



Exploring differences in spatial patterns and temporal trends of phenological models at continental scale using gridded temperature time-series

Hamed Mehdipoor¹ · Raul Zurita-Milla¹ · Ellen-Wien Augustijn¹ · Emma Izquierdo-Verdiguier^{1,2}

Received: 1 October 2018 / Revised: 15 October 2019 / Accepted: 24 October 2019
© The Author(s) 2019

Abstract

Phenological models are widely used to estimate the influence of weather and climate on plant development. The goodness of fit of phenological models often is assessed by considering the root-mean-square error (RMSE) between observed and predicted dates. However, the spatial patterns and temporal trends derived from models with similar RMSE may vary considerably. In this paper, we analyse and compare patterns and trends from a suite of temperature-based phenological models, namely extended spring indices, thermal time and photothermal time models. These models were first calibrated using lilac leaf onset observations for the period 1961–1994. Next, volunteered phenological observations and daily gridded temperature data were used to validate the models. After that, the two most accurate models were used to evaluate the patterns and trends of leaf onset for the conterminous US over the period 2000–2014. Our results show that the RMSEs of extended spring indices and thermal time models are similar and about 2 days lower than those produced by the other models. Yet the dates of leaf out produced by each of the models differ by up to 11 days, and the trends differ by up to a week per decade. The results from the histograms and difference maps show that the statistical significance of these trends strongly depends on the type of model applied. Therefore, further work should focus on the development of metrics that can quantify the difference between patterns and trends derived from spatially explicit phenological models. Such metrics could subsequently be used to validate phenological models in both space and time. Also, such metrics could be used to validate phenological models in both space and time.

Keywords Spring plant phenology · Citizen science · Climate change · Spatio-temporal trends · Cloud computing · Gridded temperature time-series

Introduction

Climate change is influencing the timing of key biological events. For example, increasingly warm springs are advancing the time of leaf onset of plants (Ellwood et al. 2013; Schwartz et al. 2013) and the migration of animals (Marra et al. 2005; Ault et al. 2011). Monitoring and analysing the timing of plants and animal development events is therefore essential

to a better understanding of the Earth system and defining climate change adaptation strategies (Briske et al. 2015; Gerst et al. 2016; Labe et al. 2017). Phenology is the science that studies periodic plant and animal life cycle events (phenophases) and how seasonal and inter-annual variations in environmental conditions affect both (Lieth 1974; Chmielewski 2013). Phenology is also an useful proxy of the impact of climate change on our planet (Doi and Katano 2008; Gordo and Sanz 2010; Zurita-Milla et al. 2017).

Plant phenological models support the study of the impacts of climate change and inter-annual weather variability on vegetated canopies (Badeck et al. 2004; Schwartz et al. 2006; Allstadt et al. 2015). Phenological models can also be used to reconstruct and qualify ground observations (Chuine et al. 2004; Menzel 2005), to estimate species-specific phenology (Chuine et al. 2000), and species performance (Basler 2016). Phenological models are often calibrated using ground and weather observations. Spring

✉ Hamed Mehdipoor
h.mehdipoor@utwente.nl

¹ Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands

² Image Processing Laboratory (IPL), Universitat de Valencia, Valencia, Spain

plant phenology (SPP) models are particularly interesting (Cayan et al. 2001; Schwartz et al. 2006; Matzarakis et al. 2010; Allstadt et al. 2015). SPP models are widely used to support management decisions and to design suitable adaptation strategies for ecological and agricultural systems (Enquist et al. 2014; Gerst et al. 2016). These uses require phenological models of high quality.

The process of quality control of models starts with their calibration and continues with their validation. Calibration is used to find the values of the model parameters that minimise the error of the model. These values are often derived using specific phenological and environmental datasets. This makes the comparisons between calibrated models challenging. Validation is used to check the error of the calibrated model. The calibration and validation of phenological models over large areas are now possible because we have access to continental-scale gridded weather time series such as daily temperature and large amounts of contemporary volunteered phenological observations (VPOs; (Rosemartin et al. 2015)). Gridded temperature time-series are key inputs to SPP models; they help to generate spatially continuous data from which patterns and temporal trends can be extracted to help evaluate the influence of climate change on plant development (Mehdipoor et al. 2018; Izquierdo-Verdiguier et al. 2018).

SPP models vary in their mathematical formulation as well as in their input variables (Schwartz and Marotz 1988; McCabe et al. 2012; Allstadt et al. 2015). Hufkens et al. (2018) developed a modelling framework that includes a comprehensive list of SPP models and data. They describe historical SPP models such as extended spring indices (SI-x; (Schwartz et al. 2013)), thermal time (TT; (Cannell and Smith 1983)) and photothermal-time (PTT; (Masle et al. 1989; Črepinšek et al. 2006)) and lilac leafing observations. These models have been widely used to study the effect of climate change and global warming on both natural and agricultural ecosystems (Schwartz et al., 1988; Schwartz 1999; Brunsdon and Comber 2012; Richardson et al. 2013; Rosemartin et al. 2015). The framework facilitates both reproducibility and community driven phenology model evaluation and development.

Exploring the effects of using these phenological models improves our understanding of the outputs from SPP models, and consequently improves the quality of the management decisions based on the outputs. Geoscientists have explored the effects of using various models at large scales. However, high resolution and open daily gridded weather data and extensive phenological observations only recently became available (Abatzoglou 2013). As a result, few studies have comprehensively explored the effects of using one or another model on phenological patterns and trends at a high spatial resolution and over large regions. The improvements in large-scale distributed computing and cloud computing facilitate the qualification of SPP models at higher spatial resolution, by using gridded input data (Guo et al. 2010; Mehdipoor et al. 2017).

The quality of phenological models is often assessed using measures of the differences such as the root-mean-square error (RMSE) between the estimated and observed day of the year (DOY) of a phenophase at a number of observations sites. The RMSE is widely used to measure the performance of the models. However, RMSE might be a misleading indicator of average error or variability because RMSE is a function of both the mean of a set of error and their variability that influence the value of RMSE. This prevents a clear assessment of RMSE interpretations (Willmott and Matsuura 2005, Willmott et al. 2009). Moreover, RMSE needs observed DOYs for every grid cell to continuously measure the performance of the models in space, which are not always available. RMSE is often calculated using estimated DOY at a set of locations for which reference observations are available. The RMSE is an average model performance measure, and it can therefore overestimate or underestimate the average model error as a result of the spatial distribution of reference observations (Willmott et al. 2017).

Several studies have shown that phenological patterns and trends derived from various kinds of models may vary considerably (Janssen and Heuberger 1995; García de Cortázar-Atauri et al. 2009; Basler 2016; Chuine et al. 2016). Although researchers and decision makers use metrics such as RMSE to select models, they might not be aware of the effect of selecting models that use error measures such as RMSE. Further, researchers aiming to model plant phenology needs to be cautious in interpreting the results, both in a historical context and especially when looking forward over the next century. Moreover, global change models are able to integrate sophisticated information about land surface model components such as phenological model outputs (Willmott et al. 2009). The output of phenological models must be explored and evaluated to give confidence that global change models will accurately simulate future changes in the land-atmosphere coupling, carbon storage and feedbacks that may affect both weather and ecology.

This study evaluates various SPP models and explores the impact of using one or another model on the phenological patterns and trends that can be extracted by running these models at continental scales. We go beyond model selection and showed various model formulations for the same event. The workflow uses VPOs and gridded time-series temperature over the conterminous US (CONUS). We illustrate the workflow by exploring the effect of using the SI-x, TT and PTT groups of models on the estimation of patterns and trends in DOY of lilac leafing.

Materials and methods

Volunteered observations and temperature data

Two datasets were used to calibrate and validate the selected SPP models. For calibration of the TT and PTT models, we collected

historical lilac (*Syringa Chinensis* ‘Red Rothomagensis’) VPOs and their corresponding temperature data. These VPOs, collected over the period 1961–1994 and containing the geographic location and DOY of first leaf (FL), were originally used to calibrate the SI-x for leafing onset (Schwartz 1997; Ault et al. 2015a, b). In this study, we use the original SI-x lilac leafing model (SI-xLM). The day of lilac FL provides a standard reference of spring plant phenology that can be compared with various locations and years (Caprio 1974; Santer 1985). Furthermore, this phenophase responds directly to changes in temperature and day-length as opposed to changes in other environmental cues (Caprio 1993; Schwartz et al. 2012). The lilac VPOs dataset has a total of 2321 observations collected across 193 sites (known as phenological stations) distributed over the continental US. The corresponding daily maximum and minimum temperatures and a day-length dataset include records from the nearest weather stations

to the VPO sites. These weather stations are part of the Global Historical Climatology Network (GHCN).

To validate the selected SPP models and to explore the patterns and trends they produce, we used VPOs from the USA National Phenology Network (USA-NPN) and Daymet gridded temperature time-series, from 2000 to 2014. The USA-NPN dataset contains 899 lilac FL observations that were checked for consistency (Mehdipoor et al. 2015; Rosemartin et al. 2015). The gridded daily maximum and minimum temperatures and day-lengths from Daymet have been available at a 1-km spatial resolution for most of North America since 1980 (Daly et al. 2008; Thornton et al. 2014). Daymet dataset uses spatially-referenced surface measurements of daily maximum and minimum temperature and precipitation from the Global Historical Climatology Network (GHCN) and the land/water mask derived from the NASA

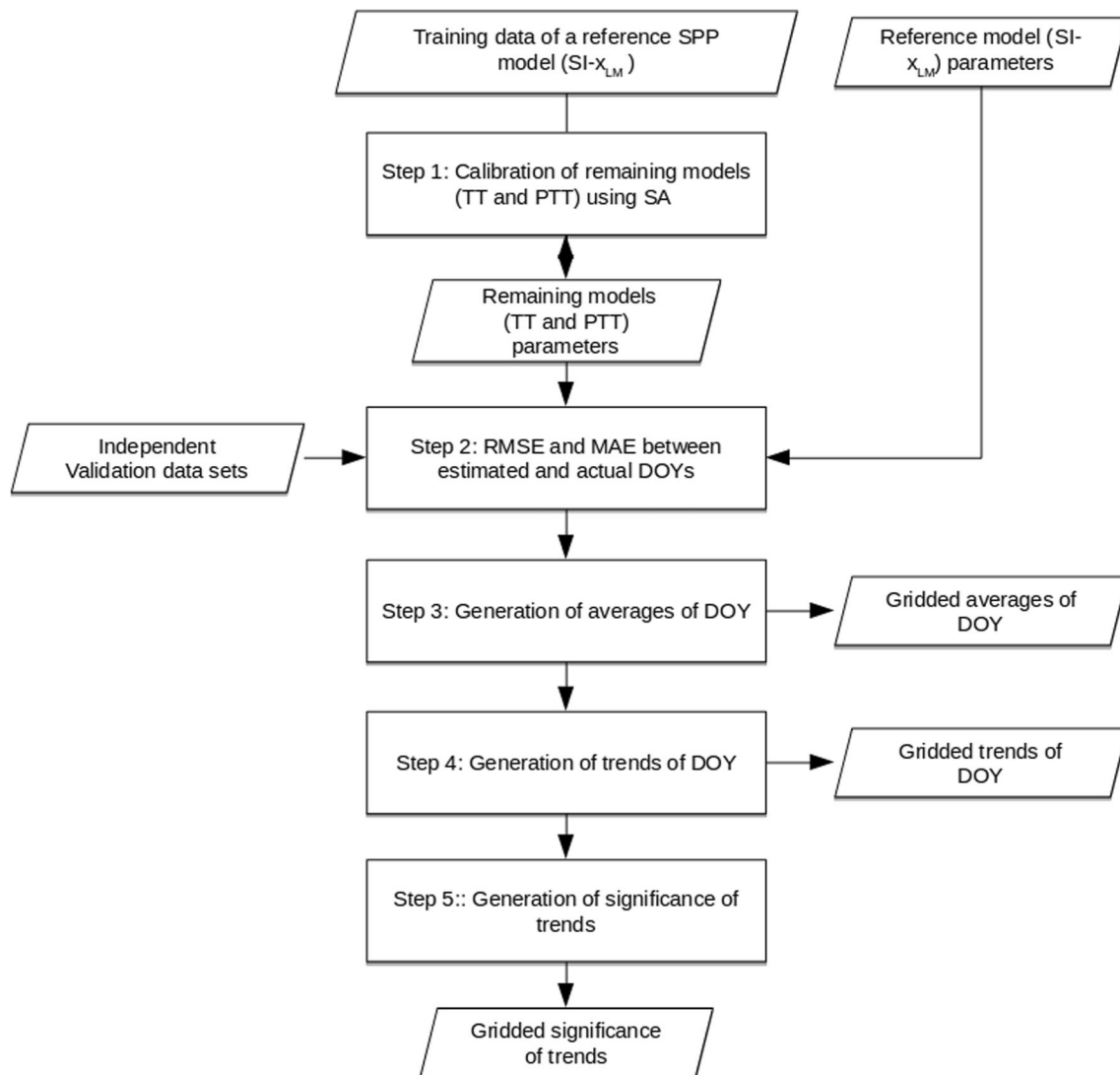


Fig. 1 The main analysis steps for calibrating and comparing models as well as for generating and comparing the spatio-temporal trends

Table 1 Calibrated parameters for the TT, TTS, photothermal TT, photothermal TTS models and the corresponding RMSE of the calibration dataset

Model name	Model parameters				RMSE	MAE
	T_b	d	e	F^*		
TT	2.34			220	11.48	6.61
TTS	7.22	-0.35	4.75	28	11.92	6.64
Photothermal TT	0.26			151	12.57	7.17
Photothermal TTS	6.23	-0.75	3.04	14	12.63	6.89

SRTM 30 arc second DEM as major inputs (Daly et al. 2008; Thornton et al. 2014). Daymet daily maximum and minimum temperatures and day-length, and SI-x Daymet-based products (Izquierdo-Verdiguier et al. 2018) are available at a 1-km spatial resolution for the continental US from 1980 to 2016. Compared with other datasets such as GridMET (Mitchell et al. 2004) and PRISM (Daly et al. 2008), the higher spatial and temporal resolutions of Daymet helps to better capture temperature regimes across complex topographies, particularly in the western USA (Mehdipoor et al. 2018).

SSP models used

We selected SSP models that were based on temperature and photoperiod because plant phenophases often respond to changes in the daily values of these variables (Caprio 1993; Schwartz et al. 2012). Our selection includes

$$\frac{1000}{DOY_{Observed}} = A_1 \frac{MDS0}{DOY_{Observed}} + A_2 \frac{SYNOP}{DOY_{Observed}} + A_3 \frac{DD57}{DOY_{Observed}} + A_4 \frac{DDE2}{DOY_{Observed}} \tag{1}$$

$$3.306 * MDS0 + 13.787 * SYNOP + 0.201 * DDE2 + 0.153 * DD57 \geq 1000 \tag{2}$$

Thermal time (TT) and thermal time calibrated using a sigmoid temperature response (TTS)

Phenology models are the second group of SPP models considered in this study. These models consider only the role of forcing temperatures (i.e. temperatures at which the plant develops). These models assume that phenophases such as the FL phenophase occur when a critical state of forcing is reached. The state of forcing is modelled as the sum of the

SI-xLM

The SI-x models are widely used to study the timing of plant leafing and its changes in the northern hemisphere (Linkosalo et al. 2008; Mehdipoor et al. 2016; Wu et al. 2016; Belmecheri et al. 2017; Hufkens et al. 2018). The outputs of the SI-x model, namely the estimated DOY of FL and first flower of indicator plants such as lilac, are used as an official indicator of climate change in the USA (Schwartz et al. 2006, 2013; Crimmins et al. 2016). SI-xLM was first calibrated about 25 years ago, using the original VPOs and daily minimum and maximum temperatures (Schwartz 1997; Ault et al. 2011; Schwartz et al. 2013; Ault et al. 2015a, b).

The SI-xLM is based on the measure of growing degree hours (GDH: the sum of the hourly temperatures above 31 °F (- 0.5 °C), which is calculated from daily minimum and maximum temperatures. These GDHs are used to define the accumulation of short and long-term variables (Schwartz et al. 2013). Estimators of SI-xLM include days since the 1st of January (MDS0), 5–7 day GDH accumulations (DD57), 0–2 day GDH accumulations (DDE2) and accumulation of the number of high-energy synoptic events, which is defined as a DDE2 greater than 637, (SYNOP). A regression of the form of Eq. 1 was fitted to calibrate SI-xLM coefficients (A1, A2, A3 and A4), using estimators’ values at observed DOYs. For prediction, the calibrated coefficients are used to operationally check the inequalities of Eq. (2) on a daily basis, starting on the 1st of January. For underlying assumptions and a more detailed definition of the SI-x models, please see (Ault et al. 2015a, b).

daily rate of forcing (R_f), which is purely a function of temperature. We calibrated TT and TTS using Eqs. 3 and 4:

$$R_{f_{TT}}(T) = \begin{cases} 0 & \text{if } T < T_b \\ T - T_b & \text{if } T \geq T_b \end{cases} \tag{3}$$

$$R_{f_{TTS}}(T) = \begin{cases} 0 & \text{if } T < T_b \\ \frac{1}{1 + \exp(d(T - e))} & \text{if } T \geq T_b \end{cases} \tag{4}$$

where the T_b is the base temperature (i.e. the minimum temperature required for plant development), T is the daily average temperature, d and e correspond to the slope at the inflection point (width) and the temperature of the mid-response (centre) of the sigmoidal function. For both

temperature response functions, the summation of the daily rate of forcing was calculated from the 1st of January (DOY=1) as in inequality 5:

$$S_f(t_s) = \sum_1^{t_s} R_f(T) \geq F^* \quad (5)$$

where t_s represent the DOY at which FL is reached, and F^* shows the amount of heat that must be accumulated by the plant to reach that state of forcing $S_f(t_s)$.

Photothermal TT and photothermal TTS

Photothermal TT and photothermal TTS are the third group of models considered in this study. These models depend on the average temperature during day-length (Burghardt et al. 2015). The summation of R_f was converted to photothermal units by adding a photoperiod variable to increase the biological relevance of inequality 5 (Črepinšek et al. 2006). The daily rate of forcing for the PTT models (R_{fPTT}) is defined as the multiplication of the light period as a proportion of a day to R_f of TT and TTS (Eq. 6). The PTT models, photothermal TT and photothermal TTS, apply the same approach to check if the plant received the amount of heat that is needed to reach a certain state of forcing, as defined in Eqs. 3 and 4, where L_i is the length of day expressed in hours.

$$R_{fPTT} = \frac{L_i}{24} R_f \quad (6)$$

Exploring patterns and trends of models

We use the workflow presented in Fig. 1 to analyse and compare the discussed above models. First, we searched for the optimal set of parameters for the TT, TTS, photothermal TT and photothermal TTS models using the VPOs, temperature and day-length datasets used to calibrate $SI-x_{LM}$, from 1961 to 1996. In particular, in the first step, we used simulated annealing (SA) to find the models' parameters (i.e. T_b , e , d and F^*). SA is a probabilistic optimization algorithm that performs a search in large multidimensional space (Aarts and Korst 1988). This algorithm is robust against entrapment in local optima (van Laarhoven and Aarts 1987), so we used it to find the optimal set of coefficients that minimizes the objective function. Here, our objective function is the RMSE between the observed and estimated DOY from the calibration dataset. We also used the mean absolute error (MAE) between the observed and estimated DOY from the calibration dataset (Willmott and Matsuura 2005) because this error metric is often used to assess phenological models (Schwartz and Marotz 1988; Zavalloni et al. 2006; Matzarakis et al. 2010). SA uses an initial random set of parameters for the objective function to start its search, then it follows steps within

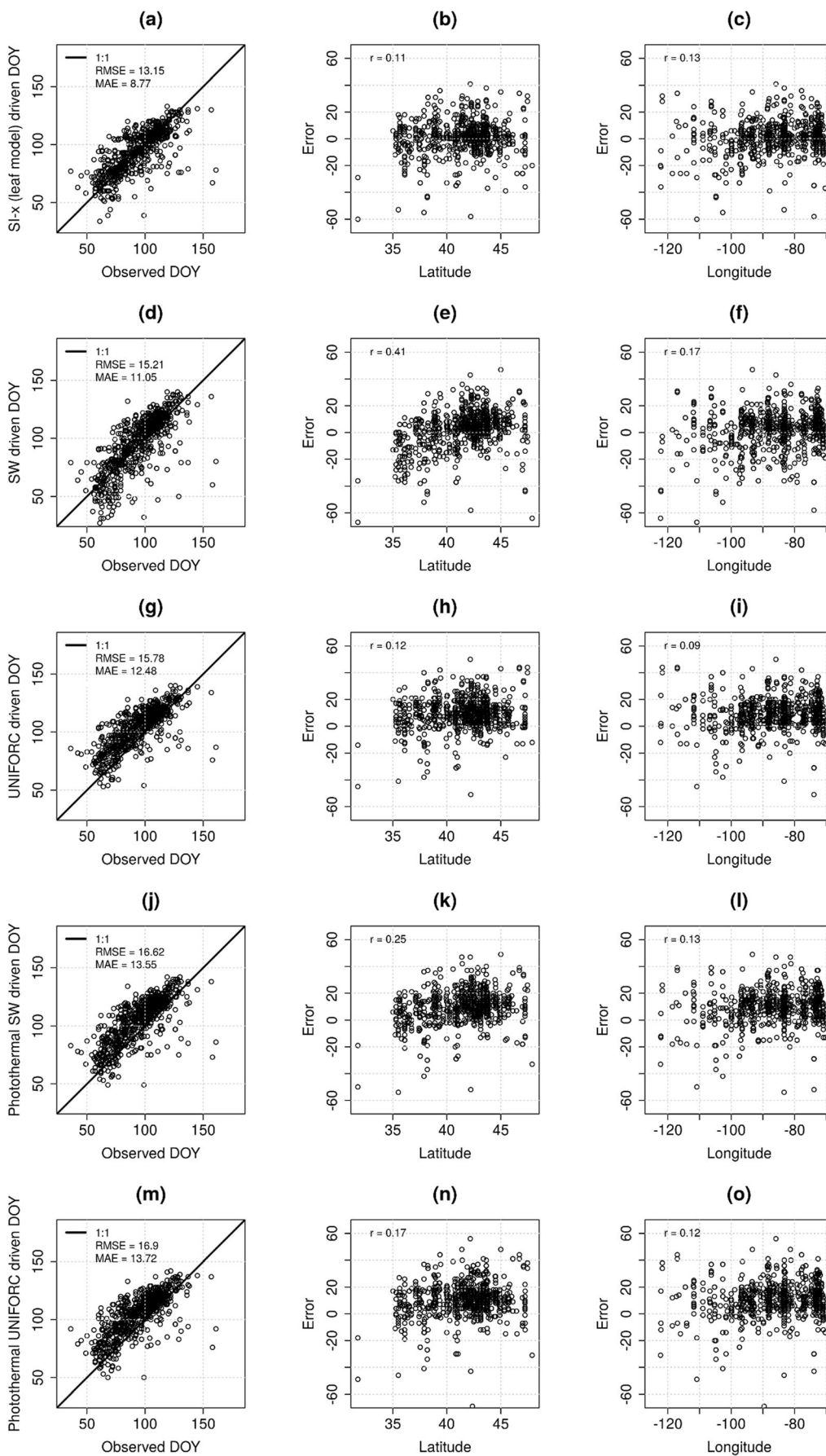
predetermined ranges. In this case, we used a range between $[0, 10]$ for T_b , $[-2, 0]$ for d , $[0, 1000]$ for F^* and $[0, 5]$ for e .

Next, in the second step, $SI-x_{LM}$ and the calibrated TT, TTS, photothermal TT and photothermal TTS models were validated using contemporary VPOs and Daymet data over the period 2000–2014. This period was selected because consistent VPOs were only available for these years. Daily temperature and day length were extracted for the VPO locations. These data were used to estimate DOY at VPOs locations. The RMSE and MAE of all models were calculated by comparing the observed and estimated DOY. Scatterplots of the observed and estimated DOYs were generated to qualitatively explore the effect on the RMSE and MAE of the model. The scatterplots between the model errors (i.e. subtraction of observed DOY from estimated ones) and geographic coordinates (i.e. latitude and longitude) were generated to analyse the model output further and to help better understand how model errors propagate over space.

In the third step, the average annual rate of change of DOY was calculated for each grid cell across CONUS to reveal variations in spatial and spatio-temporal trends between various geographic locations. Spatially continuous model outputs were only obtained for the two most accurate models to compare the effects of these models on the estimation of their average rate of change. Since we need to calculate daily variables for so many spatial locations (at 1-km resolution for the complete CONUS), we used the Google Earth Engine (GEE) cloud computing platform. The implementation of the $SI-x_{LM}$ model in the GEE (Gorelick et al. 2017; Izquierdo-Verdiguier et al. 2018) was used to generate annual and average outputs for the models. This helped to explore and compare regional variations between the gridded model outputs. The histograms of the differences were plotted to provide a geographic representation of their distribution.

In the third steps, the annual differences between the two model outputs were spatio-temporally clustered to provide an overview of the effect of the models on spatial and temporal trends in DOY. The annual differences are the pairwise subtraction of grid cell values (i.e. 1-km by 1-km temperature-driven DOY) of the TT product from the $SI-x_{LM}$ product, from 2000 to 2014. We applied k means clustering in GEE, which is a widely used clustering technique that seeks to minimize the average squared distance between data in the same cluster. For underlying assumptions and details of the GEE k means clustering, please see Arthur and Vassilvitskii 2007. Since the number of clusters cannot be known a priori, we set it to seven as an example. Clusters define regions for which the difference between the estimates of the two models was similar over the selected years. Clusters were mapped and explored to help understand the spatial variation of the regions.

Finally, for both models, the temporal trend in the gridded products was calculated and compared. For each grid cell, the temporal trend was obtained by fitting a linear regression line to



◀ **Fig. 2** Scatterplots between observations and predictions by SPP models (first column), latitude and model error (second column) and longitude and models' error (third column). Estimation errors were calculated by subtracting observed DOY from the estimated ones

the annual products from 2000 to 2014. The slope of the line is the rate of change of the models per year. We calculated the difference between the trends, by subtracting grid cell values of the two models. This highlighted regions where the estimated rates of change were highest. The statistical significance (p value) of these trends was analysed and mapped to show areas showing clear phenological changes. We applied the 2-sided p value test to see if the estimated trend is significantly greater than 0 and if the mean significantly less than 0.

Results and discussion

The SA-calibrated parameters of the TT and PTT models are shown in Table 1. The train RMSE of the models is similar to that which Hunter and Lechowicz (1992) reported. The difference in model parameters indicates that models can be parameterized to provide good predictions; however, their parameters are “biologically meaningless” (Hunter and Lechowicz 1992). The calibrated parameters result in similar minimum values of the RMSE objective function (12 days). These values are similar to the RMSE value calculated using the SI- x_{LM} (11 days). Such similarities indicate that the results given by selected phenological models can fit the historical data equally well. Similarly, SW has the smallest MAE compared with those of the other models. However, it is worth noting that using the RMSE or the MAE to evaluate the trained models results in a different ranking for the photothermal SW and photothermal UNIFORC models. Further analysis could help to assess if the performance of the selected models is significantly different, for example, via bootstrapping of model statistics (Leung et al. 2003; Noguchi et al. 2011).

The results of the validation (Fig. 2) shows that the SI- x_{LM} is more accurate by 2 to 3 days in its estimates than the other models. The MAE also highlighted the similar difference in the error of the models. The visual exploration of scatterplots of both observed and estimated DOY from TTS, photothermal TT and photothermal TTS estimations are more biased to later DOYs than are those of TT and SI- x_{LM} (Fig. 2). Although there was no significant correlation between the model's error and geographic gradients (in terms of latitude and longitude), correlation coefficients are higher for TTS than for other leaf models. Figure 2e shows that the TT model produces a higher error in the northern and southern CONUS compared with the centre. This is because for high and low latitudes, the temperature, which is the basis of the TT model, might not be the only driver of lilac FL.

The average DOY from SI- x_{LM} and TT were mapped, generalizing grid cell values into half-months of DOYs (Fig. 3). For both models, DOYs range from January to June across CONUS. Moreover, latitudinal patterns can be observed in the eastern and elevational patterns in the western CONUS. The visual exploration of the generated products shows that DOYs from the SI- x_{LM} model (Fig. 3a) are different from the results from the TT model (Fig. 3b). In the eastern CONUS, estimated DOYs from TT are earlier in the south, and later in the north than those estimated from SI- x_{LM} . For example, TT mostly estimates DOY in early May while SI- x_{LM} estimates it to be in late April. Similarly, in the western CONUS, DOYs driven from TT are earlier in low altitude regions, and later at high altitude regions, compared with DOYs driven from SI- x_{LM} . TT estimates DOY in early January for most locations in California; however, SI- x_{LM} estimates this in early February.

The histogram of difference in DOY for SI- x_{LM} and TT (Fig. 4) indicates that for most locations in CONUS, the estimated average DOY is 11 days different. At these locations, TT estimates are for earlier dates than those produced by SI- x_{LM} . Although the RMSE of SI- x_{LM} and TT were only different by two days, the estimates can show up to a month difference in the West and North West of the CONUS (Fig. 5). For

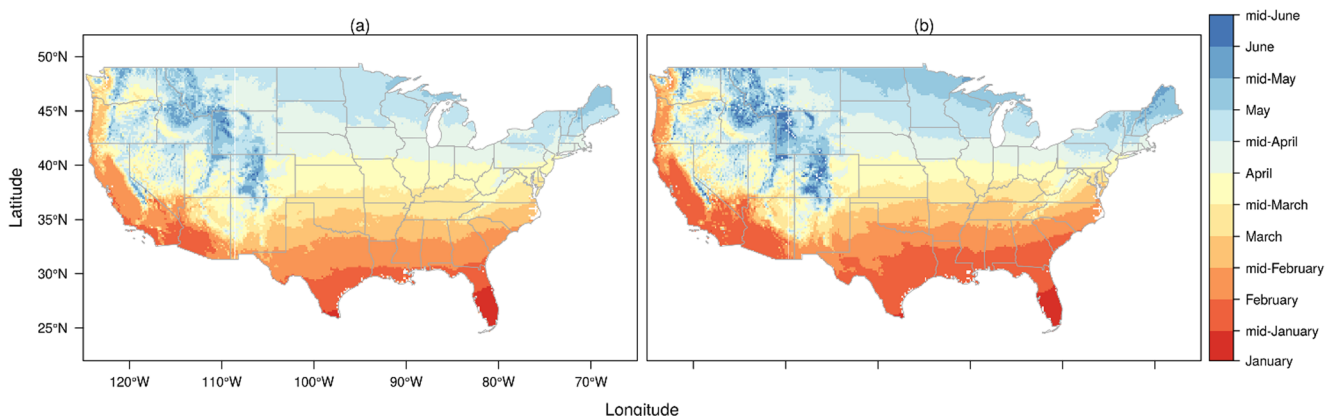
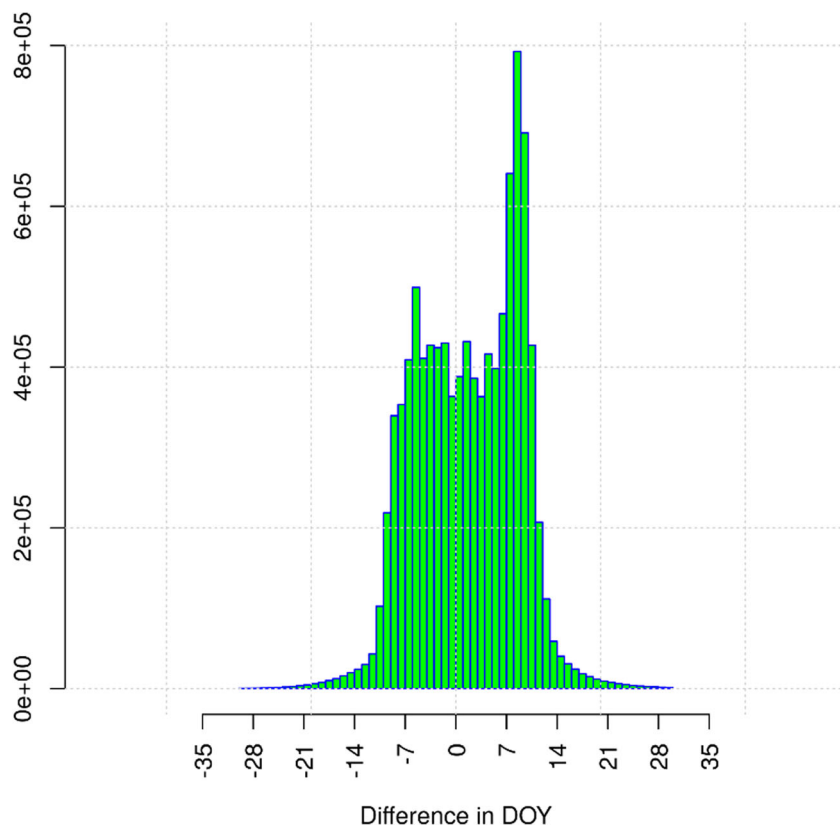


Fig. 3 Average DOY of lilac FL generated from **a** SI- x_{LM} and **b** TT, from 1980 to 2014

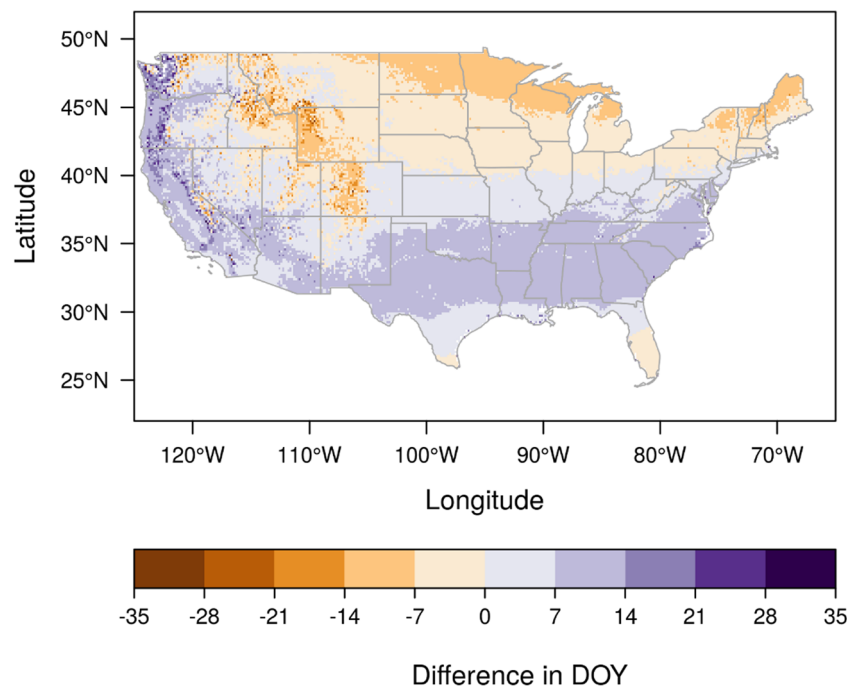
Fig. 4 Histogram of the difference between products generated from SI-x_{LM} and from TT. The differences are the pairwise subtraction of grid cell values of the TT product extracted from the SI-x_{LM} product.



example, in California and Washington states, TT estimates are about 1 month earlier. In the Rocky Mountains, TT estimates are as much as one month later. These differences can be explained by the fact that TT uses only Daymet temperature, for which the interpolation method uses elevation as a

key covariant (Daly et al. 2008). Moreover, TT considers only the long-term effects of temperature while SI-x_{LM} considers both the short-term and long-term effects and the day length. Although these exploratory results tend to show that the performance of models varies in both space and time, further

Fig. 5 Map of the difference between products generated from SI-x_{LM} and from TT. The differences are the pairwise subtraction of grid cell values of the TT product resulting from the SI-x_{LM} product



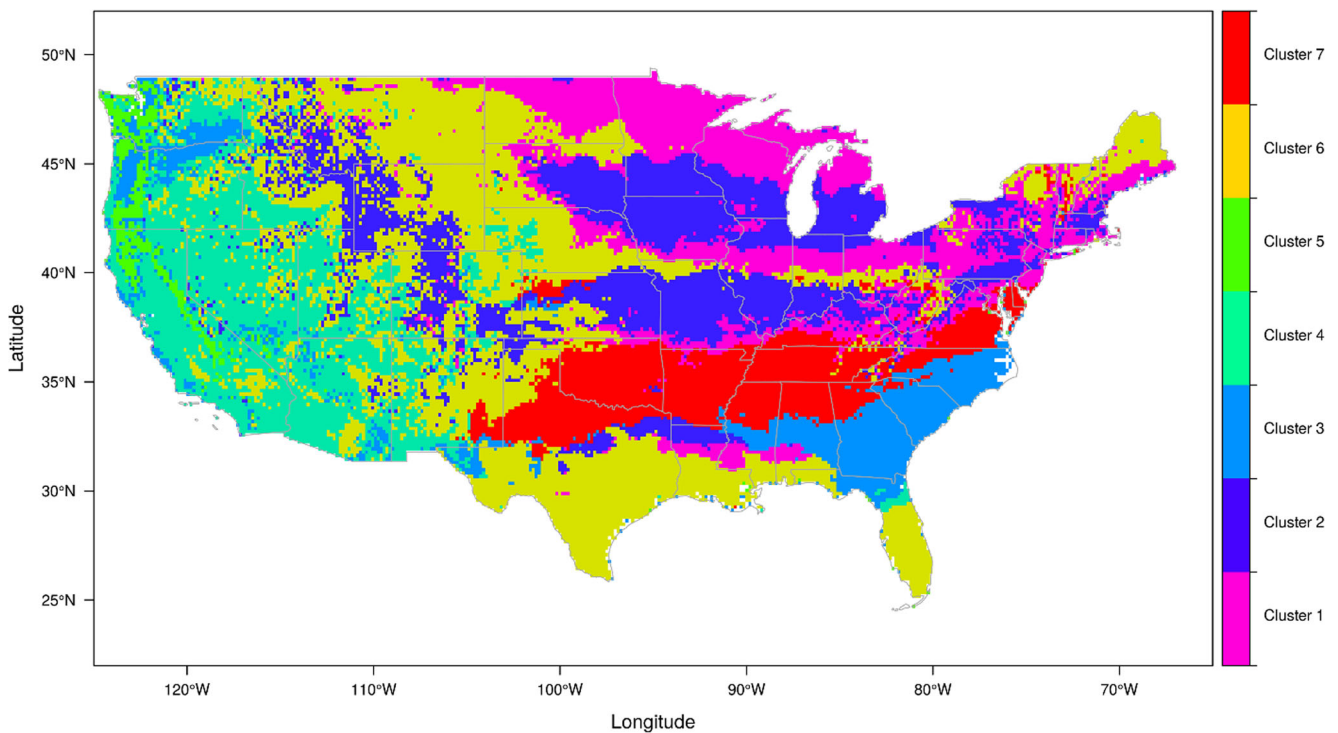


Fig. 6 Clustered regions for which the difference between estimates of $SI-x_{LM}$ and TT remains similar over the years

metrics need to be developed to quantify the significant dissimilarity of spatial and temporal variations that are derived from models.

Spatio-temporal clustering of annual difference in DOY between $SI-x_{LM}$ and TT resulted in grouping of seven regions that have a similar difference over space and time (Fig. 6). The variability of the cluster type is larger in the eastern CONUS than in the western CONUS, which shows that $SI-x_{LM}$ and TT perform substantially different in the eastern CONUS. In the eastern CONUS, the elevation gradient and consequently the temperature gradient are lower than in the western CONUS. Thus, compared with the western CONUS, the importance of day length that changes with latitude is higher in the eastern CONUS. As a result, the variability of cluster types is higher

in the eastern CONUS, and the clusters have a latitudinal pattern. This is because TT is only based on temperature while $SI-x_{LM}$ takes the influence of both temperature and day length into account.

The regression line fitted to the annual outputs of $SI-x_{LM}$ and TT, helps to explore spatial variations in the temporal trend (Fig. 7). The slope of the regression lines was mapped by generalizing them in 0.7 increments, which indicates about one week's change per decade. Both models show advancement in DOY in the most western CONUS, ranging from 1 day to 1 month from 1980 to 2014. Temporal trends (Fig. 7a) driven from $SI-x_{LM}$ show both advancements and delay in the eastern CONUS while TT driven trends show mostly delay in this part of the CONUS. The geographic pattern of the

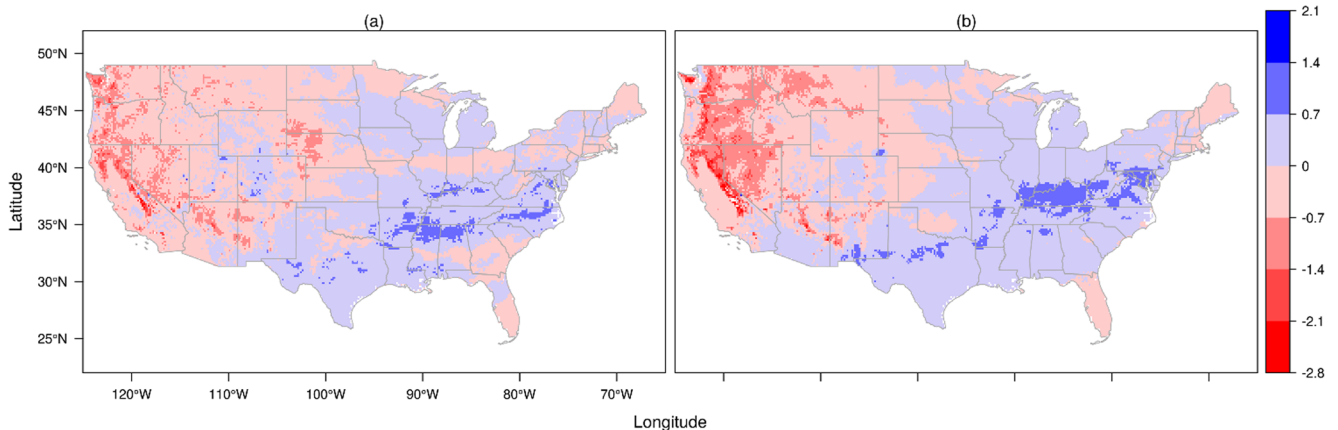


Fig. 7 Trend maps of DOY of lilac FL from **a** $SI-x_{LM}$ and **b** TT. The trend values show the rate of change of DOY per year

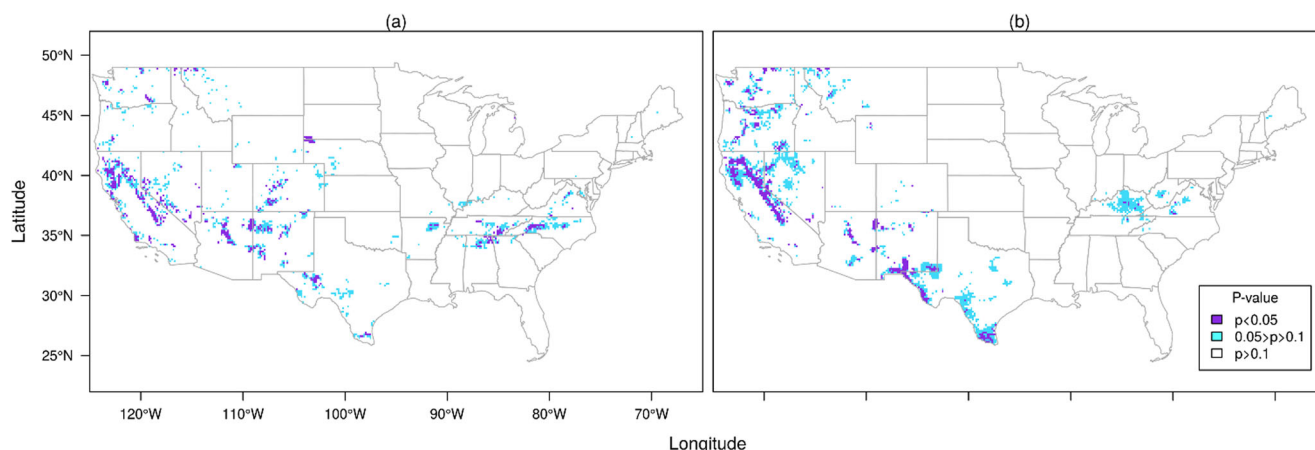


Fig. 8 Statistical significance (range of p values) for the trends in DOY of lilac FL from **a** SI- x_{LM} **b** TT, shown in Fig. 7

delay in DOY is consistent with the well-documented ‘warming hole’ in the southeastern USA (Robinson 2002; Kunkel et al. 2006; Meehl et al. 2012; Schwartz et al. 2013).

There is a significant trend variation between the outputs from SI- x_{LM} and TT (Fig. 8). TT shows larger areas with a significant trend in the west coast and south central CONUS compared with SI- x_{LM} . For example, TT shows significant advancement for most locations in Oregon, while SI- x_{LM} does not show such advancement in that state. Or, TT estimates significant delay in the southern and western part of Texas. The explanation for this is that TT uses average temperature for which annual variation is higher in the abovementioned regions. In the eastern part of the USA, significant trends between SI- x_{LM} and TT do not match. Outputs from these

models show completely different spatial patterns in estimated delay over the period 2000–2014. However, both models highlight areas in the south eastern CONUS, close to the ‘warming hole’ region, where the secular trend during the past century has been towards later DOY (Meehl et al. 2012; Schwartz et al. 2013).

The histogram of difference between the trends in SI- x_{LM} and TT estimates can be up to a week per decade (Figs. 9 and 10). Difference values are the pairwise subtraction of grid cells values of the trend in SI- x estimated from the trend in TT estimates. The positive differences are more dominant in the western CONUS where the SI- x trend product shows less advancement in DOY compared with the SI- x trends. However, the negative difference values are more concentrated in the eastern CONUS where trends estimated from TT products often show a delay compared with estimated trends from SI- x .

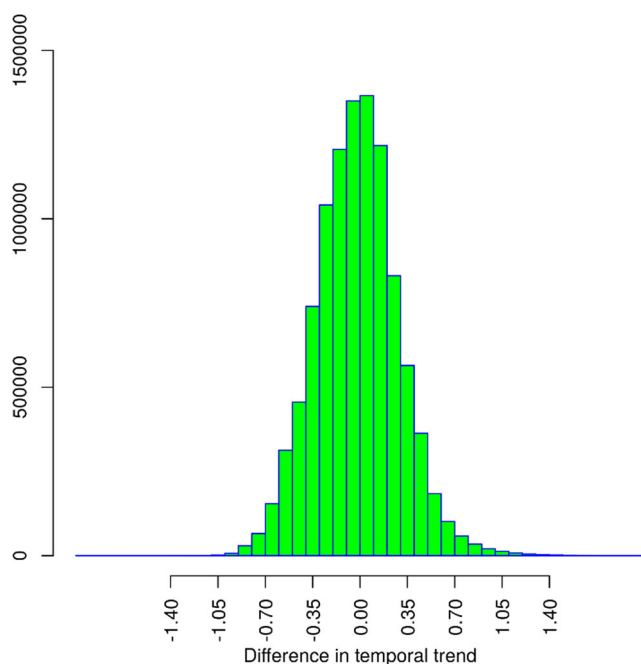


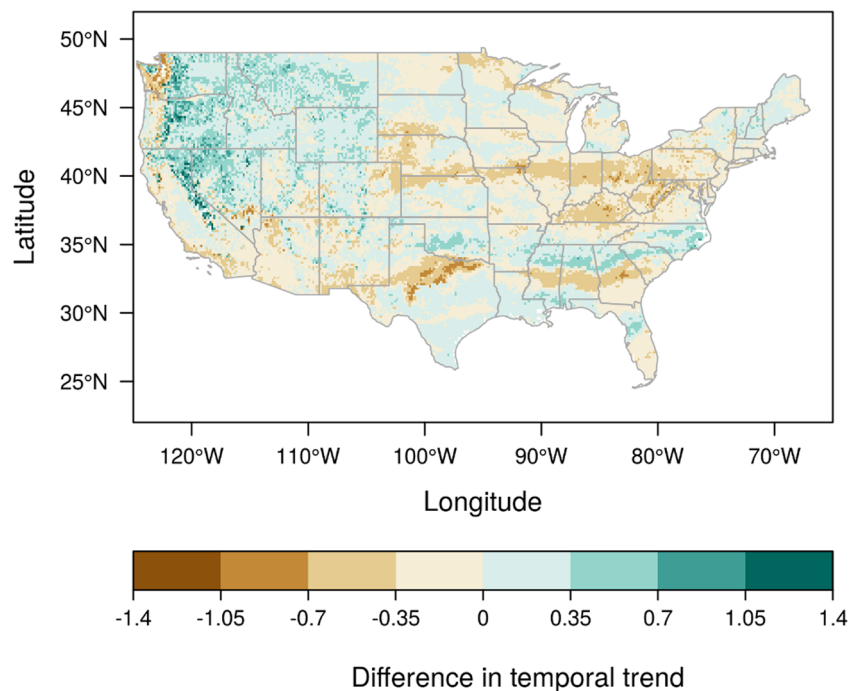
Fig. 9 Histogram of the difference between the trends in SI- x_{LM} and TT estimates. The differences are the subtraction of the trend in SI- x estimates from the trends in TT estimates

Conclusions

The analysis of spatial patterns and temporal trends in phenological model outputs is a necessary step in validating them, and to obtain more reliable model predictions that can be used to study climate change and to design management and adaptation strategies for ecological and agricultural systems. This paper not only analyses patterns and trends from observational sites but also analyses spatio-temporal patterns for the complete spatial conterminous US and at a very fine spatial resolution (1 km). Due to the large volume of data, our workflow uses cloud computing and gridded temperate time-series to study phenology at continental scales. Volunteered phenological observations are also used to compare and validate the average and rate of change of DOY across CONUS.

Our results show that errors produced by running SI- x_{LM} and TT models are similar, and that these models are 2 days more accurate than those provided by other spring phenology

Fig. 10 Map of the difference between the trends in SI-x_{LM} and TT estimates. The differences are the subtraction of the trend in SI-x estimates from the trends in TT estimates



models considered in this study. However, patterns of DOY which are derived from SI-x_{LM} and TT models are 11 days different in the CONUS. Given that the period under consideration only contains 15 years, this difference is considerable. This demonstrates the importance of spatial-temporal patterns in the model assessment. The spatial variability of the SI-x_{LM} and TT models is higher for the eastern CONUS. However, further correction and tests such as autocorrelation correction and bootstrapping are needed to analyse and compare the effect of the individual models statistically. Further, a longer period of data needs to be analysed to explore field significance further, which evaluate the performance of each model individually. The results also indicate that the estimated rate of change in DOY from SI-x_{LM} and TT can be up to 1 week per decade different across the CONUS. Moreover, our results show that the significance of the rate of change from SI-x and TT is spatially variable in the CONUS.

Therefore, current approaches for validating phenological models based on global statistics such as RMSE cannot be used to gain information about the variability of patterns and trends in different regions. Studies using phenological models and gridded input data to study climate change impact on plant seasonality, and the eventual consequences on other living organisms should check both the spatial and temporal variability at large-scale. Using a model that is found less valid across the study area than another model (i.e. with a “worse” RMSE) may still provide more realistic patterns and trends when compared with the use of large-scale phenological data and/or information. Hence, we recommend applying our workflow to check the reliability of phenological models calibrated for large scale applications because it will help guard against the

selection of overfitted models. The workflow presented in this paper can be applied to other phenophases, species and models to explore spatial phenological patterns and trends, and to better understand the impact of climate change on the Earth. It also highlights the need to develop metrics that can quantify the difference between patterns and trends derived from phenological models driven by gridded datasets.

Acknowledgements We thank volunteers and Google Earth Engine for the data availability and computation power. We thank Prof. Mark D Schwartz for providing historical datasets used to calibrate SI-x_{LM}.

Funding information This research was supported in part by a Google Faculty Research Award to Prof. Raul Zurita-Milla. Dr. Emma Izquierdo-Verdiguier is supported by the APOSTD/2017/099 Generalitat Valenciana grant (Spain).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aarts E, Korst J (1988) Simulated annealing and Boltzmann machines. John Wiley and Sons Inc., New York
- Abatzoglou JT (2013) Development of gridded surface meteorological data for ecological applications and modelling. *Int J Climatol* 33(1): 121–131. <https://doi.org/10.1002/joc.3413>
- Allstadt AJ, Vavrus SJ, Heglund PJ, Pidgeon AM, Thogmartin WE, Radeloff VC (2015) Spring plant phenology and false springs in

- the conterminous US during the 21st century. *Environ Res Lett* 10(10):104008. <https://doi.org/10.1088/1748-9326/10/10/104008>
- Arthur D, Vassilvitskii S (2007) K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 1027–1035
- Ault TR, Macalady AK, Pederson GT, Betancourt JL, Schwartz MD (2011) Northern Hemisphere modes of variability and the timing of spring in western North America. *J Clim* 24(15):4003–4014
- Ault TR, Schwartz MD, Zurita-Milla R, Weltzin JF, Betancourt JL, Ault TR et al (2015a) Trends and natural variability of spring onset in the coterminous United States as evaluated by a new gridded dataset of spring indices. *J Clim* 28(21):8363–8378. <https://doi.org/10.1175/JCLI-D-14-00736.1>
- Ault TR, Zurita-Milla R, Schwartz MD (2015b) A Matlab© toolbox for calculating spring indices from daily meteorological data. *Comput Geosci* 83:46–53. <https://doi.org/10.1016/J.CAGEO.2015.06.015>
- Badeck F-W, Bondeau A, Bottcher K, Doktor D, Lucht W, Schaber J, Sitch S (2004) Responses of spring phenology to climate change. *New Phytol* 162(2):295–309. <https://doi.org/10.1111/j.1469-8137.2004.01059.x>
- Basler D (2016) Evaluating phenological models for the prediction of leaf-out dates in six temperate tree species across central Europe. *Agric For Meteorol* 217:10–21. <https://doi.org/10.1016/j.agrformet.2015.11.007>
- Belmecheri S, Babst F, Hudson AR, Betancourt J, Trouet V, Belmecheri S et al (2017) Northern hemisphere jet stream position indices as diagnostic tools for climate and ecosystem dynamics. *Earth Interact* 21(8):1–23. <https://doi.org/10.1175/EI-D-16-0023.1>
- Briske DD, Joyce LA, Polley HW, Brown JR, Wolter K, Morgan JA et al (2015) Climate-change adaptation on rangelands: linking regional exposure with diverse adaptive capacity. *Front Ecol Environ* 13(5):249–256. <https://doi.org/10.1890/140266>
- Brunsdon C, Comber A (2012) Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach. *Geoinformatica* 16(4):675–690. <https://doi.org/10.1007/s10707-012-0159-6>
- Burghardt LT, Metcalf CJE, Wilczek AM, Schmitt J, Donohue K (2015) Modeling the influence of genetic and environmental variation on the expression of plant life cycles across landscapes. *Am Nat* 185(2):212–227. <https://doi.org/10.1086/679439>
- Cannell MGR, Smith RI (1983) Thermal time, chill days and prediction of budburst in *Picea sitchensis*. *J Appl Ecol* 20(3):951. <https://doi.org/10.2307/2403139>
- Caprio JM (1974) The solar thermal unit concept in problems related to plant development and potential evapotranspiration. Springer, Berlin, pp 353–364. https://doi.org/10.1007/978-3-642-51863-8_29
- Caprio JM (1993) Western regional phenological summary of information on honeysuckle and Lilac first bloom phase covering the period 1956–1991. *Mont. Agric. Exp. Stn. State Clim. Cent. Circ., No. 3*, 92 pp
- Cayan DR, Dettinger MD, Kammerdiener SA, Caprio JM, Peterson DH (2001) Changes in the onset of spring in the Western United States. *Bull Am Meteorol Soc* 82(3):399–415. [https://doi.org/10.1175/1520-0477\(2001\)082<0399:CITOOS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<0399:CITOOS>2.3.CO;2)
- Chmielewski F-M (2013) Phenology in agriculture and horticulture. In: Schwartz MD (ed) *Phenology: An Integrative Environmental Science*. Springer Netherlands, Dordrecht, pp 539–561. https://doi.org/10.1007/978-94-007-6925-0_29
- Chuine I, Cambon G, Comtois P (2000) Scaling phenology from the local to the regional level: advances from species-specific phenological models. *Glob Chang Biol* 6(8):943–952. <https://doi.org/10.1046/j.1365-2486.2000.00368.x>
- Chuine I, Yiou P, Viovy N, Seguin B, Daux V, Ladurie ELR (2004) Grape ripening as a past climate indicator. *Nature* 432(7015):289–290. <https://doi.org/10.1038/432289a>
- Chuine I, Bonhomme M, Legave J-M, García de Cortázar-Atauri I, Charrier G, Lacoïnte A, Améglio T (2016) Can phenological models predict tree phenology accurately in the future? The unrevealed hurdle of endodormancy break. *Glob Chang Biol* 22(10):3444–3460. <https://doi.org/10.1111/gcb.13383>
- Črepinšek Z, Kajfež-Bogataj L, Bergant K (2006) Modelling of weather variability effect on fitophenology. *Ecol Model* 194(1–3):256–265. <https://doi.org/10.1016/J.ECOLMODEL.2005.10.020>
- Crimmins A, Kolian M, Bacanskas L, Rosseel K (2016) *Climate Change Indicators in the United States 2016* (fourth edition). <https://doi.org/10.13140/RG.2.2.30480.20487>
- Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH et al (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int J Climatol* 28(15):2031–2064. <https://doi.org/10.1002/joc.1688>
- Doi H, Katano I (2008) Phenological timings of leaf budburst with climate change in Japan. *Agric For Meteorol* 148(3):512–516. <https://doi.org/10.1016/j.agrformet.2007.10.002>
- Ellwood ER, Temple SA, Primack RB, Bradley NL, Davis CC (2013) Record-breaking early flowering in the Eastern United States. *PLoS One* 8(1):e53788. <https://doi.org/10.1371/journal.pone.0053788>
- Enquist CAF, Kellermann JL, Gerst KL, Miller-Rushing AJ (2014) Phenology research for natural resource management in the United States. *Int J Biometeorol* 58(4):579–589. <https://doi.org/10.1007/s00484-013-0772-6>
- García de Cortázar-Atauri I, Brisson N, Gaudillere JP (2009) Performance of several models for predicting budburst date of grapevine (*Vitis vinifera* L.). *Int J Biometeorol* 53(4):317–326. <https://doi.org/10.1007/s00484-009-0217-4>
- Gerst KL, Kellermann JL, Enquist CAF, Rosemartin AH, Denny EG (2016) Estimating the onset of spring from a complex phenology database: trade-offs across geographic scales. *Int J Biometeorol* 60(3):391–400. <https://doi.org/10.1007/s00484-015-1036-4>
- Gordo O, Sanz JJ (2010) Impact of climate change on plant phenology in Mediterranean ecosystems. *Glob Chang Biol* 16(3):1082–1106. <https://doi.org/10.1111/j.1365-2486.2009.02084.x>
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens Environ* 202:18–27. <https://doi.org/10.1016/J.RSE.2017.06.031>
- Guo W, Gong J, Jiang W, Liu Y, She B (2010) OpenRS-Cloud: a remote sensing image processing platform based on cloud computing environment. *Sci China Technol Sci* 53(S1):221–230. <https://doi.org/10.1007/s11431-010-3234-y>
- Hufkens K, Basler D, Milliman T, Melaas EK, Richardson AD (2018) An integrated phenology modelling framework in R. *Methods Ecol Evol* 9(5):1276–1285. <https://doi.org/10.1111/2041-210X.12970>
- Hunter AF, Lechowicz MJ (1992) Predicting the timing of budburst in temperate trees. *J Appl Ecol* 29(3):597. <https://doi.org/10.2307/2404467>
- Izquierdo-Verdiguier E, Zurita-Milla R, Ault TR, Schwartz MD (2018) Development and analysis of spring plant phenology products: 36 years of 1-km grids over the conterminous US. *Agric For Meteorol* 262:34–41. <https://doi.org/10.1016/J.AGRFORMET.2018.06.028>
- Janssen PHM, Heuberger PSC (1995) Calibration of process-oriented models. *Ecol Model* 83(1–2):55–66. [https://doi.org/10.1016/0304-3800\(95\)00084-9](https://doi.org/10.1016/0304-3800(95)00084-9)
- Kunkel KE, Liang X-Z, Zhu J, Lin Y, Kunkel KE, Liang X-Z et al (2006) Can CGCMs simulate the twentieth-century “Warming Hole” in the Central United States? *J Clim* 19(17):4137–4153. <https://doi.org/10.1175/JCLI3848.1>
- Labe Z, Ault TR, Zurita-Milla R (2017) Identifying anomalously early spring onsets in the CESM large ensemble project. *Clim Dyn* 48(11–12):3949–3966. <https://doi.org/10.1007/s00382-016-3313-2>
- Leung LR, Qian Y, Bian X, Leung LR, Qian Y, Bian X (2003) Hydroclimate of the Western United States based on observations

- and regional climate simulation of 1981–2000. Part I: Seasonal Statistics. *J Clim* 16(12):1892–1911. [https://doi.org/10.1175/1520-0442\(2003\)016<1892:HOTWUS>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<1892:HOTWUS>2.0.CO;2)
- Lieth H (1974) Purposes of a phenology book. Springer, Berlin, pp 3–19. https://doi.org/10.1007/978-3-642-51863-8_1
- Linkosalo T, Lappalainen HK, Hari P (2008) A comparison of phenological models of leaf bud burst and flowering of boreal trees using independent observations. *Tree Physiol* 28(12):1873–1882. <https://doi.org/10.1093/treephys/28.12.1873>
- Marra PP, Francis CM, Mulvihill RS, Moore FR (2005) The influence of climate on the timing and rate of spring bird migration. *Oecologia* 142(2):307–315. <https://doi.org/10.1007/s00442-004-1725-x>
- Masle J, Doussinaut G, Farquhar GD, Sun B (1989) Foliar stage in wheat correlates better to photothermal time than to thermal time. *Plant Cell Environ* 12(3):235–247. <https://doi.org/10.1111/j.1365-3040.1989.tb01938.x>
- Matzarakis A, Mayer H, Chmielewski F-M (2010) Berichte des Meteorologischen Instituts der Albert-Ludwigs-Universität Freiburg
- McCabe GJ, Ault TR, Cook BI, Betancourt JL, Schwartz MD (2012) Influences of the El Niño southern oscillation and the Pacific decadal oscillation on the timing of the North American spring. *Int J Climatol* 32(15):2301–2310. <https://doi.org/10.1002/joc.3400>
- Meehl GA, Arblaster JM, Branstator G, Meehl GA, Arblaster JM, Branstator G (2012) Mechanisms contributing to the warming hole and the consequent U.S. East–west differential of heat extremes. *J Clim* 25(18):6394–6408. <https://doi.org/10.1175/JCLI-D-11-00655.1>
- Mehdipoor H, Zurita-Milla R, Rosemartin A, Gerst KL, Weltzin JF (2015) Developing a workflow to identify inconsistencies in volunteered geographic information: a phenological case study. *PLoS One* 10(10):e0140811. <https://doi.org/10.1371/journal.pone.0140811>
- Mehdipoor H, Zurita-Milla R, Augustijn E, van Vliet A (2016) Analyzing phenological synchronicity using volunteered geographic information. Association of Geographic Information Laboratories for Europe (AGILE)
- Mehdipoor H, Izquierdo-Verdiguier E, Zurita-Milla R (2017) Continental-scale monitoring and mapping of false spring : a cloud computing solution + powerpoint
- Mehdipoor H, Zurita-Milla R, Izquierdo-Verdiguier E, Betancourt JL (2018) Influence of source and scale of gridded temperature data on modelled spring onset patterns in the conterminous United States. *Int J Climatol*. <https://doi.org/10.1002/joc.5857>
- Menzel A (2005) A 500 year pheno-climatological view on the 2003 heatwave in Europe assessed by grape harvest dates. *Meteorol Z* 14(1):75–77. <https://doi.org/10.1127/0941-2948/2005/0014-0075>
- Mitchell KE, Lohmann D, Houser PR, Wood EF, Schaake JC, Robock A et al (2004) The multi-institution North American Land Data Assimilation System (NLDAS): utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J Geophys Res* 109(D7):D07S90. <https://doi.org/10.1029/2003JD003823>
- Noguchi K, Gel YR, Duguay CR (2011) Bootstrap-based tests for trends in hydrological time series, with application to ice phenology data. *J Hydrol* 410(3–4):150–161. <https://doi.org/10.1016/J.JHYDROL.2011.09.008>
- Richardson AD, Keenan TF, Migliavacca M, Ryu Y, Sonnentag O, Toomey M (2013) Climate change, phenology, and phenological control of vegetation feedbacks to the climate system. *Agric For Meteorol* 169:156–173. <https://doi.org/10.1016/J.AGRFORMET.2012.09.012>
- Robinson WA (2002) General circulation model simulations of recent cooling in the east-central United States. *J Geophys Res* 107(D24):4748. <https://doi.org/10.1029/2001JD001577>
- Rosemartin AH, Denny EG, Weltzin JF, Lee Marsh R, Wilson BE, Mehdipoor H, Zurita-Milla R, Schwartz MD (2015) Lilac and honeysuckle phenology data 1956–2014. *Sci Data* 2:150038. <https://doi.org/10.1038/sdata.2015.38>
- Santer B (1985) The use of general circulation models in climate impact analysis: a preliminary study of the impacts of a CO₂-induced climatic change on West European agriculture. *Clim Chang* 7(1):71–93. <https://doi.org/10.1007/BF00139442>
- Schwartz MD (1997) Spring index models: an approach to connecting satellite and surface phenology. *Phenol Seas Clim I*, 23–38
- Schwartz MD (1999) Advancing to full bloom: planning phenological research for the 21st century. *Int J Biometeorol* 42(3):113–118. <https://doi.org/10.1007/s004840050093>
- Schwartz MD, Marotz GA (1988) Synoptic events and spring phenology. *Phys Geogr* 9(2):151–161. <https://doi.org/10.1080/02723646.1988.10642345>
- Schwartz MD, Ahas R, Aasa A (2006) Onset of spring starting earlier across the Northern Hemisphere. *Glob Chang Biol* 12(2):343–351. <https://doi.org/10.1111/j.1365-2486.2005.01097.x>
- Schwartz MD, Betancourt JL, Weltzin JF (2012) From Caprio’s lilacs to the USA National Phenology Network. *Front Ecol Environ* 10(6):324–327. <https://doi.org/10.1890/110281>
- Schwartz MD, Ault TR, Betancourt JL (2013) Spring onset variations and trends in the continental United States: past and regional assessment using temperature-based indices. *Int J Climatol* 33(13):2917–2922. <https://doi.org/10.1002/joc.3625>
- Thornton PE, Thornton MM, Mayer BW, Wilhelm N, Wei Y, Devarakonda R, Cook RB (2014) Daymet: daily surface weather data on a 1-km grid for North America, Version 2
- van Laarhoven PJM, Aarts EHL (1987) Simulated annealing. In: *Simulated Annealing: Theory and Applications*. Springer Netherlands, Dordrecht, pp 7–15. https://doi.org/10.1007/978-94-015-7744-1_2
- Willmott C, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res* 30(1):79–82. <https://doi.org/10.3354/cr030079>
- Willmott C, Matsuura K, Robeson SM (2009) Ambiguities inherent in sums-of-squares-based error statistics. *Atmos Environ* 43(3):749–752. <https://doi.org/10.1016/J.ATMOSENV.2008.10.005>
- Willmott C, Robeson S, Matsuura K (2017) Climate and other models may be more accurate than reported. *Eos (Washington DC)*. <https://doi.org/10.1029/2017EO074939>
- Wu X, Zurita-Milla R, Kraak M-J (2016) A novel analysis of spring phenological patterns over Europe based on co-clustering. *J Geophys Res Biogeosci* 121(6):1434–1448. <https://doi.org/10.1002/2015JG003308>
- Zavalloni C, Andresen JA, Flore JA (2006) Phenological models of flower bud stages and fruit growth of ‘Montmorency’ sour cherry based on growing degree-day accumulation. *J Am Soc Hortic Sci* 131(5):601–607
- Zurita-Milla R, Goncalves R, Izquierdo-Verdiguier E, Ostermann FO (2017) Exploring vegetation phenology at continental scales : linking temperature-based indices and land surface phenological metrics