EXPLORING VEGETATION PHENOLOGY AT CONTINENTAL SCALES: LINKING TEMPERATURE-BASED INDICES AND LAND SURFACE PHENOLOGICAL METRICS

*R. Zurita-Milla *, R. Goncalves **, E. Izquierdo-Verdiguier *, F.O. Ostermann **

Faculty ITC - University of Twente * and NLeSC **, the Netherlands

ABSTRACT

Phenology is the science that studies the timings of recurring biological events such as leafing and blooming as well as their causes and variations in space and time. Spatially explicit environmental datasets and are key to understand phenological dynamics at continental to global scales. Here we present a novel exploratory analysis where we link temperature-based phenological indices and land surface phenological metrics derived from remotely sensed images. Our exploratory analysis, illustrated with two multi-decadal and high-spatial resolution phenological products for continental USA, focuses on identifying phenological regions and on mapping the coherence between phenological products. To cope with the computational challenges of analyzing big geo-datasets, we executed our analysis on a cloud platform running Apache Spark. First results show that weather, climate and land cover variability modulate phenological patterns in contrasting ways, and we believe that our computational solution work paves the path towards the analysis of global vegetation phenology at very high spatial resolution.

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

1. INTRODUCTION

Phenology studies the timing of recurring plant and animal biological phases, their causes, and their interrelations [1]. This seasonal timing varies from place to place and from year to year because it is strongly influenced by environmental conditions. Understanding this variability is critical to quantify the impact of climate change on our planet. In this work we present a novel exploratory analysis of two of the most important sources of spatio-temporal phenological data: phenological models based on weather- and location-related factors, and land surface phenological metrics derived from Earth observation sensors.

Phenological models. The Extended Spring Indices (SI-x; [2]) are a suite of models that transform daily temperatures into consistent phenological metrics that can be used to study the impact of global warming on vegetated canopies. [1]. More precisely, the SI-x models predict the day of the year (DOY) of first leaf and of first bloom for three key indicator species [3]. These phenological dates can be used to track spring onset at specific locations by using data from weather stations [4] or at continental scales by using gridded weather and/or climatic datasets [5].

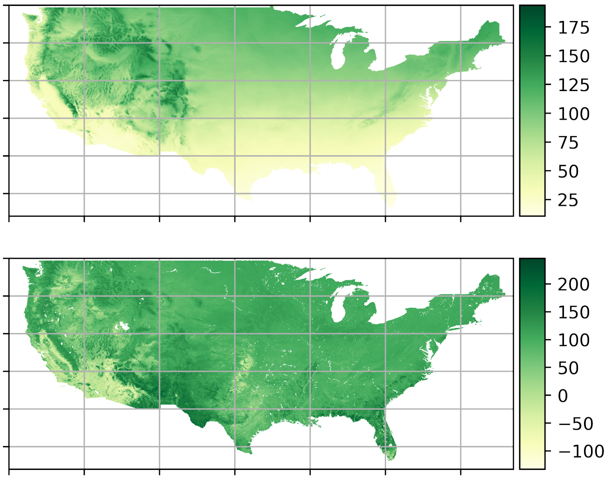

Fig. 1: Average of Leaf index [Top] and AVHRR SOS [Bottom] maps of contiguous North-America from 1989 to 2014.

In this work, we use a new long-term (1980 to 2015) and high spatial resolution ($1km$) version of the Leaf and Bloom indices, which was recently generated for the coterminous US by adapting the SI-x models to a cloud computing environment [6]. Figure 1 [Top] illustrates, as an example, the average of the Leaf index from 1989 to 2014. This map shows a clearly noticeable spring gradient, with low values in the South and high DOY values in the North.

Land surface phenology. Time series of remotely sensed images can be used to derive various land surface phenological metrics. One of these metrics is the so-called Start of Season (SOS), which indicates the beginning of photosynthetic activity in plants. Several SOS products exits in literature. Often linked to a particular sensor or study. Here we use a SOS product specifically made for the US by processing

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

doi: 10.2760/383579

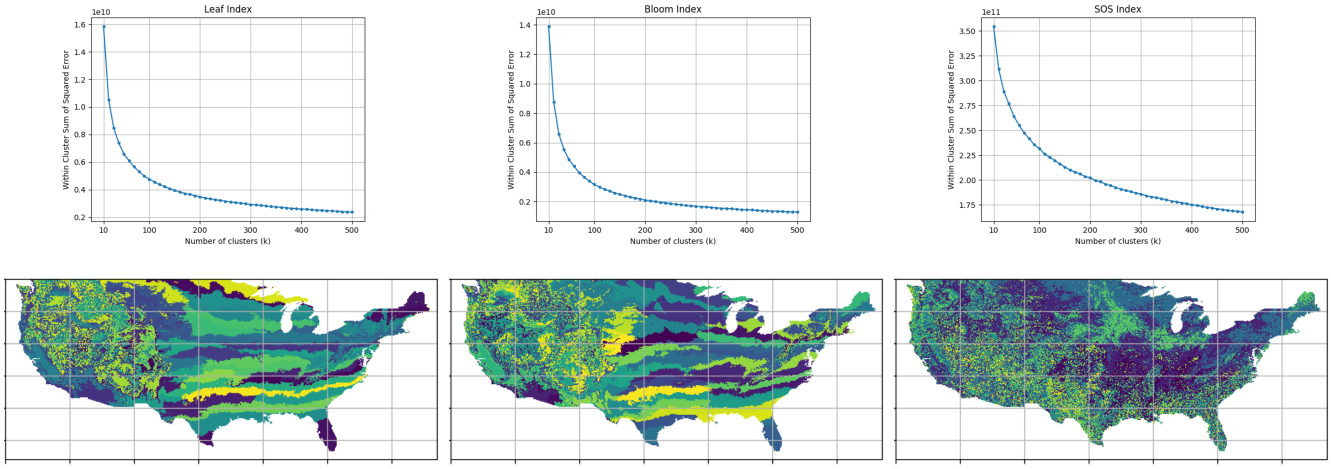

Fig. 2: Within cluster sum of squared Errors vs the number of clusters for the Leaf and Bloom indices and the SOS metric [Top row]. Clustering maps for the Leaf and bloom indices (k=70) and the SOS metric (k=100) [Bottom row]

time series of the Advanced Very High Resolution Radiometer (AVHRR) sensor [2]. The AVHRR images were first transformed into a smooth time series of Normalized Difference Vegetation Index (NDVI). Then a curve derivative method was applied to predict NDVI values based on the previous observations Finally, the SOS day was determined by identifying the day when the smoothed NDVI values become larger than the predicted NDVI values [7].

The spatial resolution of this product matches that of the SI-x but it is only available for the period 1989 - 2014 [3]. Hence our exploratory analysis is based on the products available for this period. Again, as an example Figure 1 [Bottom] illustrates the average SOS values from 1989 to 2014. In this case the spring phenological gradient is less visible as the SOS depends on both the land cover and the weather conditions. Notice that the negative values in the SOS map indicate that the SOS took place the year before (i.e. in 1988).

Computational solution. Analyzing multi-decadal and very high spatial resolution phenological products at continental scales remains a challenging task. In this work we use a cloud-based solution based on Apache Spark [8] and its scalable machine learning library MLlib [9] to perform our exploratory data analysis. Given the lack of well-tested Spark solutions in the domain of big geo-data, a secondary aim of our work is to evaluate the potential of such a computational solution to analyze big raster datasets, in both local and cloud-based environments.

With the data stored in the original file formats, such as GeoTiff and HDF, users are able analyze the data through Jupyter notebooks running either Python, R or Scala. These notebooks are not only used to share results among scientists but also as a provenance method for the scientific results.

Using the phenological products described above and our computational platform, we first identify regions with similar phenology (Section 2) and then study their correlation (Section 3). After that, we provide additional details on our computational platform (Section 4) and, finally, we summarize our findings and present follow up activities (Section 5).

2. MAPPING PHENOREGIONS

Clustering is a popular exploratory data analysis method that allows analysts to study their datasets at a higher level of abstraction [10]. Here we use K-means to identify regions with similar phenology (i.e. phenoregions) The three phenological products were clustered into k groups (with k values ranging from 10 to 500 in steps of 10) and the optimal k value was identified by the ”elbow” of the Within Cluster Sum of Squared Error (WCSSE) graph. Figure 2 shows the WCSSE plots and the clustering results.

The optimal number of phenoregions is 70 for the Leaf and Bloom indices and 100 for the SOS metric. This indicates that land cover phenological variability is larger than the one caused by temperature differences. However, the phenological regions derived from the spring indices have a much stronger spatial coherence, especially on the East Small scale differences in elevation and land cover lead to much more scattered phenoregions in the American West.

3. SPATIO-TEMPORAL CORRELATION

The ecological meaning of land surface phenological metrics is not fully clear yet [11]. To shed light on this, we performed a spatio-temporal correlation analysis between the Leaf and Bloom indices and the SOS metric. Figure 3 shows that large areas exhibit moderate to high positive correlations. This confirms that temperature is, indeed, one of the main drivers of

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

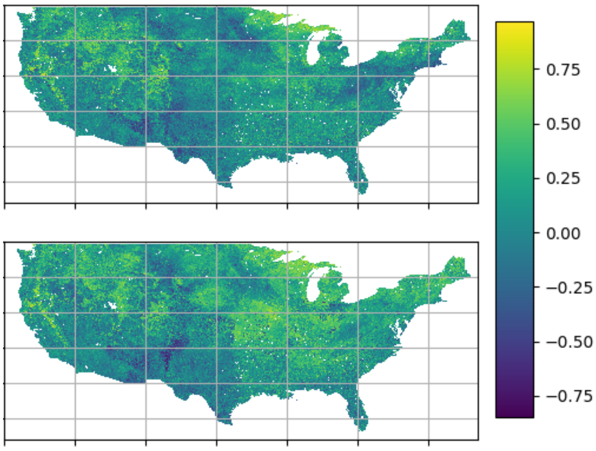

Fig. 3: Correlation between the Leaf index and SOS [Top] and between the Bloom index and SOS [Bottom]

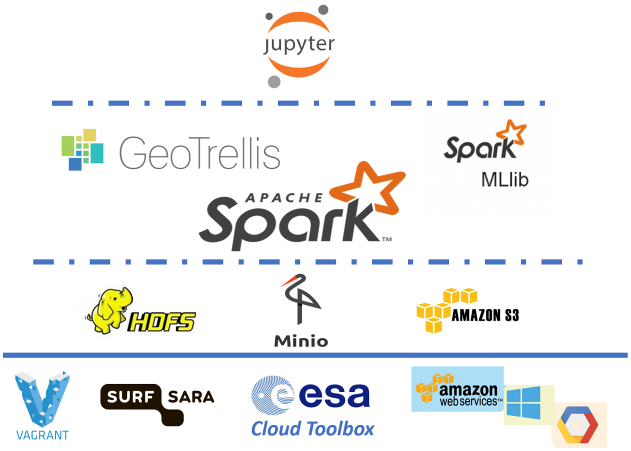

Fig. 4: Computational platform

phenological development. 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 can be seen from space.

Interestingly, Figure 3 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) and to areas where the SOS happens in the second half of the year.

4. COMPUTATIONAL PLATFORM

Our research work is conducted in a open-source platform using cloud-based infra-structures. With the aim to either do massive data analysis or a simple exploratory one, our computational platform is designed for easy user interaction and scalability. Users interact with the platform through Jupyter notebooks and computations are pushed down to a remote cluster. The computations are designed to use distributed data structures and Spark internals for efficient distributed processing. For its deployment and management we use Emma [12], a project to create a platform for development of applications for Spark and DockerSwarm clusters.

A cloud-based platform. The platform runs on an infrastructure composed by local or virtual machines attached to a large object storage with an Amazon Simple Storage Service (S3). The latter is becoming a de-facto API standard for objects-storage. It is supported by Google and Microsoft cloud services for easy port of cloud-based applications. The machines are prepared/constructed by either preparing cloud virtual machine or constructing using Vagrant [13] boxes. The latter allows the platform to be simulated on a local machine, i.e., provide a local development environment.

Once the machines are prepared the servers are provisioned using Ansible, an automation tool for IT infra-structure. Ansible [14] playbooks are used to create a storage layer, processing layer and JupyterHub [15] services. With Ansible we are able to deploy a platform with the same features at different locations, such as local cluster, national infra-structure or even a commercial cloud provider. Such feature allows us to have tool-provenance for easily repeatability of experiments between Scientists.

The platform's architecture is organized in three layers: storage layer, processing layer and JupyterHub services for user-interaction, (Figure 4). The storage layer offers two flavors of storage, file-base by Hadoop Distributed File System (HDFS), and object-based by Amazon S3 service. For local environments we use Minio [16], an open source object storage server with Amazon S3 compatible API, to avoid application re-write when moving to a cloud provider. HDFS is used by Apache Spark [8] to exploit data locality and to store intermediates to avoid re-computations. The object storage is used to store the phenology data products and other remote sensing data products.

At the processing layer we have Spark with its machine learning library SparkMLlib [9] and GeoTrellis [17] for high-performance geographic data processing. With GeoTrellis GeoTiffs are directly read from the S3 storage into Resilient Distributed Datasets (RDDs). With the phenology data products loaded as RDDs we then exploit Spark's internal for distributed data processing. One example is the the mapping of pheno regions in Section 2.

For the data analysis the user expresses the operations either in Scala, R or Python using Jupyter notebooks. Hence, with a browser and remote connection the user is able to express a research question or collect an insight over large data sets. All computations are pushed down to the computational platform and results fetched back for data visualization.

Scalability. On our platform computations are not only pushed down for remote processing, but they are also designed to exploit Spark's cluster computational features. To achieve that data is always loaded into memory-based data structure such as RDD, DataFrames and distributed matrices. With the data loaded into Spark's memory-based structures, distributed task scheduling and fault-tolerance is then handled by Spark.

Such 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., use higher resolution data from Sentinel-2, or for changes in the amount of available resources when moving to a different cloud-infrastructure. The decision of which structure to use and a study on the impact of different resource allocation, i.e., a detailed performance profile, is out of the scope of this paper.

5. CONCLUSIONS AND FUTURE WORK

In this paper we exploit the Apache Spark ecosystem for large scale distributed processing. With our phenological experiments we have demonstrated that it possible to map phenoregions at high spatial resolution and at continental scales. Moreover, we have shown that temperature-based indices are both positively and negatively correlated with the AVHRR SOS metric. Further analysis is needed to better understand the complementary and synergistic value of these two phenological products.

Future work will deal with the integration of the millions of ground phenological observations collected by citizen scientists as well as with the analysis of very high spatial resolution phenological metrics from the Sentinel missions. We plan to conduct this analysis at the ESA cloudtoolbox [4] and at different commercial cloud providers in an attempt to verify if our platform is generic enough.

Acknowledgments

This work has been partially supported by the NLeSC Project: "High spatial resolution phenological modelling at continental scales" [5] and it was carried out using the Dutch national e-infrastructure provided by 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.

[4] `http://eogrid.esrin.esa.int/cloudtoolbox`
[5] `https://github.com/phenology`

6. REFERENCES

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

[2] 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*, 2013.

[3] 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 19562014," *Scientific Data*, vol. 2, 2015.

[4] 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.

[5] E. Izquierdo-Verdiguier, R. Zurita-Milla, T. R. Ault, and M. D. Schwartz, "Development and analysis of spring plant phenology: 36 years of 1-km grids over the conterminous us," *Agricultural and Forest Meteorology*.

[6] 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," *Third International Conference on Phenology*, 2015.

[7] 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.

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

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

[10] X. Wu, R. Zurita-Milla, and M. J. Kraak, "A novel analysis of spring phenological patterns over europe based on coclustering," *Journal of Geophysical Research: Biogeosciences*, 2016.

[11] 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," *Global Change Biology*, 2009.

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

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

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

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

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

[17] "Geotrellis," `https://geotrellis.io`.