Exploring spring onset at continental scales: mapping phenoregions and correlating temperature and satellite-based phenometrics

Raul Zurita-Milla, Romulo Goncalves, Emma Izquierdo-Verdiguier and Frank O. Ostermann

Abstract—Each spring many plants put on new leaves and/or open their flowers creating a ”green-wave” that can be tracked using phenological data. Various phenological datasets can be used to study spring onset at continental to global scales. Here we present a novel exploratory analysis where we link two multi-decadal and high-spatial resolution datasets: temperature-based phenological indices and land surface phenological metrics derived from satellite images. Our exploratory analysis, illustrated with data for the conterminous US, focuses on identifying regions with similar spring onset, and on mapping the coherence between these phenological products. Our results show that the spring onset patterns captured by the satellite are more complex than the ones identified using temperature-based phenological indices. They also highlight areas with stable and unstable spring onsets (i.e. areas that tend to remain or change of phenoregion from year to year). Finally, our resutls reveal that temperature-based indices are both positively and negatively correlated with the phenological information that can be derived from satellites. This opens the door to the definition of rules to integrate multi-source phenological data. To cope with the computational challenges of analyzing big geospatial rasters, we executed our analysis on a cloud platform running Apache Spark and various of its extensions (e.g. Geotrellis, SparkMLlib). This platform performed well and allowed the execution of user-tailored analyses. Hence, we believe that our computational platform paves the path towards the efficient analysis of global vegetation phenology at very high spatial resolution and, more generally, to the analysis of the ever-increasing collections of geospatial data about our planet.

Index Terms—Extended spring indices, land surface phenology, exploratory data analysis, big geo-data, Apache Spark.

1 Introduction

EACH spring many plants begin to grow, put on new leaves, and/or open their flowers creating what is known as the ”green-wave” [1]. This wave varies from place to place and from year to year because its form and dynamics are strongly influenced by environmental conditions [2]. Phenology is the science that studies the timings of such recurring biological events, as well as their causes, interrelations, and variations in space and time [3].

With a rapidly warming world (NASA just confirmed that 2017 was the second warmest year since global estimates became feasible in 1880, and that “17 of the 18 warmest years have occurred since 2001”[1]), understanding spring onset variability is critical to quantify the impact of climate change on our planet [2], [4], [5]. Moreover, spring onset variability also affects agricultural productivity, the water, carbon, and energy cycles, and many ecological processes like bird migration and insect emergence [6], [7]. Hence, it is not surprising that several (inter-)national initiatives, programs, and projects actively support phenological data collection.

Ground observations constitute an important source of phenological data [8], especially since the popularization of citizen science phenological networks [9]. These networks not only educate the general public but also have lowered the barriers for volunteers to collect and share geo-referenced phenological observations. Although ground phenological observations continue to directly support a variety of ecological studies [8], [10], [11], they are not ideal for continental scale studies because they tend to be sparse and discontinuous (i.e. they do not have the required temporal and spatial coverage). This is recognized by phenologists who often combine ground observations with weather data to create phenological models [8]. These models “spatialize” and generalize the available observations, and this allows continental to global scale analyses. However, phenological models cannot fully capture all the phenological changes that are taking place on our planet. Fortunately, Earth observation sensors can also be used to monitor phenology and phenological changes at regional, national, continental, and global scales [12], [13], [14], [15]. Time series of satellite images, typically transformed into vegetation indices that highlight vegetation signal, are used to map what is known as land surface phenology [13].

Unfortunately, few studies have worked on bridging the (semantic) gap between the information that can be derived from phenological models and from Earth observation sensors. As a result, the general applicability of both data sources as well as their deeper ecological meaning remains poorly understood. Because of this, the seamless integration and analysis of multi-source phenological data remains challenging [15], [16]. The main objective of this pa-

- *R. Zurita-Milla and F. O. Ostermann are with the Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, the Netherlands.E-mails: {r.zurita-milla, f.o.ostermann}@utwente.nl*
- *R. Goncalves is with the Netherlands eScience Center (NLeSC), the Netherlands. E-mail: r.Goncalves@esciencecenter.nl*
- *E. Izquierdo-Verdiguier is with the Institute of Surveying, Remote Sensing and Land Information (IVFL), University of Natural Resources and Life Sciences (BOKU), Vienna, Austria. Email: emma.izquierdo@boku.ac.at*

[1]https://twitter.com/i/web/status/954013611395112960

per is to address this challenge by analyzing the coherence of spring onset metrics derived from a temperature-driven phenological model and from satellite images. For this, we first look at the similarity in spring onset patterns by mapping regions with similar phenology (i.e. phenoregions). Then, we investigate whether the timings of spring onset predicted by spring onset models and observed by satellites are statistically dependent. This is done by analyzing their correlation and investigating whether predictive relationships could be modeled to integrate these two phenological products.

1.1 Related work

Mapping spring onset phenoregions requires forming groups of grid cells that are more similar among themselves than to any of the cells assigned to other groups. From a methodological point of view, this task belongs to the domain of exploratory data mining and typically relies on unsupervised classification or clustering methods. The first phenoregion study dates back to 2005 when White et al. [14] used the well-known K-means algorithm to cluster 17 years of Normalized Difference Vegetation Index (NDVI) data from the Advanced Very High Resolution Radiometer (AVHRR) sensor. This 8 km global satellite data was complemented with monthly temperature and precipitation climatologies before performing the clustering to define optimal regions for phenological monitoring. The same algorithm was also used a few years later to cluster higher spatial resolution NDVI time series acquired by the Moderate-resolution imaging spectroradiometer (MODIS). The goal of that analysis was to build an early warning system for US forests based on phenological information [17], [18]. Other authors have used other clustering algorithms, e.g. Gu et al. [19] used the ISODATA algorithm to cluster phenological metrics derived from MODIS NDVI composites. Their results were used to create a new regionalization of the US, which in turn contributed to a new land cover map. Silva et al. [20] attempted to discover regions with similar phenology in the Amazon forests with the spectral angle mapper based on AVHRR NDVI data, in order to study the environmental drivers behind the phenoregions. Zhang et al. [21], [22] created phenoregions in the state of Colorado (US) with K-means++, a specialized form of the K-means algorithm that provides better initialization seeds. They used both AVHRR and MODIS NDVI time series and climatological information. Finally, Wu et al. [23] used an information theory co-clustering algorithm to map European spatio-temporal regions with similar spring onsets. However, that work did not use any satellite imagery but long-term simulations of spring onset at a relatively coarse spatial resolution (25 km).

Given the proliferation of phenological data and the lack of a universal approach to measure spring onset, more work is needed to evaluate existing phenological products. A good example of this research direction are White at al. [24], who compared 10 methods to derive spring phenology from the same satellite data (15-day composites of 8 km AVHRR NDVI images). Their results indicate that the ecological meaning and driving causes of land surface phenometrics are unclear because they found differences in the predicted start of the vegetation season of up to 60 days. Other authors have focused on the impact of scale (and indirectly of mixed pixels) on the values of land surface phenometrics [25], and many authors employ ground phenological observations to validate and/or benchmark their phenological products [26], [27]. Yet only few studies have compared phenological models with land surface phenological metrics. Duchemin et al. [28] found that the budburst timing predicted from air temperatures using the thermal time model correlates well with the AVHRR-based budburst predictions over monospecies forests in France. Schwartz et al. [29], [30] compared the original spring indices (c.f. section 3) against two satellite-derived start of season metrics in the Eastern US. Although the results of these studies were encouraging for connecting surface and sensor phenometrics, the analysis was limited to 10 by 10 km windows around weather stations (possibly because of the lack of high spatial resolution phenological products). Finally, in [30], the original spring indices were again compared with satellite derived phenometrics using multiple methods to estimate the start of the season from MODIS data, but the analysis was limited to just six phenological sites in the Eastern US. Results indicate that the original spring indices better match the phenology observed on the ground than the relatively coarse phenological metrics that could be derived from MODIS.

1.2 Goals, contribution and outlook

Our work aims to take the previously mentioned phenological studies to the next level by presenting a novel exploratory analysis of spring onset models and of land surface phenological metrics, which constitute two of the most important sources of spatio-temporal phenological data. Our analysis is performed at a continental scale (the conterminous 48 US states), and is based on the longest possible time series at the highest possible spatial resolution. In particular, our analysis focuses on mapping long-term and annual spring onset phenoregions, and on studying the coherence and statistical predictability of multi-source phenological products.

To realize this analysis, we designed and implemented an open-source, distributed, and cloud-based computational platform that allows storing, analyzing, and visualizing large collections of geospatial rasters. This platform represents another major contribution of our work. Given the lack of well-tested open, distributed, and cloud-enabled solutions in the domain of big geo-data, a secondary aim of our work is to evaluate its potential to analyze big geospatial rasters in both local and cloud-based environments. We also think that the proposed platform makes our work future-proof because it can deal with the new generation of very high spatial resolution phenological datasets that are currently being produced by, for instance, the Sentinel-1 and 2 missions [31].

The rest of this paper is organized as follows: In section 2, we introduce the philosophy, design principles, and characteristics of our computational solution. Then, in section 3, we describe the phenological datasets selected for this study and our approach to deal with large geospatial datasets. In section 4, we present our first case study on mapping

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2019.2926292, IEEE Transactions on Big Data

phenoregions over the 48 conterminous US states. This is followed by our second case study, in section 5, where we study the correlation between temperature- and satellite-based phenometrics. Finally, section 6 contains a summary of our main findings, and presents follow-up research ideas.

2 Computational Solution

Analyzing long-term and high spatial resolution phenological products at continental scales is a computationally demanding task. Hence, we designed and implemented an open-source and cloud-enabled solution relying on Apache Spark [32] and various of its extensions, like the scalable machine learning library MLlib [33].

This platform allows conducting the case studies presented in sections 4 and 5 in a transparent and cloud-based infra-structure. Moreover, our computational solution is designed for easy user interaction and scalability, supporting both (simple) data exploration and massive data analysis. User interaction is realized through Jupyter notebooks that push the computations to a remote cluster. Scalability is ensured by making sure that all computations are based on distributed data structures and Spark internals designed for efficient data processing.

2.1 The platform's architecture

Our computational solution remains application-independent and is designed to store the data in well-known file formats like GeoTIFFs and Hierarchical Data Format (HDF). The platform's architecture is organized into three layers: a storage layer, a processing layer, and JupyterHub services for user-interaction (Figure 1). The storage layer offers two flavors of storage: file-based by Hadoop Distributed File System (HDFS), and object-based by Amazon S3 service. For local environments we use Minio [34], an open-source object storage server with an Amazon S3 compatible API, to avoid modifying it when moving to a cloud provider. HDFS is used by Apache Spark [32] to exploit data locality and to store intermediates to avoid re-computations. In our case, the object storage is used to store the phenological products.

For the data analysis the user expresses the operations in Jupyter notebooks as Python, R, or Scala code. Hence, with a browser and remote connection, the user can express a research question and/or collect insights from large data sets. Computations flow from the Jupyter notebooks to the processing layer through Spark jobs (Figure 1 orange arrow), data is then retrieved from the storage layer for in-memory distributed processing, and results are fetched back for data visualization (Figure 1 green arrow).

At the processing layer we have Spark with its machine learning library SparkMLlib [33], [35] and GeoTrellis [36] for high-performance raster data processing. With GeoTrellis, GeoTIFF files are read directly from the S3 storage into Resilient Distributed Datasets (RDDs). With the phenology data products loaded as RDDs, we exploit Spark's internals for distributed data processing. One example is the mapping of phenoregions presented in Section 4.

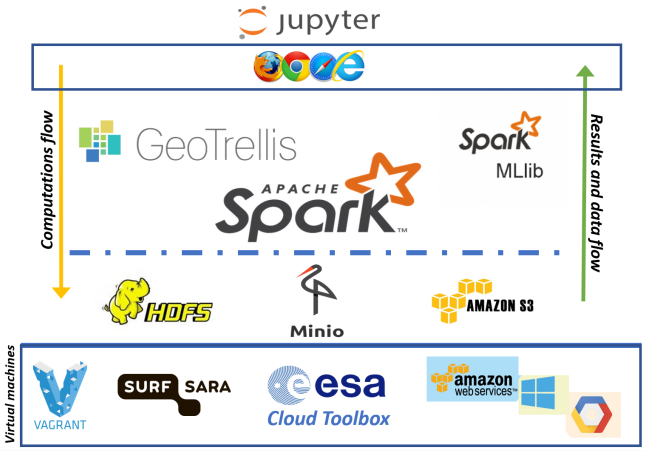

Fig. 1: Computational platform

2.2 Platform's deployment

The actual deployment and management of our computation solution is done through Emma [37], an open-source project to create a platform for development of applications for Spark and DockerSwarm clusters. The platform runs on an infra-structure composed of virtual machines that must be reachable by SSH. The machines are either cloud virtual machines or Vagrant [38] machines (see Figure 1). The latter tool allows the platform to be simulated on a local machine, i.e. in a local development environment.

Once the machines are prepared, the servers are provisioned using Ansible, an automation tool for IT infra-structure. Ansible [39] playbooks are used to create a storage layer, processing layer, and JupyterHub [40] services. With Ansible we are able to deploy a platform with the same features at different locations, such as local clusters, national infra-structure, or even a commercial cloud provider. Such a feature allows us to have tool-provenance for easy repeatability of experiments between scientists.

2.3 Provenance

Precise data citation and provenance are essential for the reproducibility of any scientific research. Earth science data are often the result of applying complex data transformation and analysis work flows to large amounts of data. Provenance information is thus essential to ensure the reproducibility of the results [41], [42].

The process to obtain provenance information should be simple and lead to a clear and human readable description. In case of having an incomplete piece of information, the user should be able to complete it. Hence, our goal is to collect sufficient information to reproduce data close enough to come to the same conclusion, but also to be able to describe the difference between two processing scenarios if there is a difference in the result. In that case, we should be able to initiate an analysis for determining the nature of the difference.

Our analysis is implemented in Jupyter notebooks. These notebooks are not only used to share results among scientists but also as a provenance and reproducibility mechanism. Furthermore, the cluster configuration, cloud setup,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2019.2926292, IEEE Transactions on Big Data

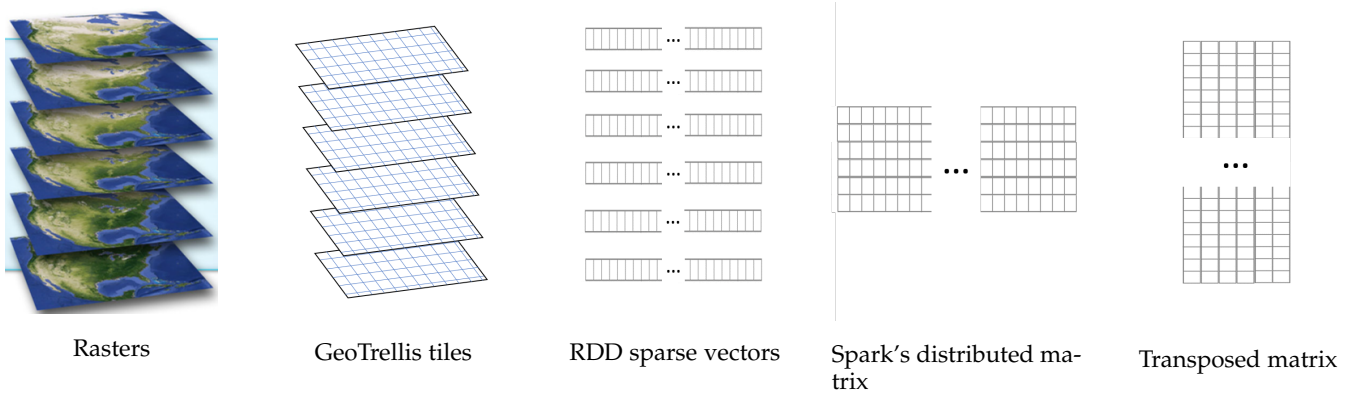

Fig. 2: From GeoTIFFs to a Spark's distributed matrix.

and platform status are traceable, and virtualization facilitates repetition and deployment of the experiment elsewhere. To assist with its deployment, users should use Emma's Ansible notebooks and roles, because they contain a detailed description of the platform, i.e., libraries' names and versions.

2.4 Platform in action

On our platform, computations are not only pushed down for remote processing, but they are also designed to exploit Spark's computational features. To achieve that, data is always loaded into memory-based data structures such as RDDs, DataFrames, and distributed matrices. With the data loaded into these structures, distributed task scheduling and fault-tolerance is then handled by Spark. Such a strategy is crucial to achieve efficiency and scalability. It also releases the user from the burden of re-writing an application in case the problem size increases, e.g. when using higher resolution data from the Copernicus Sentinel-2 mission, or for changes in the amount of available resources when moving to a different cloud-infrastructure. The decision about which structure to use and a study on the impact of different resource allocations, i.e. a detailed performance profile, is out of the scope of this paper.

The datasets (c.f. Section 3) chosen to illustrate our phenological case studies are stored as GeoTiffs. These GeoTiffs are either loaded into HDFS or to the S3 Storage. We use GeoTrellis to read each GeoTiff band as a Tile, i.e. a grid of points. For distributed processing, GeoTrellis reads time-series as an RDD of Tiles. To identify phenoregions and to study the spatio-temporal correlation between phenological time-series, each Tile is converted to a sparse vector by flattening it row-wise. The sparse vectors are used to create distributed matrices for linear algebra transformations such as Transpose. All these steps are summarized in Figure 2. Section 3 contains more details about data preparation.

Besides data preparation, Spark's internal functionality is also useful for efficient data analysis. Spark-MLlib version 2.1.1 provides a rich set of methods for data analysis ranging from basic statistics to dimensionality reduction. In Section 4.2, we show how Spark-MLlib was used for an iterative and stepwise clustering analysis. Nevertheless, it is important to mention that RDDs give an enormous flexibility to the user, and allow using analysis methods not yet covered by Spark-MLlib. In Section 5.2, we show how easy it is to introduce user-tailored data analyses.

3 Phenological Data

In this section we describe the temperature-based phenological indices (3.1) and the satellite-derived phenometric (3.2) chosen for this study. Finally, in section (3.3) we explain our approach to iteratively analyze big geospatial rasters.

3.1 The Extended Spring Indices

The Extended Spring Indices (SI-x; [43]) are a suite of mathematical models that transform daily minimum and maximum temperatures into consistent phenological metrics that indicate spring onset. The SI-x can be used to better understand and manage climate change impacts, as demonstrated by their inclusion in the list of official US global change indicators [2]. These phenological models are called "extended" because they were derived from the "original" spring indices ([5]) after lifting the "chilling" requirements, so that the Leaf and Bloom indices could be calculated across broad geographic areas [43].

From a mathematical point of view, the SI-x models were created by combining volunteered observations of leafing and blooming for three key indicator species (two varieties of honeysuckle and one of cloned lilacs [9]) with daily weather station data. With this data, a linear regression model was built to predict the day of the year (DOY) of first leaf and of first bloom for each of these plant species by using a set of regressors based on short and long term accumulations of growing degree days and changes in length of day, calculated from the latitude of the weather station. Next, the Leaf and Bloom indices are calculated as the averages of the corresponding predictions for each of the plants. From a practical point of view, the Leaf index indicates an early spring onset typically linked to the activation of grasses and shrubs. The Bloom index, which occurs a few days or weeks after the Leaf index, indicates a later phase of spring onset, associated with the moment of leafing out of deciduous trees [26]. For additional details

[2]http://www.globalchange.gov/explore/indicators

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2019.2926292, IEEE Transactions on Big Data

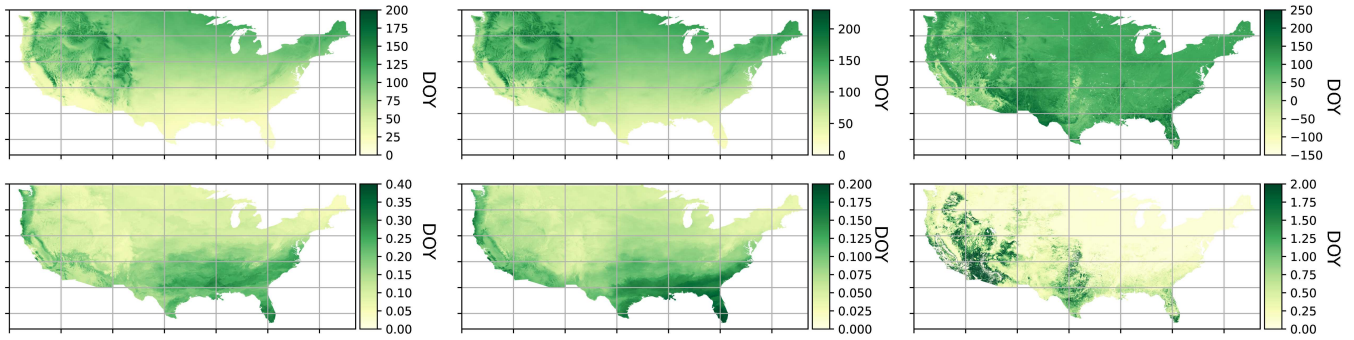

Fig. 3: Top: Average of Leaf and Bloom indices and AVHRR Start of Season (SOS) maps of conterminous US from 1989 to 2014. Bottom: The coefficient of variation of Leaf and Bloom indices and AVHRR SOS maps of conterminous US from 1989 to 2014.

including the exact model definition, the coefficients of each model, and the computer code to calculate them, see [44]. The SI-x models have been extensively used to track spring onset at specific locations by using temperature data from weather stations [43], [44], and at continental scales by using gridded temperature data [26], [27].

In this study we use a high spatial resolution (1 km) and long-term (1980 to 2015) version of the Leaf and Bloom indices, which was recently generated for the conterminous US by adapting the SI-x models to a cloud computing environment [27], [45].

3.2 Start of season

Time series of remotely sensed data can be used to derive various land surface phenological metrics that indicate phenological changes in vegetated canopies. One of these metrics is the so-called Start of Season (SOS), which indicates the beginning of photosynthetic activity. Several SOS products exist in literature, often linked to a particular sensor or study [24], [30]. Here we use a SOS product specifically made for the US by the Earth Resources Observation and Science (EROS) Center of the US Geological Survey. This product, available since 1989, is made from a special 1 km AVHRR dataset created from Local Area Cover (LAC) images provided by the AVHRR sensors aboard various National Oceanic and Atmospheric Administration (NOAA) satellites (NOAA-11, 14, 16, 17, 18, and 19). We selected this phenological product for our study because it offers the longest possible SOS time series at the best possible spatial resolution. Other AVHRR-phenological products start a few years earlier (in 1981), but at the cost of using data with a reduced spatial resolution (typically between 4 and 8 km). MODIS also offers a high spatial resolution SOS product, but the latest collection (i.e. v005) of this product is only available for the period of 2001-2012.

At the EROS center, the AVHRR LAC images were first transformed into a smooth time series of NDVI values by using the weighted least-squares regression method. This processing reduces cloud and atmospheric effects ([46]) so that a curve derivative method can be applied to predict future NDVI values based on the previous 9 observations. Finally, the SOS day is determined by identifying the DOY when the smoothed NDVI values become larger than the predicted values [47]. This SOS prediction method, also known as the delayed moving average method, indicates the occurrence of the first sustained positive change in the NDVI signal during spring [29]. Eight other phenometrics (e.g. time of maximum NDVI or end of season) are derived by the EROS center but are not used here because their timings fall outside of the spring windows, and a recent study has shown that they are negatively affected by the orbital drift of the various AVHRR sensors used to create these phenometrics [46]. The SOS product is, in fact, the least affected product and only 4-5% of the pixels are influenced by orbital drift [46]. A final limitation of the selected SOS product is that although its spatial resolution matches that of the SI-x products, SOS data is only available for the period of 1989 - 2014 [3]. Therefore, our exploratory analysis is constrained to this period.

We refer to Figure 3 for an illustration of the average Leaf and Bloom indices and of the SOS phenometric from the study period. The SI-x maps show a clearly noticeable spring gradient, with low values in the south and high DOY values in the north, although the Bloom gradient is abrupter than the Leaf gradient. This gradient is much less visible in the average SOS metric as the values are much more dependent on the actual distribution of land cover types than the SI-x, which are mostly driven by the evolution of daily temperatures. Two characteristics of the SOS product are remarkable: 1/ the SOS map contains negative values that indicate that the season started sometime during the second half of the previous calendar year, and 2/ that the product is spatially continuous, indicating that the NDVI composite method and the SOS extraction algorithm work well over cloudy areas. The coefficients of variation (CV) of the chosen phenometrics indicate that the variability of the Leaf index is twice as large as that of the Bloom index. This is not surprising because weather variability is larger at the beginning of the seasonal transition from winter to spring. The CV of the SOS product is much larger than that of the SI-x, particularly for areas in the south west which often

[3]https://lta.cr.usgs.gov/avhrr_phen

have negative SOS values.

3.3 Data challenges

For large data sets it is important to effectively reduce the data input size by filtering out all irrelevant data points for a specific analysis. In the context of our analysis, users can easily define their area of interest by applying a mask so that they can start their analysis for a single state and then shift to the conterminous US. The same applies to the spatial and/or temporal resolution of the datasets. All the analyses were designed to work at different resolutions without requiring heavy modifications of the code. This allows users to collect insights or test hypothesis in an agile fashion.

4 Mapping phenoregions

Two types of phenoregions can be obtained from our phenological time series: First, all data can be clustered along the temporal dimension to find regions (i.e. groups of grid cells) that share the same phenological trajectory across time (i.e. similar SOS values for the complete study period). Second, the data can be re-arranged to cluster regions based on their annual phenology, characterized by the annual values of the Leaf and Bloom indices. As described in the introduction, the former type of clustering is more common and leads to the identification of long-term phenological regions, and the latter clustering is normally used when grouping actual NDVI values instead of phenometrics. Here, we perform both types of clustering by using the K-means implementation available in SparkMLlib, which is a parallelized variant of the k-means++ algorithm called k-means|| [4].

First the Leaf, Bloom, and SOS time series were clustered to create long-term phenoregions. We decided to cluster each dataset independently so what we could visualize and compare the spatial patterns associated with each spring onset indicator. Then, we combined and reshaped the Leaf and Bloom datasets to create annual phenoregions so that we could visualize the stability of the temperature-based phenoregions. Both analyses were run multiple times to find the optimum k-value for each type of clustering. For the multi-temporal phenoregions, we tested values of k from 5 to 500 in steps of 10, and for the annual clusterings we tested k values from 5 to 100 in steps of 5. The optimum k-value was found at the elbow of the within-cluster sum of squared errors (WCSSE). The WCSSE metric measures the distance of each grid-cell to the centroid it has been assigned to. As such, the WCSSE monotonically decreases as k increases. However, the rate of decrease slows down after passing the optimal number of clusters.

4.1 Results and discussion

The results of the multi-temporal clustering of the Leaf, Bloom, and SOS products are summarized in Figure 4. This figure contains the WCSSE plots and two visualizations of the phenoregions (one based on centroid IDs and another one based on the actual DOYs of the centroids). The optimal number of phenoregions is 70 for both indices and 100 for the SOS metric. This indicates that land cover phenological variability is much larger than the one caused by temperature differences. However, the phenological regions derived from the spring indices have a much stronger spatial coherence as they are derived from a field with a high degree of autocorrelation. This is especially visible in the east where they strongly depend on the latitude. Topographic heterogeneity caused by the Rocky Mountains, the intermontane plateaus and the Pacific Mountain System lead to much more scattered climatic phenoregions in the west. The SOS metric is clearly dominated by the interactions between topography and land cover composition, leading to smaller phenoregions and difficult identification of clear large-scale phenological patterns. Yet the large forests in the north and east of the US as well as the so-called Corn Belt are recognizable. Finally, the bottom row of this figure shows the typical DOY of start of spring in each cluster. These values were calculated as the average of the annual DOY values that are returned as cluster centroids. Again, the clustering of the extended spring indices is smoother and resembles the temperature zonation of the USA. This was expected as these models are driven by daily temperature values. The SOS clusters resemble the land cover and ecosystem composition of the US, which was expected because land surface phenology is different for each land cover type, justifying the importance of phenological information to create land cover maps.

The results of the annual phenoregions are illustrated in Figure 5. The WCSSE plot shows that the optimal number of SI-x phenoregions is 15. This relatively low number indicates that the integration of early and later spring onset metrics leads to less variability than their independent analysis. Yet, the spatial patterns of the mode of each grid cell strongly resemble the spatial patterns found in the previous analysis. Finally, the bottom panel shows the percentage of times that the modal cluster was found in each grid cell. This figure shows that there is a large inter-annual (natural) variability in the way spring arrives to the US. Only the Sonora desert, the south of Florida and Texas, and parts of the Rocky Mountains are consistently clustered in the same group. The rest of the US shows moderate to high inter-annual variability, especially visible in the north and northwest of the US (Great Basin area, Cascades, and Pacific Coast Ranges).

Finally, it is important to realize that phenoregions are identified by clustering data and, as such, they represent a high level abstraction of the phenological patterns found in the data. This abstraction drastically reduces the amount of data that the analylist needs to look at and reduces the cognitive load required to understand the data. Therefore, we believe that annotated phenoregion maps are an ideal means to communicate changes (e.g. trends) in spring onset.

4.2 Iterative and step-wise analysis

The identification of the number of clusters in a data set is one of the most difficult problems in cluster analysis. One of the approaches is to use the *elbow curve* as used above for our phenoregion identification. Such approach requires the user to run a series of k-means clusterings, obtain their WCSSE, and create the plot shown on Figure 4. With Spark

[4] https://spark.apache.org/docs/2.2.0/mllib-clustering.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2019.2926292, IEEE Transactions on Big Data

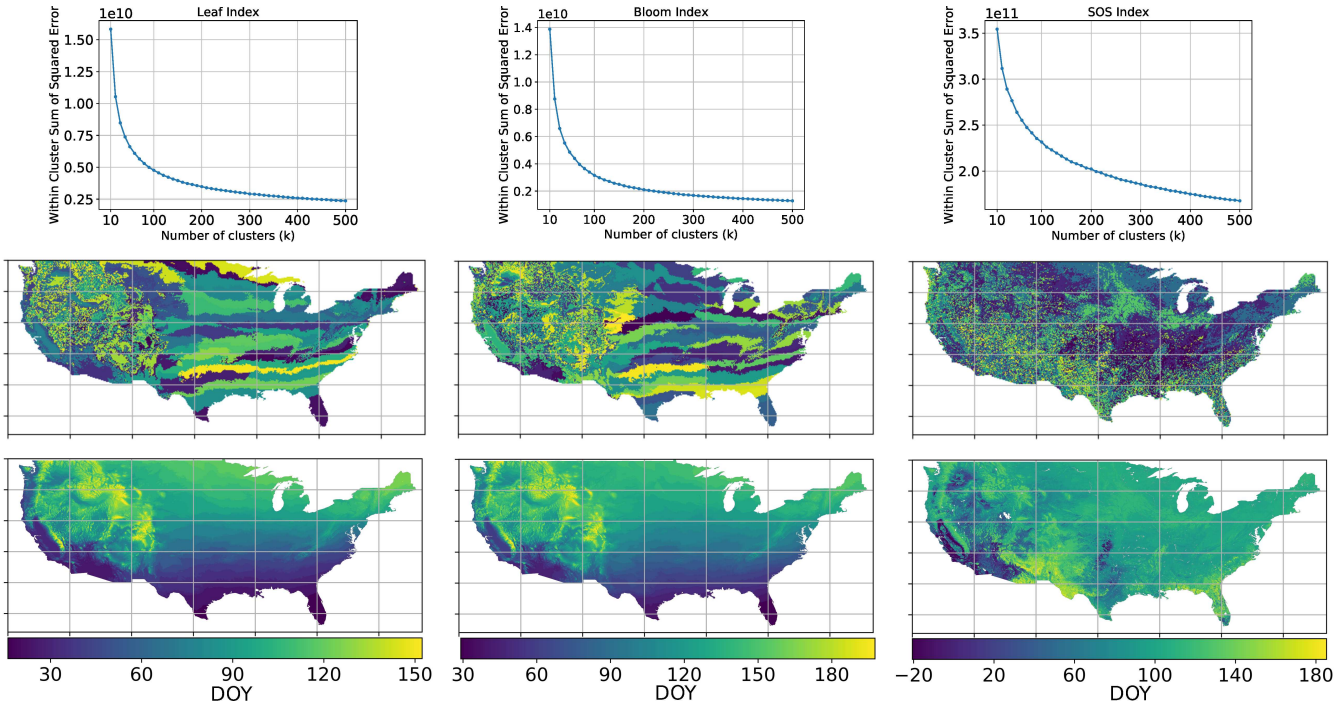

Fig. 4: Within-cluster sum of squared errors vs the number of clusters for the Leaf and Bloom indices and the SOS metric [Top row]. Clustering maps for the Leaf and Bloom indices (k=70) and the SOS metric (k=100) [Middle row]. Centroids maps for the Leaf and Bloom indices (k=70) and the SOS metric (k=100)[Bottom row].

it is possible to obtain such series in an efficient and effective way. The RDDs are cached in the nodes' memory to increase efficiency. With data cached in memory and each iteration being sent as independent job to the Spark cluster, it is possible to run loops with hundreds of iterations in an efficient way.

With RDDs cached, multiple k-means are requested with coarse iterations to quickly have an estimation of the ideal number of clusters. If the *elbow curve* is not clear due to the use of coarse iteration steps, the user can request a re-run with smaller iteration steps. Once the ideal number of clusters is identified, the user runs the k-means model in the input data and obtains the GeoTiffs in Figure 4. Such iterative and step-wise approach is not available in other cloud platforms specialized in the processing of geospatial data such as Google Earth Engine [48].

5 Correlating Temperature- and Satellite-Based Phenological Metrics

As stated in the introduction, the general applicability and deeper ecological meaning of phenological models and of satellite-derived land surface phenological metrics is not fully clear yet [24], [30]. To shed light on this, we performed a Pearson correlation analysis between the Leaf and Bloom indices and the AVHRR SOS phenometric. This analysis was made at a per-pixel level using the complete period of overlap between these phenological products (i.e. 1989-2014). After that, we identified areas that exhibit high positive and high negative correlation and cross-plotted the data to evaluate the coherence and the statistical predictability of the phenological metrics at hand.

5.1 Results and discussion

Figure 6 shows the results of the Pearson correlation analysis where we see that large areas exhibit moderate to high positive correlations. This confirms that temperature is one of the main drivers of phenological development at mid-latitudes. Our analysis also shows that the Leaf index is, in general, less correlated with the SOS than the Bloom index. This could indicate that satellites cannot detect the very early leaf onset, and that a certain amount of leaves (vegetation activity) is needed before spring onset can be detected from space using this sensor and method to estimate SOS. Interestingly, Figure 6 also shows areas with moderate to high negative correlation. These areas correspond to locations where phenology seems to be driven by other environmental factors (e.g. water). The SOS values of the grid cells with negative correlations tend to be negative. These results seem to indicate that late autumn and early winter SOS values in the year y are associated with early SI-x values for the year $y+1$. In terms of land cover, high positive correlations tend to occur in the evergreen needle leaf forests located in the northwest, the mixed forests around the Great Lakes, and mixed croplands and natural vegetation areas in the Midwest. Strong negative correlations mostly occur in the Corn Belt region and in the woody savannas of the southwest.

To explore the complementarity and predictive power of the chosen phenological indices and metrics, we selected 6 locations across the conterminous US (see lower panel in Figure 6) and fitted a linear regression model between the SI-x and the SOS. Figure 7 shows the results of this analysis together with the regression lines, the 1-to-1 line,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2019.2926292, IEEE Transactions on Big Data

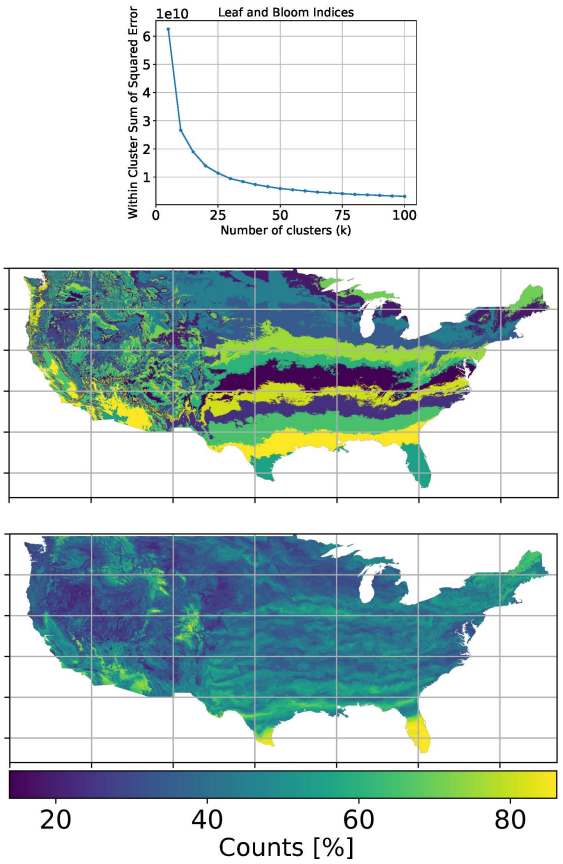

Fig. 5: Within-cluster sum of squared Errors vs the number of clusters for Leaf and Bloom [Top]. Clustering maps for mode of Leaf and Bloom clustering [Middle]. Percentage of counts for each mode [Bottom].

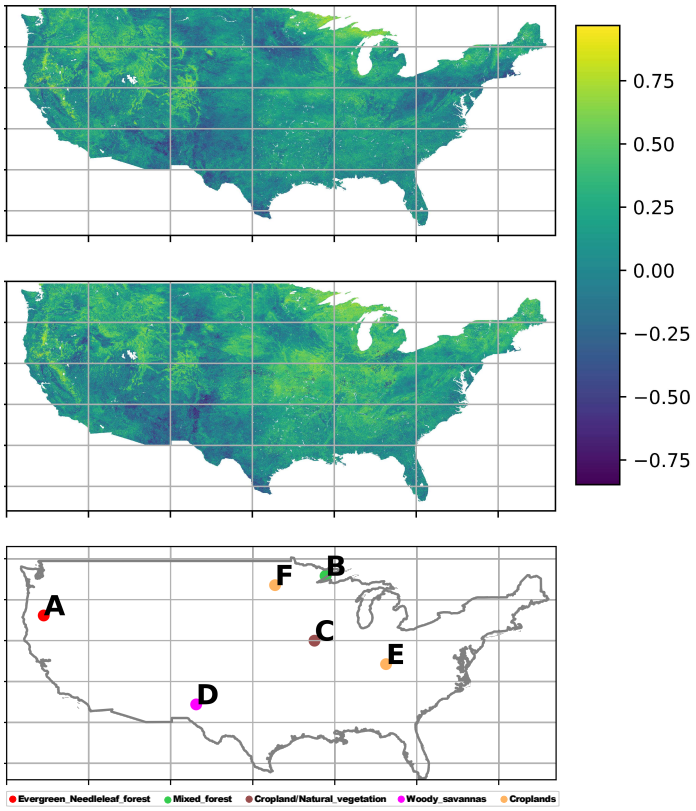

Fig. 6: Correlation between the Leaf index and SOS [Top], between the Bloom index and SOS [Middle], and sampled grid cells for further analysis [Bottom].

and the land cover of the selected cells. This figure shows that relative order between the Leaf and Bloom indices and the SOS metric depends on both the geographic location and the land cover type because the regression line is sometimes above and sometimes below the 1-to-1 line. This figure also illustrates that a high positive or negative correlation does not mean that the products are not biased. For instance, SOS values of about 140 for the evergreen needleaf forest correspond to Leaf values of around 120 and to Bloom values of about 170. Such "rules" could be used to combine multi-source phenological data and to create early anomaly detection systems.

5.2 User-tailored analysis

Although the identification of phenoregions works well with an RDD of sparse vectors, our correlation analysis requires the use of Spark's internal algorithms that only work with two RDDs of doubles or an RDD of vectors. [5]. Hence, calculating the per-pixel correlation between two time-series would require either millions of RDDs or the creation of a correlation matrix. To avoid that, we joined two RDDs into a single RDD with a tuple of sparse vectors per row. Using a map function, we then calculated the Pearson correlation between the two time series. This example shows how easy it is for a user to express the necessary calculations to conduct his/her research while exploiting effectively the Spark distributed framework.

[5]https://spark.apache.org/docs/2.1.1/mllib-statistics.html#correlations

6 Conclusions and Future Work

Climate change is modifying the timing of recurring biological events and, hence, that of seasonal transitions like spring onset. Understanding these changes is critical to design better climate change adaptation and management strategies. In this paper, we explore spring onset over the conterminous US by means of two case studies. In the first one, we mapped high-spatial resolution phenoregions and their changes in time. We found that the temperature-based phenoregions have a higher spatial coherence and are less complex than the phenoregions derived from the SOS metric. Through the annual analysis, we also found areas that tend to have stable timings for spring onset and areas where this timing is highly variable. We believe that both the multi-temporal and the annual phenoregions could be used to design more targeted climate change monitoring strategies because they identify homogeneous regions. They could also be used to (visually) communicate changes (e.g. trends) in spring onset to larger audiences because they represent a high-level abstraction of the data, which is easier to grasp. In the second case study, we analyzed the coherence

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2019.2926292, IEEE Transactions on Big Data

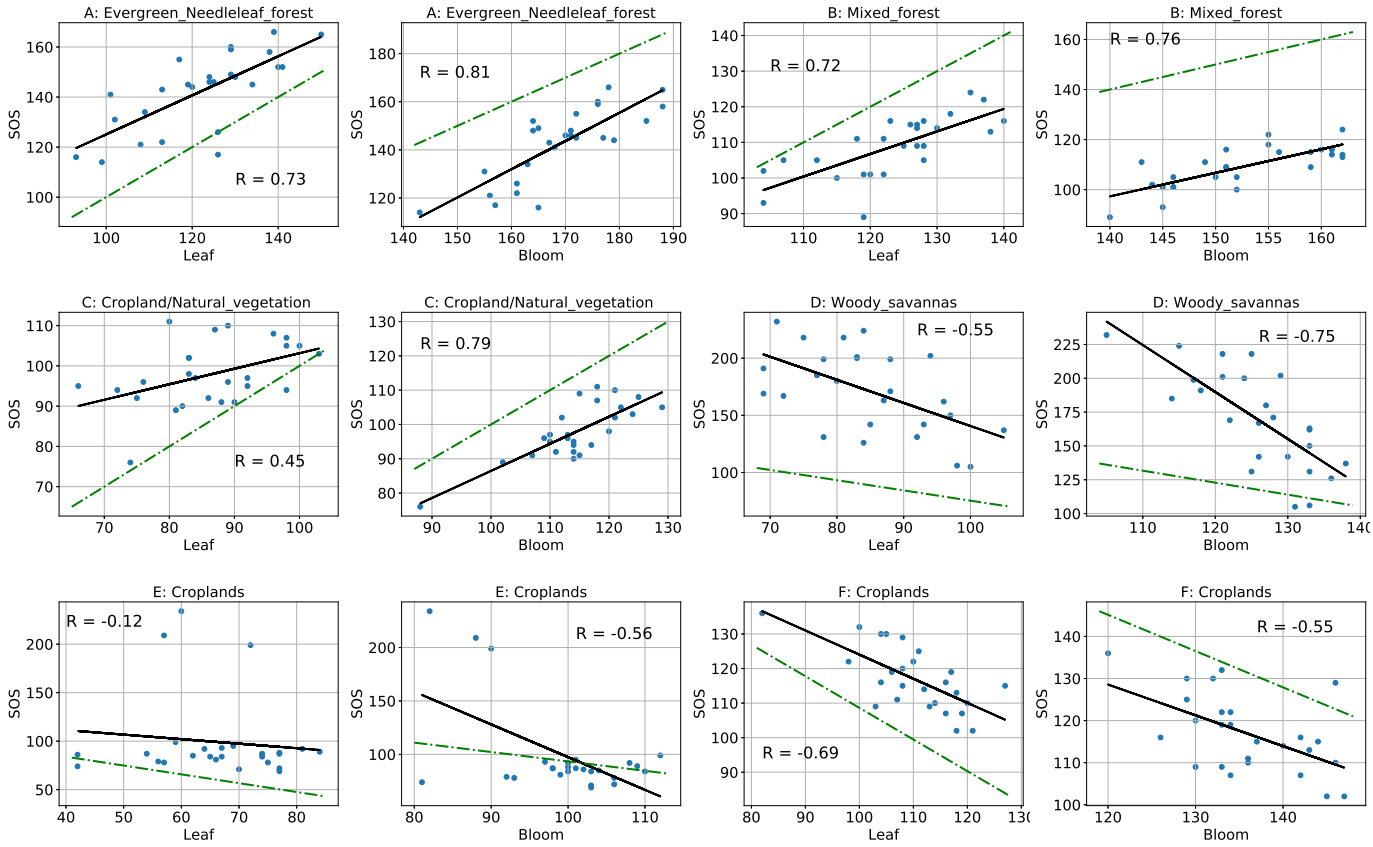

Fig. 7: Crossplots between the Leaf/Bloom index and SOS for the grid cells shown in Fig. 6. The black lines show the corresponding linear fits, the dashed green lines indicate the 1:1 line, the R values are the Pearson correlation coefficients and the titles list the MODIS-extracted land cover types.

between phenological indices and land surface metrics by mapping their correlation and investigating their statistical relationships. We found that there are several areas in the US where the spring indices are moderately to highly correlated (both positively and negatively) with the SOS metric. This indicates that the phenology of certain areas and land cover types is mostly driven by temperature. We also found that the Bloom index, which captures later spring onsets, is more correlated with this particular SOS product than the Leaf index. These correlations were analyzed in more detail by plotting the available data for six locations across the US. These plots confirm the moderate to high correlations and could be used to derive "mapping rules" to integrate both datasets or to predict the timing of one indicator by the value of the other one. Further analysis is still needed to better understand the complementary and synergistic value of these semantically distinct phenological products.

Both case studies are based on long-term (36 years) and high spatial resolution (1 km) simulations of the leaf and bloom indices, and on the longest possible (26 years) time series of the SOS metric at such a high spatial resolution. The efficient analysis of these relatively big datasets required the design and implementation of a distributed and cloud-based computational platform. For this, we used the Apache Spark ecosystem including its scalable machine learning library (SparkMLlib) and its geographic processing engine (GeoTrellis). Our case studies demonstrate the power of this solution for dealing with large geospatial grids and its versatility to create user-tailored analysis tools. We are confident that our open-source and scalable solution is future-proof and will be able to deal with the ever-increasing amounts of geospatial data about our living environment.

Future work will deal with the analysis of other spring onset indicators (either alternative SOS extraction methods or the derivation of very high spatial resolution metrics from the Sentinel missions), and with the integration of the millions of ground phenological observations collected by citizen scientists over the US, so that we can use them to benchmark our future work. Moreover, we are currently working on expanding the analytical functionality of our platform, e.g. to include a co-clustering algorithm that will allow the identification of phenoregions that are valid for a particular period [23]. Last but not least, we are working on including our computational platform and its philosophy in our educational curriculum so that new generations of geoinformation specialists become familiar with big data solutions.

ACKNOWLEDGMENTS

This work has been partially supported by the NLeSC Project: "High spatial resolution phenological modelling at continental scales" [6] and it was carried out using the Dutch

[6]https://github.com/phenology

national e-infrastructure with the support of the SURF Cooperative. The extended spring indices were computed in the framework of the "Green-wave" project, funded via a Google Faculty Award to the first author of this paper.

REFERENCES

[1] Mark D. Schwartz, "Green-wave phenology," *Nature*, vol. 394, pp. 839840, 1998.

[2] A.D. Richardson, T.F. Keenan, M. Migliavacca, Y. Ryu, O. Sonnentag, and M. Toomey, "Climate change, phenology, and phenological control of vegetation feedbacks to the climate system," *Agricultural and Forest Meteorology*, vol. 169, pp. 156–173, 2013.

[3] H. Lieth, "Purposes of a phenology book," in *Phenology and seasonality modeling*. Springer, 1974.

[4] N. B. Villoria, J. Elliott, C. Mller, J. Shin, L. Zhao, and C. Song, "Rapid aggregation of global gridded crop model outputs to facilitate cross-disciplinary analysis of climate change impacts in agriculture," *Environmental Modelling & Software*, vol. 75, pp. 193 – 201, 2016.

[5] M. D. Schwartz, R. Ahas, and A. Aasa, "Onset of spring starting earlier across the Northern Hemisphere," *Global Change Biology*, vol. 12, no. 2, pp. 343–351, 2006.

[6] J. E. Olesen and M. Bindi, "Consequences of climate change for European agricultural productivity, land use and policy," *European Journal of Agronomy*, vol. 16, no. 4, pp. 239 – 262, 2002.

[7] A. Barr, T. A. Black, and H. McCaughey, "Climatic and phenological controls of the carbon and energy balances of three contrasting boreal forest ecosystems in western Canada," in *Phenology of ecosystem processes*, pp. 3–34. Springer, 2009.

[8] E.E. Cleland, I. Chuine, A. Menzel, H.A. Mooney, and M.D. Schwartz, "Shifting plant phenology in response to global change," *Trends in Ecology and Evolution*, vol. 22, no. 7, pp. 357–365, 2007.

[9] A. H. Rosemartin, E. G. Denny, J. F. Weltzin, R. L. Marsh, B. E. Wilson, H. Mehdipoor, R. Zurita-Milla, and M. D. Schwartz, "Lilac and honeysuckle phenology data 1956-2014," *Scientific Data*, vol. 2, pp. 150038, 2015.

[10] M. D. Schwartz, "Phenology: an integrative environmental science," Springer, 2003.

[11] H. Mehdipoor, R. Zurita-Milla, E-W. Augustijn, and A. J. H. van Vliet, "Checking the consistency of volunteered phenological observations while analysing their synchrony," *ISPRS International Journal of Geo-Information*, vol. 7, no. 12, 2018.

[12] R. Zurita-Milla, J. A. E. van Gijsel, N. A. S. Hamm, P. W. M. Augustijn, and A. Vrieling, "Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 1914–1921, April 2013.

[13] K. M. de Beurs and G. M. Henebry, "Land surface phenology, climatic variation, and institutional change: Analyzing agricultural land cover change in Kazakhstan," *Remote Sensing of Environment*, vol. 89, no. 4, pp. 497 – 509, 2004.

[14] M. A. White, F. Hoffman, W. W. Hargrove, and R. R. Nemani, "A global framework for monitoring phenological responses to climate change," *Geophysical Research Letters*, vol. 32, no. 4, pp. 295–309, 2005, L04705.

[15] J. M. Fitchett, S. W. Grab, and D. I. Thompson, "Plant phenology and climate change: Progress in methodological approaches and application," *Progress in Physical Geography*, vol. 39, no. 4, pp. 460–482, 2015.

[16] M. D. Schwartz and B. C. Reed, "Surface phenology and satellite sensor-derived onset of greenness: An initial comparison," *International Journal of Remote Sensing*, vol. 20, no. 17, pp. 3451–3457, 1999.

[17] W.W. Hargrove, J.P. Spruce, G.E. Gasser, and F.M. Hoffman, "Toward a national early warning system for forest disturbances using remotely sensed canopy phenology," *Photogrammetric Engineering and Remote Sensing*, vol. 75, no. 10, pp. 1150–1156, 2009.

[18] R. T. Mills, F. M. Hoffman, J. Kumar, and W. W. Hargrove, "Cluster analysis-based approaches for geospatiotemporal data mining of massive data sets for identification of forest threats," 2011, vol. 4, pp. 1612–1621.

[19] Y. Gu, J. F. Brown, T. Miura, W. J. D. Van Leeuwen, and B. C. Reed, "Phenological classification of the United States: A geographic framework for extending multi-sensor time-series data," *Remote Sensing*, vol. 2, no. 2, pp. 526–544, 2010.

[20] F. B. Silva, Y. E. Shimabukuro, L. E. O. C. Arago, L. O Anderson, G. Pereira, F. Cardozo, and E. Arai, "Large-scale heterogeneity of Amazonian phenology revealed from 26-year long AVHRR/NDVI time-series," *Environmental Research Letters*, vol. 8, no. 2, pp. 024011, 2013.

[21] Y. Zhang, G. F. Hepner, and P. Dennison, "Delineation of phenoregions in geographically diverse regions using k-means++ clustering: A case study in the Upper Colorado River Basin," *GIScience and Remote Sensing*, vol. 49, no. 2, pp. 163–181, 2012.

[22] Y. Zhang and G. F. Hepner, "The dynamic-time-warping-based k-means++ clustering and its application in phenoregion delineation," *International Journal of Remote Sensing*, vol. 38, no. 6, pp. 1720–1736, 2017.

[23] X. Wu, R. Zurita-Milla, and M. J. Kraak, "A novel analysis of spring phenological patterns over Europe based on co-clustering," *Journal of Geophysical Research: Biogeosciences*, vol. 121, no. 6, pp. 1434–1448, 2016.

[24] M. A. White, de K. M. Beurs, K. Didan, D. W Inouye, A. D Richardson, O. P. Jensen, J. O'Keefe, G. Zhang, R. R. Nemani, et al., "Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982–2006," Wiley Online Library, 2009.

[25] X. Zhang, J. Wang, F. Gao, Y. Liu, C. Schaaf, M. Friedl, Y. Yu, S. Jayavelu, J. Gray, L. Liu, D. Yan, and G. M. Henebry, "Exploration of scaling effects on coarse resolution land surface phenology," *Remote Sensing of Environment*, vol. 190, pp. 318 – 330, 2017.

[26] T. R. Ault, M. D. Schwartz, R. Zurita-Milla, J. F. Weltzin, and J. L. Betancourt, "Trends and natural variability of spring onset in the Coterminous United States as evaluated by a new gridded dataset of spring indices," *Journal of Climate*, vol. 28, no. 21, pp. 8363–8378, 2015.

[27] E. Izquierdo-verdiguier, R. Zurita-milla, T. R. Ault, and M. D Schwartz, "Development and analysis of spring plant phenology products: 36 years of 1-km grids over the conterminous US," *Agricultural and forest meteorology*, vol. 262, pp. 34–41, 2018.

[28] B. Duchemin, J. Goubier, and G. Courrier, "Monitoring phenological key stages and cycle duration of temperate deciduous forest ecosystems with NOAA/AVHRR data," *Remote Sensing of Environment*, vol. 67, no. 1, pp. 68 – 82, 1999.

[29] M. D. Schwartz, B. C. Reed, and M. A. White, "Assessing satellite-derived start-of-season measures in the conterminous USA," *International Journal of Climatology*, vol. 22, no. 14, pp. 1793–1805, 2002.

[30] M. D. Schwartz and J. M. Hanes, "Intercomparing multiple measures of the onset of spring in eastern North America," *International Journal of Climatology*, vol. 30, no. 11, pp. 1614–1626, 2010.

[31] A. Vrieling, M. Meroni, R. Darvishzadeh, A. K. Skidmore, T. Wang, R. Zurita-Milla, K. Oosterbeek, B. O'Connor, and M. Paganini, "Vegetation phenology from Sentinel-2 and field cameras for a Dutch barrier island," *Remote Sensing of Environment*, vol. 215, pp. 517 – 529, 2018.

[32] "Apache spark," https://spark.apache.org.

[33] "Apache spark-mllib," https://spark.apache.org/mllib.

[34] "Minio," https://www.minio.io.

[35] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, DB Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar, "MLlib: Machine learning in apache spark," *Journal of Machine Learning Research*, vol. 17, no. 34, pp. 1–7, 2016.

[36] "Geotrellis," https://geotrellis.io.

[37] R. Goncalves, S. Verhoeven, N. Drost, and J. Attema, "Emma," doi:10.5281/zenodo.996308.

[38] "Vagrant," https://www.vagrantup.com.

[39] "Ansible," https://www.ansible.com.

[40] "Jupyterhub," https://github.com/jupyterhub/jupyterhub.

[41] C. Tilmes, Y. Yesha, and M. Halem, "Tracking provenance of earth science data," *Earth Science Informatics*, pp. 59–65, 2010.

[42] C. Tilmes, Y. Yesha, and M. Halem, "Distinguishing provenance equivalence of earth science data," *Procedia Computer Science*, vol. 4, pp. 548 – 557, 2011.

[43] M. D. Schwartz, T. R. Ault, and J. L. Betancourt, "Spring onset variations and trends in the continental United States: past and regional assessment using temperature-based indices," *International Journal of Climatology*, , no. 33, pp. 2917 – 2922, 2013.

[44] T. R. Ault, R. Zurita-Milla, and M. D. Schwartz, "A Matlab© toolbox for calculating spring indices from daily meteorological data," *Computers & Geosciences*, vol. 83, pp. 46 – 53, 2015.

[45] E. Izquierdo-Verdiguier, R. Zurita-Milla, T. R. Ault, and M. D. Schwartz, "Using cloud computing to study trends and patterns in the Extended Spring Indices," in *Phenology 2015: Third International Conference on Phenology*, p. 1. Humboldt-University of Berlin and the Adnan Menders University Aydin, 2015.

[46] L. Ji and J. F. Brown, "Effect of NOAA satellite orbital drift on AVHRR-derived phenological metrics," *International Journal of Applied Earth Observation and Geoinformation*, vol. 62, pp. 215 – 223, 2017.

[47] B. C. Reed, J. F. Brown, D. VanderZee, T. R. Loveland, J. W. Merchant, and D. O. Ohlen, "Measuring phenological variability from satellite imagery," *Journal of Vegetation Science*, vol. 5, no. 5, pp. 703–714, 1994.

[48] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18 – 27, 2017.