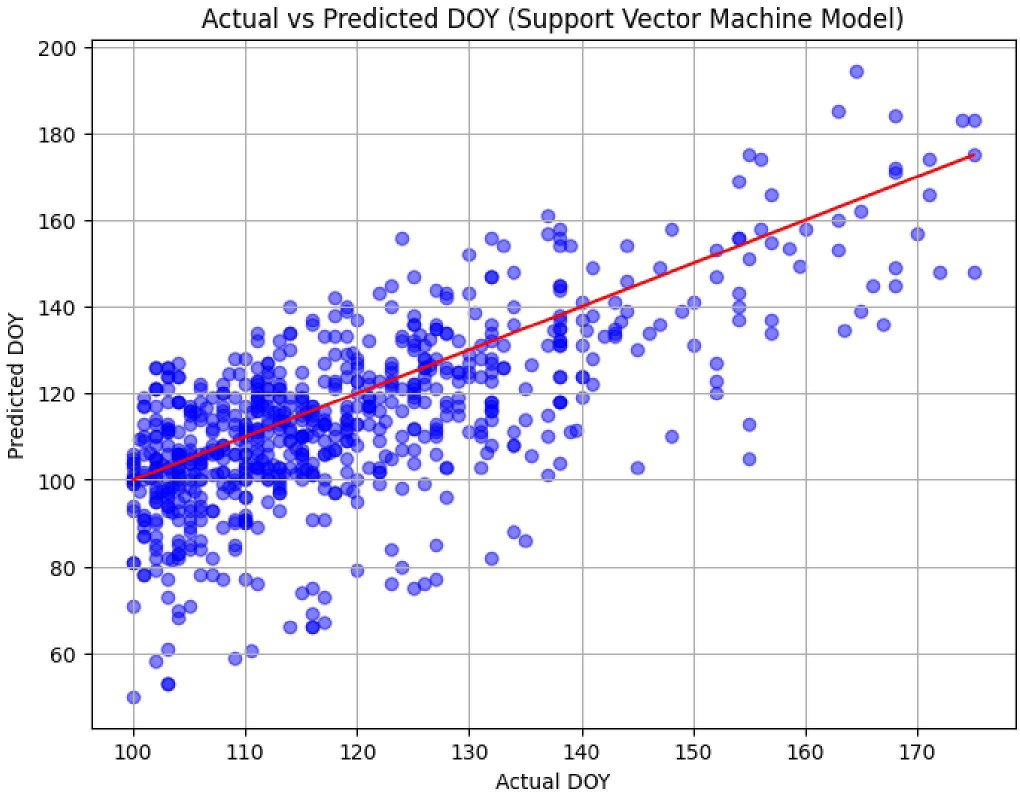


Figure 23: Scatterplot of the predicted DOY vs. the actual DOY for the SVM model

Comparison between process-based models and machine learning models

The ML model predictions are translated into RMSE scores according to the method explained in the methodology. Figure 24 shows a bar plot of the RMSE scores for all six models. For the TT model, the S model, the UC model, the GB model, the RF model, and the SVM model these values translate to 12.957, 13.307, 13.019, 16.024, 16.130, and 15.888 days respectively. As can be seen, the PB models all perform relatively similarly, as do the ML models, with the TT model outperforming all models. Additionally, the Mean Absolute Error (MAE) is calculated to show the difference between days better. These translate to 9.845, 9.515, 10.082, 11.832, 12.193, and 11.847 respectively (Figure 25). Within this metric, the S model performs best and the RF model worst.

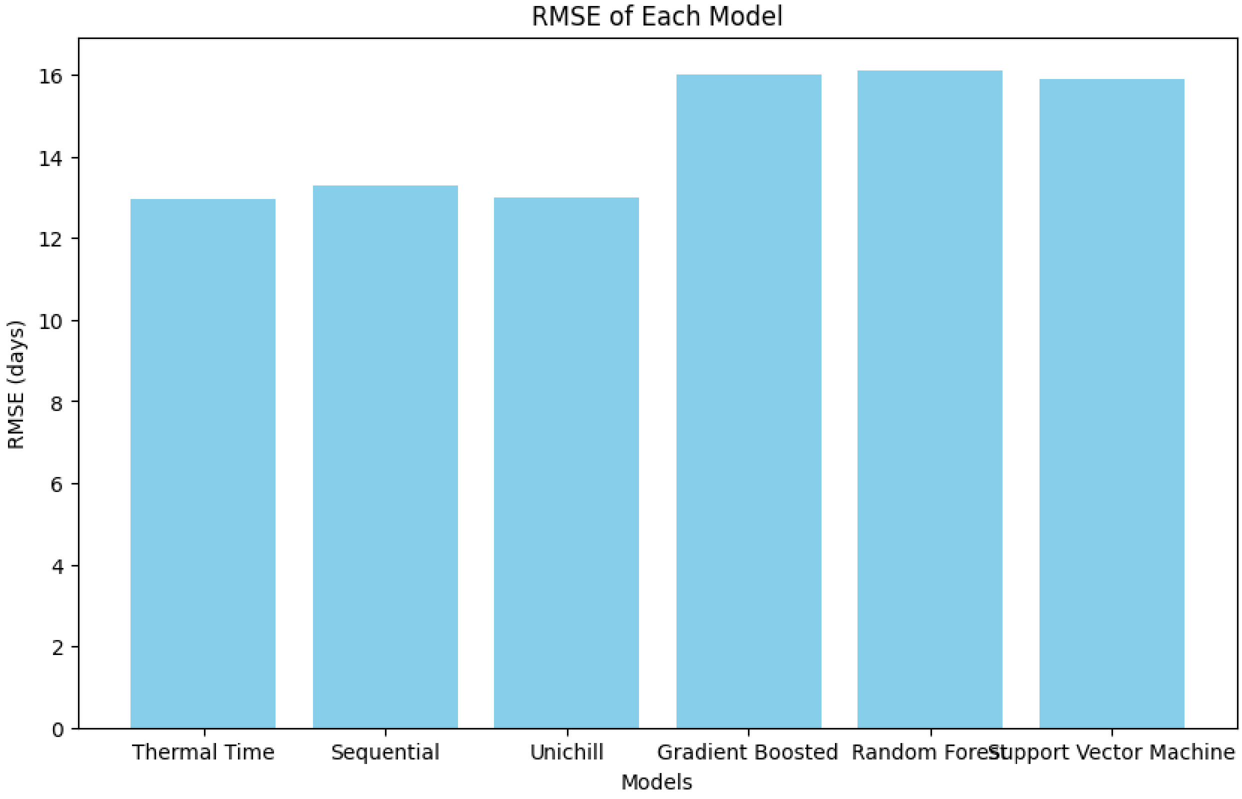


Figure 24. Bar plot of RMSE values for all models

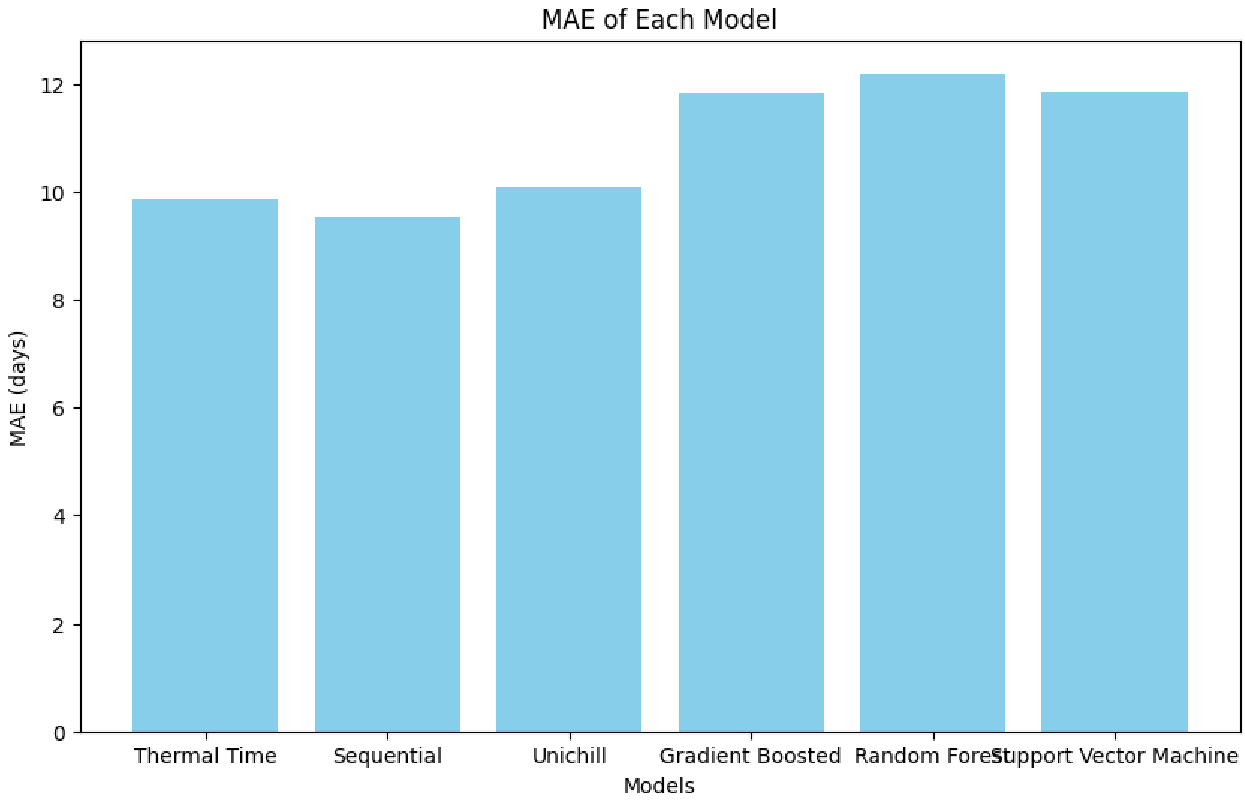


Figure 25. Mean Absolute Error per model

Figure 26 shows a boxplot of the distribution of the prediction errors, which is the absolute difference between a predicted leaf emergence date and an actual leaf emergence date, for every model. The median line for all models is relatively close to 0, generally meaning a good model performance. The further away from 0 this line is, the lower the model performance on average. The box size for all models is again relatively even, except for the S model, which is significantly smaller. This indicates that the prediction errors are more tightly clustered around the median, often suggesting more consistent performance. A larger box would indicate more variability in prediction errors. Lastly, there is a significantly lower distance between outliers and the box for the ML models. Fewer and less extreme outliers in a model indicate that a model is more robust, which could be said for the ML models. Lastly, all PB models seem to overpredict the actual leaf emergence date and the ML models seem to underpredict the actual leaf emergence dates. It is difficult to state which models perform best according to Figure 26, so for that, the earlier calculated RMSE and MAE are the determining factors.

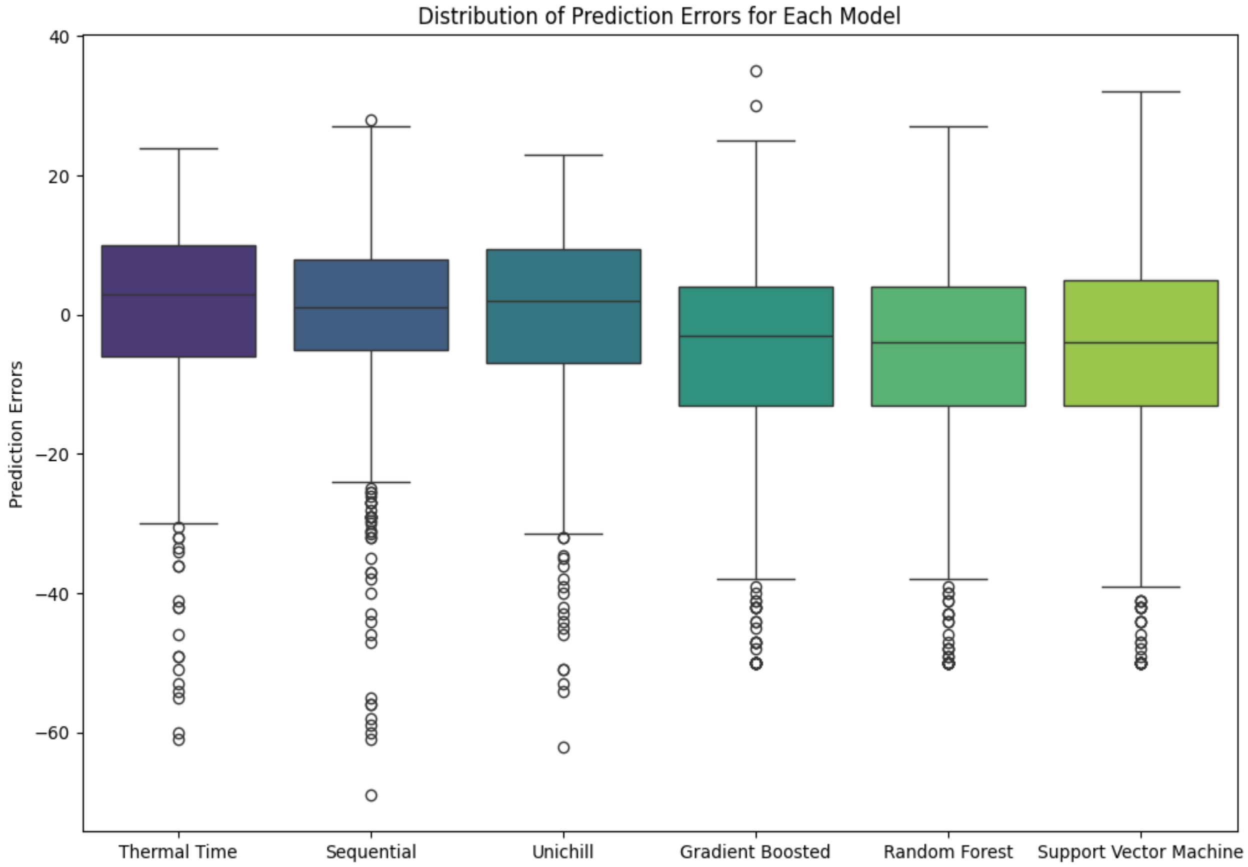


Figure 26. Box plot of prediction errors between models

Figure 27 shows a combination of Figures 15, 16, 17, 21, 22, and 23 to show the differences between models better. The main difference between the PB models and the ML models is that the predicted leaf emergence dates of the former are mostly concentrated around day 120 of the year. In general, the scatterplots of the ML models follow the perfect line better than the plots of the PB models.

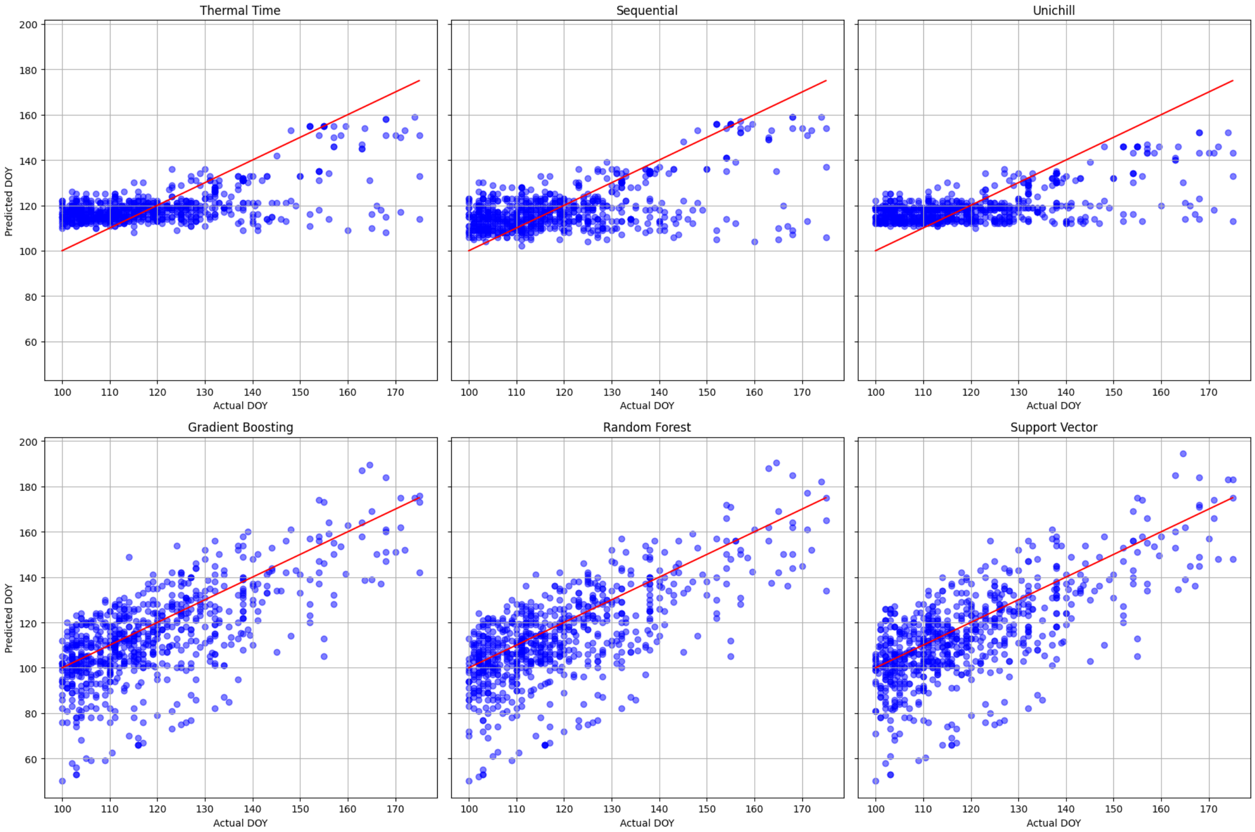


Figure 27. Scatterplots of all six models

Discussion

Within this study, the main research questions and sub-questions have been answered. However, there are a multitude of factors that may have influenced the study negatively or positively. These could be taken into account in future research or may have needed to be investigated more extensively in the current study.

From the PB models, the TT model performed best. This could be because this model is straightforward and therefore avoids potential overfitting. Additionally, it has fewer parameters than the S or UC model, making it less sensitive to errors in parameter estimation, which could lead to better performance. Another reason might be that the used data is better suited to the TT model, which is likely. Deciduous forests in this area might have a leaf emergence mechanism that aligns closely with the way the TT model calculated the leaf emergence date. It could also be that the research area has consistent temperature patterns that fit well with the TT model, or the time scale is a better fit for the model. Lastly, the TT model is one of the earliest developed and most extensively researched models. There might be an empirical fit that aligns with the research context.

The SVM (SVM) model performed best for the ML models. This could be due to a multitude of reasons. First, SVM performs well with many features as input, which was the case as a result of the binary classification of the data. Additionally, SVM is good at mitigating overfitting because of its regularization parameter. Plus, SVM is generally well-suited to binary classification problems, although this is the case with the other models as well.

In general, the PB models performed better than the ML models. This might be because PB models are built to simulate the mechanisms of plant phenology and are grounded in biological principles. The three used models all include parameters that are directly related to the physiological processes that drive leaf emergence. Additionally, PB models are specifically designed to predict leaf emergence dates, while ML models are not inherently designed for this purpose and may not capture non-linear patterns that might be essential. Another plausible reason is that PB models typically require less data for accurate predictions compared to ML models. Due to the removal of outliers, the dataset was not as large as it potentially could have been, resulting in a possibly worse performance of the ML models. Lastly, data was binary classified, which in theory should have resulted in a better performance of the ML models. This was not the case, but without further investigation, it is difficult to exactly state why this is. Most likely, data was best fitted for PB models.

Different models performed better when looking at the RSME value than other models when taking the MAE into account. This could be because MAE is less sensitive to outliers compared to RMSE. Additionally, RMSE gives higher weight to large errors due to the squaring operation and taking the square root afterward. This means that larger errors generally contribute more to overall error in RMSE compared to MAE. Another reason could be the distribution of the errors. If these are skewed, one metric might be more suitable than the others. As a result, it can be stated that PB models performed better than ML models, but differences between the types of models are small and it is difficult to determine which model performed best due to different outcomes for different metrics, such as RMSE and MAE.

When comparing the models more thoroughly more differences become clear. The median of the prediction errors for the PB models is consistently above 0 and that of the ML models below 0. This means that the PB models generally overpredict the leaf emergence date and the ML models underpredict. It is difficult to determine why this is the case, but one reason could be the fact that the training data had a slight bias towards earlier dates. The ML models do not have parameters and only use the training data, whereas the PB models have parameters that could result in this difference. The fact that outliers are generally higher before the leaf emergence dates could also be explained by this factor. Additionally, the ML models seem to perform better when looking at the scatterplots but have a higher RMSE value. These models tend to have a lower bias but higher variance. They can predict a wide range of dates, closely following the actual leaf emergence dates. However, the increased variance might lead to occasional larger errors, which significantly impact the RMSE. Lastly, the models that have lower RMSE values are not the models that perform best in the boxplot graph of prediction errors. A median prediction error close to 0 is a good sign that a model is not systematically biased in predictions but does not guarantee lower RMSE, which is affected more by the presence of large errors.

This study initiated a basic investigation into phenological predictions by using the Springtime Python package. The aim was to predict leaf emergence dates using a repeatable workflow that would allow data to be preprocessed in a structured manner and fed into different models, PB and ML. Notably, the study maintained a simplistic approach, functioning as a first attempt to build a framework using Springtime. The decision to keep the case study basic stems from the need to establish a solid foundation for more complex analyses in future research projects. For instance, this project only used weather data as a form of geospatial dataset, but other datasets can be integrated into the workflow within Springtime. Furthermore, the study's scope is limited by its focus on a single phenophase -leaf emergence- and a single meteorological feature – temperature. This leaves open the possibility of future research on other phenophases and other environmental factors such as humidity, photoperiod, or precipitation. Non-environmental features could also be taken into account, such as coordinates or altitude. Additionally, more models could have been tested so that better conclusions could have been drawn about the best-performing models. There is also an absence of hyperparameter tuning within all the methods, which could lead to less-than-optimal model performance. Hyperparameter tuning allows models to capture more subtle patterns or relationships in the data, leading to the best possible model performance. However, hyperparameter tuning was outside the scope of this study. Additionally, there has also been an inability to use ensembling techniques, which could also pose a limitation in this study. These methods are known to improve predictive performance by combining multiple models to mitigate individual model uncertainties. Consequently, models may have been more susceptible to overfitting or underfitting. Incorporating hyperparameter tuning and ensembling techniques in future research endeavors could improve the accuracy and stability of phenological predictions. Another method that could have improved results is hybrid modeling between PB models and ML models. This could include for instance adjusting the prediction of PB models using ML methods that could account for recent changes in environmental factors. Another example is sequential modeling where PB models could be used to simulate certain aspects of the process and these are given as input for the ML models. This method could improve the general predictive capabilities of the models but is outside of the scope of this research. Essentially, this study lays the groundwork for phenological forecasting using Springtime and acts as a foundation for further investigation and advancement in the field of phenological research.

Furthermore, ground-based observations made by citizens and citizen scientists could have introduced biases or limitations due to the local conditions or observer subjectivity, potentially impacting the generalization of findings. For instance, different researchers may have used different techniques in observing the DOY. Additionally, relying solely on ground-based observations could overlook the large-scale environmental factors that could influence the DOY, such as land use changes. Therefore, in future research, there is a potential to also take remote sensing or satellite data into account.

Additionally, it is essential to acknowledge that the findings within this research are inherently limited to the chosen area and specific tree species. Insights might be valuable for the understanding of leaf emergence dynamics within this context, but extrapolation to other regions or forest types should be approached with caution.

The USA-NPN dataset is strong after 2016, but before 2016 there is a lack of data availability. The reliance on the period after 2016 may overlook long-term trends, potentially limiting the depth and accuracy of the predictions. Moreover, focusing on a relatively short time frame may fail to capture gradual changes in phenological events, again potentially limiting the depth and accuracy of predictions. Additionally, short-term fluctuations in environmental factors may have influenced the results. Underlying trends could have been not shown, influencing the reliability of the conclusions. Furthermore, data availability was even less due to the necessity of taking out outliers, which were close to half of the observations.

The PB models may still fall short in correctly capturing the complex physiological processes underlying leaf emergence. To some extent, these models are simplified representations of tree physiology which could lead to unrealistic predictions, possibly weakening the validity of the results. As a result, the applicability of the PB models in forecasting scenarios might be limited.

This study was in some ways limited due to the nature of the observations, the first attempt to use Springtime to build such a workflow, and other factors as explained above. As a result, the output of the models could be limited or skewed. For instance, there is not a significant difference between the outcomes of the PB models and the outcomes of the ML models. This is also why it is difficult to make strong conclusions about the outcomes of the results. However, the main aim of this research was to build a repeatable workflow within Springtime at the hand of a case study. Therefore it can be stated that the importance of this study was not on the results of the models and case study but rather on the implementation of a functioning workflow. The results of this workflow, in the form of Python scripts, will be provided in the Springtime documentation as examples and are free for anyone to reuse and reproduce, or produce their own results.

Conclusion

Through the use of open-access geospatial information and observations from phenology networks, this study aimed to develop a workflow for phenological modeling. Using Springtime, a Python package developed by the University of Twente, a case study was conducted in which this workflow was created and made usable for different case studies with different phenological species, research areas, and phenophases. Earlier was stated that there was no standardized workflow that allows comparison of different phenological prediction models. Springtime is in the first package that allows you to easily download data, pre-process it, and compare the outcomes of different models through integrated libraries such as pyPhenology and scikitlearn. For this particular case study, a broad region and a longer time period were selected to guarantee the study's significance. In general, this should improve the dependability of the results. Deciduous tree species were selected, distinguished by their unique phenological stages, which adds to the clarity of the findings. Utilizing Springtime, an automated workflow was developed, facilitating the application of various prediction models to different regions and plant and tree species.

Within PB models, the primary focus is on temperature as a critical factor influencing the phenological phases of plants and trees. However, other factors such as photoperiod can also be included to more accurately represent the biological process occurring in these species. The accuracy of these models was assessed by comparing predicted leaf emergence dates with observed dates, allowing for a comprehensive assessment between different models. Three models were tested within this research: the TT model, the S model, and the UC model, all models focused on temperature as a determinative feature. The TT model performed best in this. For the ML models, the data was binary classified and three models were tested: the GB model, the RF classifier model, and the SVM model. In this, the SVM model performed best, although, differences between the models were small. For all models, the RMSE was calculated by using the predicted dates of leaf emergence and the actual dates of leaf emergence The PB models generally outperformed the ML models by a small margin.

A wide array of features can be extracted from geospatial datasets, mostly depending on data availability in these datasets. These factors can include temperature, precipitation, photoperiod, coordinates, altitude, location, and various derivatives of these features. In this study, the extraction process involved creating detailed recipes to assign daily temperature values to each observation for the year the observation was taken. Binary classifying the data allowed for the creation of 100 times more features (depending on the sample size S) than normal. In theory, this makes the prediction ability of the ML models significantly stronger and it is a method that could help in future phenology prediction studies with ML models.

By establishing a workflow that is reproducible for further use, this study paves the way for more extensive and varied applications of phenological prediction models. Springtime allows for the development of a workflow that can be used for different types of data and model testing. The study's methodology and findings have the potential to enhance the understanding of plant phenology because the written code in Python can be used for different research areas, phenological stages, and species. This could ultimately contribute to more accurate and reliable phenological forecasting and management.

References

Adams, S. R., & Langton, F. A. (2005). Photoperiod and plant growth: a review. The Journal of Horticultural Science and Biotechnology, 80(1), 2-10.

Asse, D., Randin, C. F., Bonhomme, M., Delestrade, A., & Chuine, I. (2020). Process based models outcompete correlative models in projecting spring phenology of trees in a future warmer climate. Agricultural and Forest Meteorology, 285, 107931.

Caffarra, A., Donnelly, A., & Chuine, I. (2011). Modelling the timing of Betula pubescens leaf emergence. II. Integrating complex effects of photoperiod into process based models. Climate research, 46(2), 159170.

Chang, Y. W., Hsieh, C. J., Chang, K. W., Ringgaard, M., & Lin, C. J. (2010). Training and testing low-degree polynomial data mappings via linear support vector machine. Journal of machine learning Research, 11(4).

Chuine, I. (2000). A unified model for leaf emergence of trees. Journal of theoretical biology, 207(3), 337-347.

Czernecki, B., Nowosad, J., & Jabłońska, K. (2018). Machine learning modeling of plant phenology based on coupling satellite and gridded meteorological dataset. International journal of biometeorology, 62, 1297-1309.

Dai, W., Jin, H., Zhang, Y., Liu, T., & Zhou, Z. (2019). Detecting temporal changes in the temperature sensitivity of spring phenology with global warming: Application of machine learning in phenological model. Agricultural and Forest Meteorology, 279, 107702.

Dai, W., Jin, H., Zhou, L., Liu, T., Zhang, Y., Zhou, Z., ... & Jin, G. (2023). Testing machine learning models on a binary classification phenological model. Global Ecology and Biogeography, 32(1), 178-190.

Ellis, R. H., Summerfield, R. J., Edmeades, G. O., & Roberts, E. H. (1992). Photoperiod, leaf number, and interval from tassel initiation to emergence in diverse cultivars of maize. Crop science, 32(2), 398403.

Evans, R. A., Holbo, H. R., Eckert, R. E., & Young, J. A. (1970). Functional environment of downy brome communities in relation to weed control and revegetation. Weed Science, 18(1), 154-162.

Gerst, K. L., Kellermann, J. L., Enquist, C. A., Rosemartin, A. H., & Denny, E. G. (2016). Estimating the onset of spring from a complex phenology database: trade-offs across geographic scales. International journal of biometeorology, 60(3), 391-400.

Groot, R., & McLaughlin, J. (2000). Geospatial Data Infrastructure: concepts, cases and good practice. Oxford university press.

Grossiord, C., Sevanto, S., Adams, H. D., Collins, A. D., Dickman, L. T., McBranch, N., ... & McDowell, N. G. (2017). Precipitation, not air temperature, drives functional responses of trees in semi-arid ecosystems. Journal of Ecology, 105(1), 163-175.

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., ... & Senthilnath, J. (2021). Integrated phenology and climate in rice yields prediction using machine learning methods. Ecological Indicators, 120, 106935.

Hill, W. R., & Dimick, S. M. (2002). Effects of riparian leaf dynamics on periphyton photosynthesis and light utilisation efficiency. Freshwater Biology, 47(7), 1245-1256.

Hill, W. R., Mulholland, P. J., & Marzolf, E. R. (2001). Stream ecosystem responses to forest leaf emergence in spring. Ecology, 82(8), 2306-2319.

Hoffmann, F., Bertram, T., Mikut, R., Reischl, M., & Nelles, O. (2019). Benchmarking in classification and regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(5), e1318.

Holway, J. G., & Ward, R. T. (1965). Phenology of alpine plants in northern Colorado. Ecology, 46(1-2), 73-83.

Hopkins, A. D. (1920). The bioclimatic law. Journal of the Washington Academy of Sciences, 10(2), 3440.

Izquierdo-Verdiguier, E., Zurita-Milla, R., Ault, T. R., & Schwartz, M. D. (2018). Development and analysis of spring plant phenology products: 36 years of 1-km grids over the conterminous US. Agricultural and forest meteorology, 262, 34-41.

Jánosi, I. M., Silhavy, D., Tamás, J., & Csontos, P. (2020). Bulbous perennials precisely detect the length of winter and adjust flowering dates. New Phytologist, 228(5), 1535-1547.

Jozefowicz, S., Stone, M., & Aravopoulou, E. (2020). Geospatial data in the UK. The Bottom Line, 33(1), 27-41.

Kalverla, P., Alidoost, F., Verhoeven, S., & Khodadadzadeh, M. (2024). Springtime (v0.2.0). Zenodo. https://doi.org/10.5281/zenodo.10610958

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Keenan, T. F., & Richardson, A. D. (2015). The timing of autumn senescence is affected by the timing of spring phenology: implications for predictive models. Global change biology, 21(7), 2634-2641.

Kramer, K. (1994). Selecting a model to predict the onset of growth of Fagus sylvatica. Journal of Applied ecology, 172-181.

Lindsey, A. A., & Newman, J. E. (1956). Use of official weather data in spring time: temperature analysis of an Indiana phenological record. Ecology, 37(4), 812-823.

Linkosalo, T., Lappalainen, H. K., & Hari, P. (2008). A comparison of phenological models of leaf bud burst and flowering of boreal trees using independent observations. Tree physiology, 28(12), 18731882.

Livingston, B. E., & Livingston, G. J. (1913). Temperature coefficients in plant geography and climatology. Botanical Gazette, 56(5), 349-375.

Lutcher, L. K., Wuest, S. B., & Johlke, T. R. (2019). First leaf emergence force of three deep-planted winter wheat cultivars. Crop Science, 59(2), 772-777.

Mehdipoor, H., Zurita-Milla, R., Izquierdo-Verdiguier, E., & L Betancourt, J. (2019). Exploring the Influence of Source and Scale of Phenological Model Inputs at Continental Scale. GeoComputation 2019.

Melaas, E. K., Friedl, M. A., & Richardson, A. D. (2016). Multiscale modeling of spring phenology across Deciduous Forests in the Eastern United States. Global change biology, 22(2), 792-805.

Merriam, C. H. (1894). Laws of temperature control of the geographic distribution of terrestrial animals and plants.

Mitchell, T. M. (1997). Machine learning.

O'Brien, O., & Cheshire, J. (2016). Interactive mapping for large, open demographic data sets using familiar geographical features. Journal of Maps, 12(4), 676-683.

Pellagatti, M., Masci, C., Ieva, F., & Paganoni, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. Statistical Analysis and Data Mining: The ASA Data Science Journal, 14(3), 241-257.

Polikar, R. (2012). Ensemble learning. Ensemble machine learning: Methods and applications, 1-34.

Réaumur, R. A. (1735). Observations du thermomètre faites pendant l'année MDCCXXXV comparées a celles qui ont été faites sous la ligne a l'Isle-de-France, a Alger et en quelques-unes de nos Isles de l'Amérique. Mémoires de l'Académie Royale des Sciences, 1735, 545-576.

Roberts, A. M. I. (2012). Comparison of regression methods for phenology. International Journal of Biometeorology, 56, 707-717. ISO 690

Rosemartin, A. H., Denny, E. G., Gerst, K. L., Marsh, R. L., Posthumus, E. E., Crimmins, T. M., & Weltzin, J. (2018). USA National Phenology Network observational data documentation (No. 20181060). US Geological Survey.

Saleska, S. R., Didan, K., Huete, A. R., & Da Rocha, H. R. (2007). Amazon forests green-up during 2005 drought. Science, 318(5850), 612-612.

Sarvas, R. (1972). Investigations on the annual cycle of development of forest trees. Active period. Investigations on the annual cycle of development of forest trees. Active period., 76(3).

Sharratt, B. S., Sheaffer, C. C., & Baker, D. G. (1989). Base temperature for the application of the growingdegree-day model to field-grown alfalfa. Field Crops Research, 21(2), 95-102.

St Peter, J. R. (2015). A Model For Determining Drivers of Phenology in Western United States Rangelands.

Tang, J., Körner, C., Muraoka, H., Piao, S., Shen, M., Thackeray, S. J., & Yang, X. (2016). Emerging opportunities and challenges in phenology: a review. Ecosphere, 7(8), e01436.

Thessen, A. E. (2016). Adoption of machine learning techniques in ecology and earth science (No. e1720v1). PeerJ PrePrints.

Thornton, M. M., Thornton, P. E., Wei, Y., Mayer, B. W., Cook, R. B., & Vose, R. S. (2018). Daymet: annual climate summaries on a 1-km grid for North America, Version 3. ORNL DAAC, Oak Ridge, Tennessee, USA.

Uçar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The effect of training and testing process on machine learning in biomedical datasets. Mathematical Problems in Engineering, 2020.

Van de Pol, M., & Bailey, L. D. (2019). Quantifying the climatic sensitivity of individuals, populations, and species. Effects of climate change on birds, 44-59.

Voigt, J. W., & Mohlenbrock, R. H. (1964). Plant communities of southern Illinois. SIU Press.

Wang, J., Zhang, X., & Rodman, K. (2021). Land cover composition, climate, and topography drive land surface phenology in a recently burned landscape: An application of machine learning in phenological modeling. Agricultural and Forest Meteorology, 304, 108432.

Warrington, I. J., & Kanemasu, E. T. (1983). Corn Growth Response to Temperature and Photoperiod II. Leaf-Initiation and Leaf-Appearance Rates 1. Agronomy Journal, 75(5), 755-761.

Wei, Q., & Dunbrack Jr, R. L. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. PloS one, 8(7), e67863.

Zhou, X., Xin, Q., Dai, Y., & Li, W. (2021). A deep-learning-based experiment for benchmarking the performance of global terrestrial vegetation phenology models. Global Ecology and Biogeography, 30(11), 2178-2199.