BACI: Towards a Biosphere Atmosphere Change Index - Detection of extreme events in the biosphere.

Yanira Guanche García[1,2] YANIRA.GUANCHE.GARCIA@UNI-JENA.DE

Maha Shadaydeh[1]

Miguel Mahecha[2,3]

Markus Reichstein[2,3]

Joachim Denzler[1,2]

1 Computer Vision Group, Friedrich Schiller University Jena, Germany

2 Michael Stifel Center for Data-driven and Simulation Science Jena, Germany

3 Max Planck Institute for Biogeochemistry, Jena, Germany Address Line 2

Abstract

Technological developments from last decades offer unprecedented opportunities to monitor the Earth system. International research projects like BACI are joint efforts to provide free-of-charge, unified and high quality Earth Observations and the development of tools to analyze them. The ability to detect and monitor anomalous behaviour in multivariate environmental time series is crucial. These events are signals of changes in the underlying dynamical system and their detection can be used as an early-warning system for land ecosystems. In this study we present a methodology to detect these anomalies in biosphere data by a combination of a multivariate autoregressive model together with a distance measure. This work is framed within the EU-funded project BACI 'Detecting changes in essential ecosystem and biodiversity properties - towards a Biosphere Atmosphere Change Index'.

Keywords: Anomaly detection, Autoregressive model, Mahalanobis distance

1. Introduction

Space data archives and space-borne Earth observations play an essential role in monitoring ecosystems. Their transformations and responses to human interventions or climate extremes can now be studied in more detail than ever. This development is complemented by an increasing availability of a wide range of ground data. They cover many aspects of ecosystem functioning, structure, and other parameteres relevant to fully describe the functional biogeography of ecosystems. The BACI project aims to tap into the yet-to-be realized potential of existing and scheduled space-borne Earth observations.

The BACI consortium, formed by 10 institutions from 7 European countries consists of terrestrial remote sensing experts, experts in ecosystems, biodiversity and socio-economical modeling and observations of different types and experts on machine learning and big data processing techniques. Within the 9 work packages in which is divided the project, the WP5 is dedicated to the development of anomaly detection techniques that allow for detecting sudden events and abnormal changes in the multivariate Earth observation data streams. Combining different machine learning methods the main goal is to detect extreme events in historical data and therefore help defining those areas with higher amount of abnormal

records. This challenging data-driven task faces the added issue of the lack of ground truth events or contrasted events where it is known what happened and which were the causes.

2. Anomaly Detection in Biosphere Parameters

An abnormal event can be defined as those points within a time series that are not well represented by a previously fitted statistical model, Chandola et al. (2009). Following this intuitive concept, we propose a methodology based on a linear regression combined with a distance measure to detect extreme events in biosphere parameters. More precisely, after preprocessing the data, we combine a Multivariate Autoregressive Model (MVAR) with the Mahalanobis distance of the residuals between the model and the data to detect those points where the model and the data significantly differ and therefore can be considered as abnormal events.

Data from the Earth System Data Cube (ESDC) developed within the ESDL project has been used as the primary source of biosphere data for this study. The ESDC comprises spatiotemporal data consisting of: time, latitude, longitude and multivariate Earth Observations. The version used in this study covers the period from January 2001 to December 2012 with 8-daily observations and a spatial grid with a resolution of 0.25°. More than 30 biosphere and atmosphere parameters are included in this database. Out of these variables, we have used those 5 that mainly measure the terrestrial biosphere activities: Gross Primary Productivity (GPP), Latent Energy (LE), Net Ecosystem Exchange (NEE), Sensible Heat (SH) and Terrestrial Ecosystem Respiration (TER), which were kindly provided by the FLUXCOM initiative (Tramontana et al. (2016)). The study area comprises Africa and Europe (see Figure 1). This area was defined as the main study area within BACI.

2.1 Preprocessing

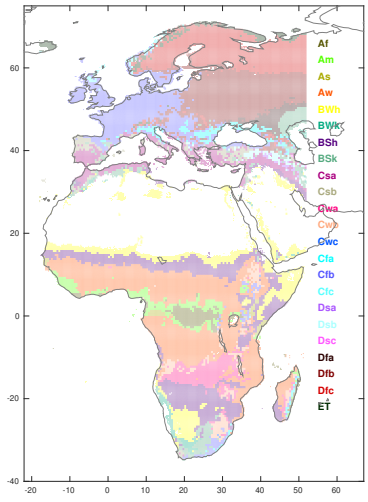
To avoid inconsistencies later, data needs to be pre-processed. We have applied techniques commonly used in environmental sciences. Additionally, to simplify computational load while the models fitting process the data was regionalized.

Deseasonalization and normalization: The mean seasonal cycle has been substracted from the variables. The remaining variables were then normalized by subtracting its mean, μ, and dividing by its variance, σ. This was done for all the 5 variables locally at each pixel of the grid.

Regionalization: The grid was clustered into regions of similar climate conditions according to the climate types defined by the Köppen Climate Classification (Chen and Chen (2013)). The Köppen Climate Classification is a widely used vegetation-based empirical clustering that divides the world in up to 31 climate regions. From these 31 climate regions, 23 are present in our study area. Figure 1 shows the climate regions with the legend explaining the codes that define them.

2.2 Multivariate Autoregressive (MVAR) Model

Let $x_i, i = 1, \cdots, N$ denote the time series of N Earth observation variables. Each time series $x_i(n), n = 1, \cdots, m$ is a realization of length m real valued discrete stationary stochastic process $X_i, i = 1, \cdots, N$. These N time series can be represented by a pth order multivariate



1^{st} letter	2^{nd} letter	3^{rd} letter
A: Tropical	f: fully humid	h: hot arid
B: Dry	m: monsoon	k: cold arid
C: Mild temperate	s: dry summer	a: hot summer
D: Snow	w: dry winter	b: warm summer
E: Polar	W: dessert	c: cool summer
	S: steppe	d: cold summer
	T: tundra	
	F: frost	

Figure 1: Area of study clustered according to the Köppen Climate Classification. Gaps represent areas where there is no data available.

autoregressive model (MVAR(p)) of the form

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_N(n) \end{bmatrix} = \sum_{r=1}^{p} A_r \begin{bmatrix} x_1(n-r) \\ \vdots \\ x_N(n-r) \end{bmatrix} + \begin{bmatrix} \epsilon_1(n) \\ \vdots \\ \epsilon_N(n) \end{bmatrix}, \tag{1}$$

The residuals $\epsilon_i, i = 1, \cdots, N$ constitute a white noise stationary process with an $N \times N$ residual covariance matrix Σ. The model parameters at time lags $r = 1, \cdots, p$ are defined by

$$A_r = \begin{bmatrix} a_{11}(r) & \cdots & a_{1N}(r) \\ \vdots & \ddots & \vdots \\ a_{N1}(r) & \cdots & a_{NN}(r) \end{bmatrix}. \tag{2}$$

For each climate region, a representative point that is geographically centered in the region and hence reflects its average behaviour, has been selected. The MVAR model order p was defined for every climate region, at each representative point, by means of a Bayesian Criterion (Schwarz et al. (1978)). Once the model order (p) was defined for each region we proceed with the entire grid, fitting an MVAR(p) model at each point.

2.3 Mahalanobis distance

The residual vector of the MVAR model is calculated as the difference between the model output and the real data for the five variables. The Mahalanobis distance (Mahalanobis (1936); Hotelling (1947)) of the residual vector is used as a measure of the deviation of the multivariate residuals at certain time step from their joint distribution. The Mahalanobis distance is defined in square unit as

$$d_m(\mathbf{E}) = (\mathbf{E} - \bar{\mathbf{E}})^{\mathbf{T}} \boldsymbol{\Sigma}^{-\mathbf{1}} (\mathbf{E} - \bar{\mathbf{E}}) \tag{3}$$

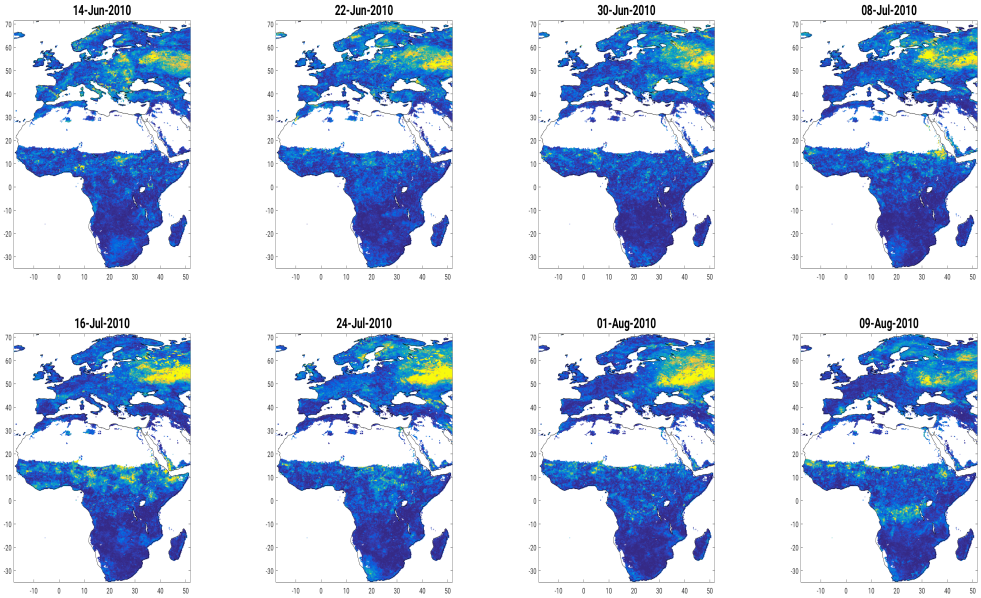


Figure 2: Heatwave in Russia, summer 2010.

where $\bar{\mathbf{E}}$ and $\mathbf{\Sigma}$ are the mean and covariance matrix of the multivariate residuals vector $\mathbf{E}$ respectively. The mean and the covariance were estimated considering the entire time series. This was the best way to do so in our case due to the short length of the time series used together with its coarse temporal resolution.

3. Results

Validating models that try to reproduce environmental processes is not a trivial task. There are no well defined ground-truth events which can be used to compare the models' performance and level of accuracy. With help of experts on the topic involved in the BACI project, some well known historical events that caused perturbations in the biosphere within the time span of our data were selected.

Figures 2-6 show the results obtained for some of the selected known historic events. Some events, such as the Russian heatwave in 2010 (Figure 2) and the drought in the horn of Africa in 2006 (Figure 3) are clearly detected by the method due to its magnitude and large temporal and spatial scale. Another particular event of interest is the volcanic eruption in the coast of the Red Sea in June 2011. This event is clearly detected despite its small spatial scale as shown in Figure 5. There are on the other hand some events such as the coldwaves (Figure 6) and cyclones (Figure 4) where the threshold for detections should be lower in winter due to the nonstationarity of the signals.

4. Conclusions

A new methodology to detect anomalies in biosphere time series has been described. Our approach comprises two main steps after preprocessing the data: a multivariate linear regression model combined with a distance measure. The combination of these techniques

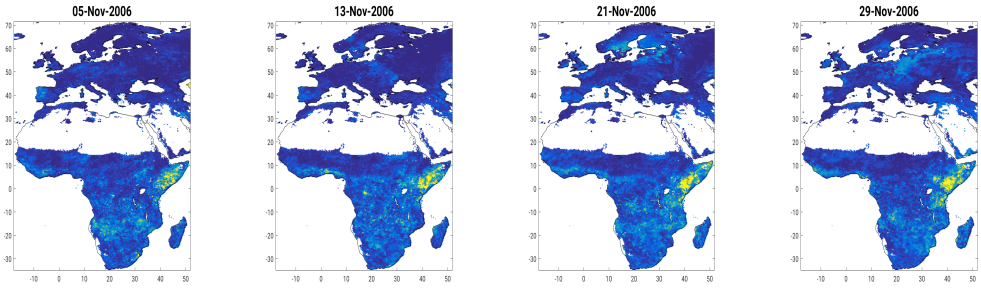


Figure 3: Drought in the horn of Africa, November 2006.

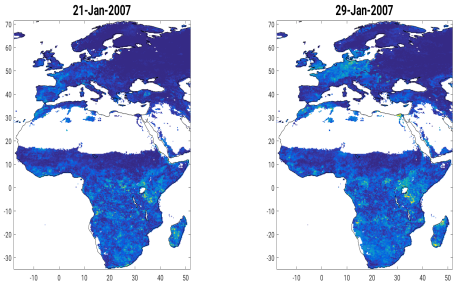


Figure 4: Cyclone in Central Europe, January 2007.

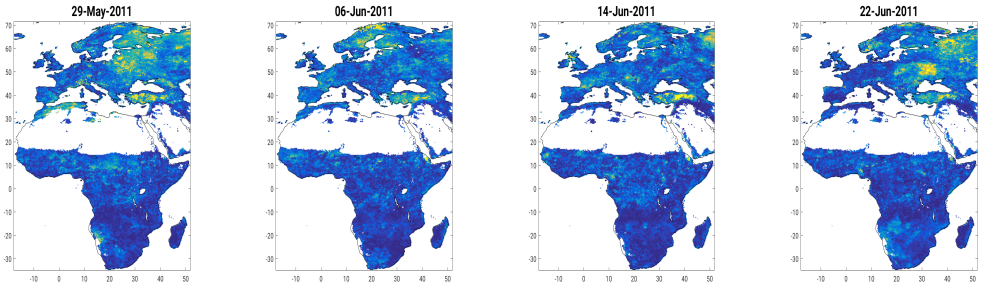


Figure 5: Volcanic eruption in the Red Sea coast, June 2011.

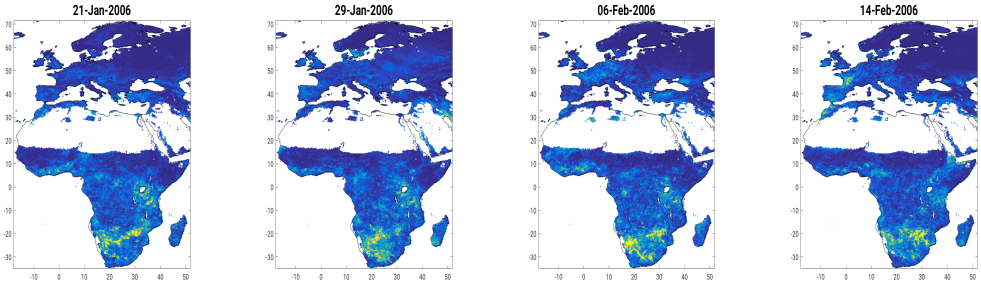


Figure 6: Coldwave in Central Europe, January-February 2006.

allow for the detection of abnormal events in the time series.

The proposed methodology has been applied to a large area that covers Europe and Africa. Results show that the method is able to detect the spatial and temporal extent of known historic events.

Acknowledgments

This study has been conducted within the framework of the project BACI: Towards a Biosphere Atmosphere Change Index, funded by the European Union's Horizon 2020 research and innovation program under the grant agreement No 640176.

References

V Chandola, A Banerjee, and V Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

D Chen and H W Chen. Using the Köppen classification to quantify climate variation and change: An example for 1901–2010. *Environmental Development*, 6:69–79, 2013.

H Hotelling. Multivariate quality control. *Techniques of statistical analysis*, 1947.

P Mahalanobis. On the generalised distance in statistics (vol.2, pp.49–55). *Proceedings National Institute of Science, India. Retrieved from http://ir. isical. ac. in/dspace/handle/1/1268*, 1936.

G Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

G Tramontana, M Jung, C R Schwalm, K Ichii, G Camps-Valls, B Ráduly, M Reichstein, M A Arain, A Cescatti, G Kiely, et al. Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, 13:4291–4313, 2016.