Discovering relationships in climate-vegetation dynamics using satellite data

Christina Papagiannopoulou[1], Diego Miralles[3,2], Mathieu Depoorter[2], Niko E.C. Verhoest[2], Wouter Dorigo[4], and Willem Waegeman[1]

[1] Depart. of Mathematical Modelling, Statistics and Bioinformatics
Ghent University, Belgium
christina.papagiannopoulou, willem.waegeman@ugent.be
[2] Laboratory of Hydrology and Water Management, Ghent University, Belgium
mathieu.depoorter, niko.verhoest, diego.miralles@ugent.be
[3] Depart. of Earth Sciences, VU University Amsterdam, the Netherlands
[4] Depart. of Geodesy and Geo-Information, Vienna University of Technology, Austria
wouter.dorigo@geo.tuwien.ac.at

Abstract. Advances in satellite Earth observation have resulted in the development of consistent global historical records of environmental and climatic variables, forming enormous amounts of multivariate time series. In this work we present a novel machine learning framework for detecting relationships between climatic time series and vegetation indices. Our pipeline consists of several components, including data fusion from various databases, time series decomposition techniques, feature construction methods and predictive modeling. Experimental results indicate that with this pipeline it is possible to detect patterns that express relationships in large-scale climate data.

Keywords: time series analysis, feauture construction, shapelets

1 Introduction

Earth Observation (EO) satellite data provide a wealth of information about the dynamics of our planet in recent decades. Independent sensors on different platforms monitor vegetation, soils, oceans and atmosphere, collecting optical, thermal, microwave, altimetry, or gravimetry information. Composite records of environmental and climatic variables now span up to 35 years, enabling the study of climate-vegetation interactions over multi-decadal scales. Such records can be interpreted as long multivariate time series with different spatial and temporal resolutions. Due to their volume and complexity, the resulting datasets pose important challenges with respect to data processing and data analysis, leading to a need for developing novel methods.

Vegetation is a major player in the global climate system by affecting the water, the energy and the carbon cycles. Plants alter climate through the transfer of water vapor from land to atmosphere, direct effects on the surface net radiation, exchange of carbon dioxide with the atmosphere, or changes in roughness

length affecting wind speed and direction. Given the crucial role of vegetation in climate, and the influence of climate on vegetation dynamics, understanding how vegetation will respond to projected climatic changes is crucial to narrow down the uncertainty in the predictions of global warming. A first and necessary step, however, is to investigate the sensitivity of vegetation to past-time climate variability. Simple correlation statistics have led to important steps forward in understanding the link between climate and vegetation while considering a few datasets (e.g. [8, 13]). But to fully use this available stream of information in constant expansion, new and more sophisticated approaches are required.

In this article we present a novel framework for finding climatic drivers that affect vegetation. We will consider a data-driven approach, analyzing satellite time series that span the entire globe and three decades. In a first step, we describe how we combine data from different sources. In a second step, we reformulate the problem of discovering relationships in climate-vegetation dynamics as a machine learning problem, where vegetation is considered as the target time series, while climate information sources serve as predictor time series. We apply time series decomposition techniques to the target vegetation time series and the various predictor climatic time series to isolate seasonal cycles, trends and residuals in order to remove unwanted correlations that originate from seasonal and trend effects. Subsequently, we explore various techniques for constructing high-level features from climatic time series using techniques that are similar to shapelets [14]. We employ standard machine learning techniques, as nonlinear autoregression method, in order to search for shapelets that are predictive with respect to the residuals of vegetation time series. As shown in the experimental results, this approach allows us to discover novel insights w.r.t. climate-vegetation dynamics, moving beyond the state-of-the-art in this application domain.

2 Methodology

2.1 Data collection and fusion

Since we aim to disentangle the effect of past-time climate variability on global vegetation, data sets have been selected from the current pool of satellite and *in situ* observations. The environmental and climate variables are collected on the basis of meeting a series of spatiotemporal requirements: (a) to span multi-decadal records, (b) to have a global coverage, and (c) to be available at an adequate spatial and temporal resolution. All these data sets span the study period 1981-2011 at the global scale, and have been converted to a common monthly temporal resolution and $1^\circ \times 1^\circ$ latitude-longitude spatial resolution. To do so, we have used averages to re-sample original data sets found at finer native resolution, and linear interpolation to resample coarser-resolution ones. Five different climatic and environmental drivers of vegetation dynamics have been considered: precipitation, temperature, radiation, snow depth (i.e. snow water equivalents) and surface soil moisture. Rather than using a single data set for each of these variables, the approach has been to collect and utilise the

largest possible number of data sets meeting the above-mentioned requirements (see Supplementary Material for details[5]).

For vegetation, we use the satellite remote sensed products of Normalized Difference Vegetation Index (NDVI), a graphical indicator which is used to assess whether the target being observed contains live green vegetation or not. Data from the Global Inventory Modeling and Mapping Studies (GIMMS) data set has been used, which is one of the most commonly used NDVI data sets [10] covering a wide time interval of 30 years (1981-2011).

2.2 Nonlinear autoregressive model

In this paper we will analyze climate-vegetation dynamics in a machine learning setting, where for each pixel the vegetation time series will be considered as target time series, using the notation Y_t, and the other time series will serve as predictor time series, using the notation X_t. This will result in a regression problem with moving windows (NDVI is a continuous variable). We hypothesize that relationships between climate and vegetation are expected to be highly nonlinear, we will replace the linear Vector Autoregressive (VAR) models with nonlinear machine learning models. We have chosen to use random forests [1], a well-known nonlinear machine learning method that has shown its merits in diverse application domains. We will evaluate the performance in terms of explained variance, R^2, defined as:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{1}$$

where $RSS = \sum_{i=1}^{n}(y_i - y_i')^2$ is the residual sum of squares, $TSS = \sum_{i=1}^{n}(y_i - \bar{y_i})^2$ the total sum of squares, the y_i and y_i' the real and the predicted value respectively, $\bar{y_i}$ the mean and n the number of the examples. In our analysis, we will treat each pixel on earth as a separate problem, because vegetation is assumed to have a very local pattern.

Our approach is visualized in Figure 1. For a given value of the target time series Y_t at time stamp t, we investigate properties of the different predictor time series X_t by considering a moving window of the months before time stamp t. Within those windows, we intend to construct higher-level features, which represent properties in each of the climate time series that are predictive w.r.t. the vegetation time series.

Before constructing those features, we first decompose the target and predictor time series into trends, seasonal cycles and anomalies. This is an important step, because the trend and seasonal component of the vegetation time series are not influenced by climatic features. In what follows we will work further with the anomalies, and the goal will be to forecast those using information from the predictor time series.

Many methods for decomposing time series have been proposed in the literature [3,11]. We decided to use an additive model without break-points, since it is

[5] `http://www.sat-ex.ugent.be/supplementary_material.pdf`

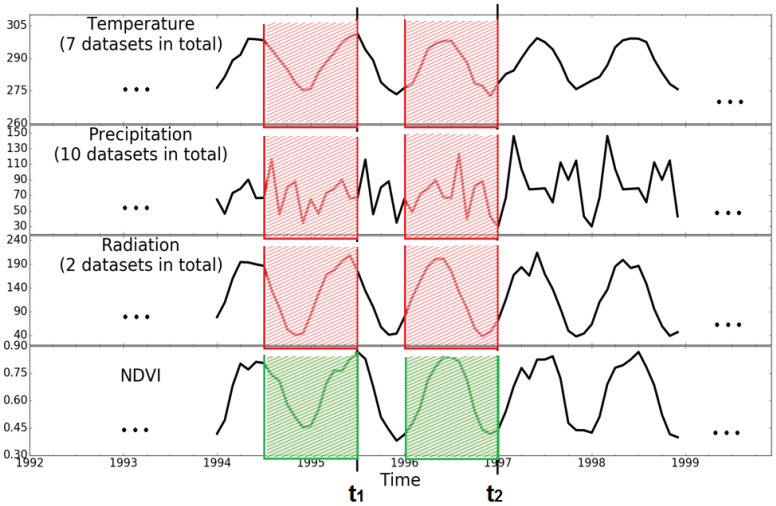

Fig. 1: An illustration of the moving window approach taken to discover relationships between vegetation time series and climatic time series. See text for details.

conceptually-simple, while delivers satisfactory results in a reasonable amount of time. In a first step, time series Y_t is at every pixel de-trended linearly based on the entire study period, using a simple linear regression model, $Y_t \approx \beta_1 \times t + \beta_0 = T_t$. In this way we obtain the de-trended time series, $D_t = Y_t - T_t$. In a second step, the seasonal cycle S_t is estimated as a monthly expectation, taking the multi-year average for each month of the year. In a last step, the anomalies are calculated by subtracting the corresponding monthly expectation from the de-trended time series, $R_t = D_t - S_t$. The same time series decomposition method is followed for all predictor time series as well (see Supplementary material for details[5]).

2.3 Feature construction from shapelets

We intend to identify patterns in the windows of Figure 1 that are predictive w.r.t. the anomalies of the target vegetation time series. To this end, we will analyze subsequences of the moving window specified for time stamp t, a technique that is similar to so-called shapelets [14]. Techniques for finding shapelets have been mainly applied to the problem of time series classification while various metrics are used for the evaluation of their quality [6, 7, 14].

Unlike most applications of feature selection, we are in our analysis less interested in discovering individual features. In contrast, we rather intend to know which types of climatic drivers affect vegetation in different regions of the world. The predictor variables will hence be grouped based on the type of the climate

variables. More specifically, three main categories are distinguished, namely radiation, water (including precipitation, soil moisture and snow-water equivalents) and temperature. Then, we run ridge regression and random forests separately for each group and we examine the value of R^2, investigating which group of predictors explains better the variability of the NDVI residuals. Each group of predictors is assessed separately because climate drivers are highly correlated and our goal is to investigate their predictive power for each region separately.

2.4 Overview of features extracted from shapelets

Besides the application to discovering relationships between time series instead of time series classification, another difference with shapelet construction papers is that we will not be interested in the shapelets itself, but aggregates that can be derived from them. In particular we will focus on three categories: lags, cumulatives and extreme indices. They are described below in more detail.

Lags: Vegetation responds to meteorological and environmental changes at different time scales. Since vegetation, soil and atmosphere have a memory, and because vegetation may require some time to adapt to environmental changes, it is necessary to explore potential lag-time responses to gain understanding of the relation between plants and their environment. While the concept of introducing lag times in the study of this relationship is not new (see e.g. [4]), it has become more extended in recent studies [2, 12]. Given the flexibility of our machine learning approach to incorporate a large number of climate-based predicting features, a large number of lag-times can be applied to the different climate variables. We experimented with time-lags covering a range of $\ell = 0, 1, ..., 12$ months (where $\ell = 0$ is the current month) for all the driving variables.

Cumulatives: Vegetation dynamics may not necessarily reflect the climatic conditions from (e.g.) three months ago, but the average of the (e.g.) three antecedent months. This integrated response to antecedent environmental and climatic conditions is referred here as 'cumulative' response. Note that, unlike in the case of lagged variables, cumulative variables include always up to present-time climate conditions.

Extreme indices: Over the last few years, many research studies have been performed on climate extremes [9, 17]. The fact that many daily data sets are freely available make the calculation of extreme indices easier. In recent years, 27 recommended indices related to temperature and precipitation have been developed [5, 15]. Related to the vegetation response, extremes on climate variables that cause a different behavior of the terrestrial ecosystems have been investigated [16].In our work, we have calculated different monthly indices on the raw data as well as on the residuals (including the trend - see Supplementary material[5]).

3 Experimental results

Putting all pieces of the above pipeline together, we end up with a dataset that has 5319 features generated on thirty-year time series with a monthly resolution.

We will analyze 13,097 land pixels independently, covering 130GB on disk. We use the implementation of scikit-learn for the random forest regressor, setting the number of trees equal to 100 and the maximum number of features per node to the square root of the total number of features. The model is assessed by means of five-fold cross-validation.

We apply ridge regression, Pearson correlation coefficients and a random forest-based Autoregressive model using as features the past values of the NDVI residuals (only the green moving window in Fig. 1) as baseline methods. We perform ridge regression using nested five-fold cross-validation for the tuning of the parameter λ on the same data sets. Pearson correlation coefficients are calculated on the training sets and the feature with the highest value is selected. Then, we calculate the squared Pearson correlation value of the selected feature on the test data.

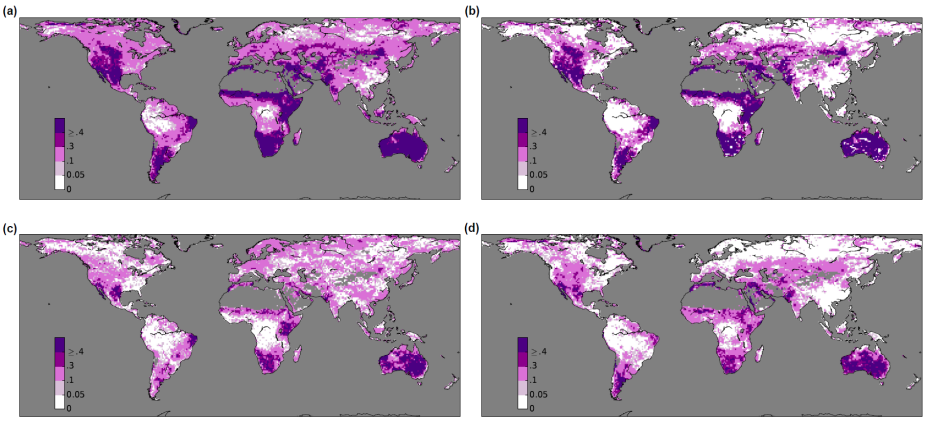

Fig. 2: Model performance. Explained variance by different families of models run with the common database of predictive features. (a) Random forest approach, (b) Ridge regressions, (c) Pearson correlation for the most correlated feature to the target variable per pixel, (d) Autocorrelations.

Figure 2, on the top, shows the result of the random forest model applied to the total number of features. As one can observe, the model has better predictive skill in Australia, in the bulk of Africa and in a portion of North and South America. Compared to the other three approaches, the results obtained by the random forest model are a big improvement. Ridge regression also performs well for the regions where the random forest yields a high R^2, in contrast to the other regions, where the map is mostly colored white. However, ridge regression in general leads to substantially worse results for almost all regions of the world. This result confirms that the relationships between the climatic variables are non-linear. The filter approach based on Pearson correlations, which is the current state-of-the-art in modelling climate-vegetation dynamics performs very

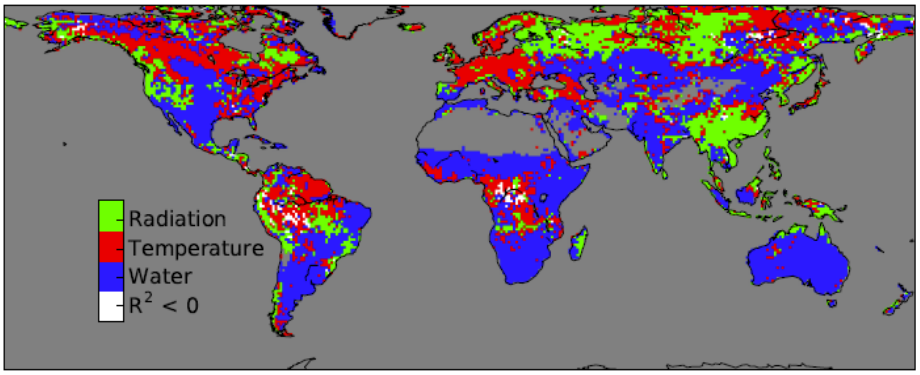

Fig. 3: Performance comparison of the three feature groups. **Blue:** Regions where water related features explain better the variance of the NDVI residuals. **Red:** Regions where temperature features outperform the other two groups. **Green:** Regions where radiation features give better result. **White:** All models give low R^2.

poorly. Finally, the Autoregressive model performs really low in almost all of the pixels. As such, we can conclude that an interplay between different variables is needed to model vegetation dynamics.

In Fig. 3 we classify each pixel to one of the three groups: temperature, radiation and water, according to the explained variance obtained from each group. It is clear that in regions where the performance of the total model is high, the water-related group of features give better results in comparison with the other two groups, radiation and temperature. In addition, in regions where radiation outperforms the other two groups, the performance of the total model is poor. Our results are quite consistent with the ones in [8], except for the result in rain forests. This is explained by the fact that NDVI is very constant in these regions, and thus none of the three groups (radiation, water, temperature) is, a priori, a clear winner for the explanation of this little NDVI variability that exists.

4 Conclusions

In this paper we presented a machine learning framework for detecting relationships in climate-vegetation dynamics. We used a wide collection of data and we created a unique database that bundles all publicly-available datasets with an appropriate spatial and temporal resolution. As such, we hope that this paper can inspire the machine learning and data mining communities to explore a new application domain with enormous potential for developing novel methods.

Our preliminary results are quite encouraging, and therefore in upcoming work the implications of our results from a climate and biophysical perspective will be examined. The framework that we propose also allows to answer other

questions w.r.t. climate-vegetation dynamics, such as the influence of lags, and the quality of the various data sources. We have already obtained interesting preliminary results in that direction, but those results are excluded due to lack of space.

References

1. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
2. T. Chen et al. Using satellite based soil moisture to quantify the water driven variability in NDVI: A case study over mainland Australia. *Remote Sensing of Environment*, 140:330–338, 2014.
3. R. B. Cleveland et al. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.
4. M.B. Davis. *Climatic instability, time, lags, and community disequilibrium*. Harper & Row, 1984.
5. M.G. Donat et al. Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The hadex2 dataset. *Journal of Geophysical Research: Atmospheres*, 118(5):2098–2118, 2013.
6. J. Hills et al. Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4):851–881, may 2013.
7. A. Mueen et al. Logical-shapelets: an expressive primitive for time series classification. In *Proc. of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 1154, New York, New York, USA, aug 2011. ACM Press.
8. R.R. Nemani et al. Climate-Driven Increases in Global Terrestrial Net Primary Production From 192 To 1999. *Science (New York, N.Y.)*, 300(5625):1560–3, 2003.
9. N. Nicholls and L. Alexander. Has the climate become more variable or extreme? progress 1992-2006. *Progress in Physical Geography*, 31(1):77–87, 2007.
10. C.J. Tucker et al. An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. *International Journal of Remote Sensing*, 26(20):4485–4498, 2005.
11. J. Verbesselt et al. Detecting trend and seasonal changes in satellite image time series. *Remote sensing of Environment*, 114(1):106–115, 2010.
12. D. Wu et al. Time-lag effects of global vegetation responses to climate change. *Global change biology*, 2015.
13. G. Wu et al. Wu et al. 2015 frontiers in plant science. may 2015.
14. L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 947, New York, New York, USA, jun 2009. ACM Press.
15. X. Zhang et al. Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, 2(6):851–870, 2011.
16. J. Zscheischler et al. A few extreme events dominate global interannual variability in gross primary production. *Environmental Research Letters*, 9(3):035001, 2014.
17. F.W. Zwiers et al. Climate extremes: challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. In *Climate Science for Serving Society*, pages 339–389. Springer, 2013.