9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

ATTRIBUTION OF MULTIVARIATE EXTREME EVENTS

Yanira Guanche García[1,3], Maha Shadaydeh[2], Miguel Mahecha[3,4], Joachim Denzler[2,3]

***Abstract*—The detection of multivariate extreme events is crucial to monitor the Earth system and to analyze their impacts on ecosystems and society. Once an abnormal event is detected, the following natural question is: what is causing this anomaly? Answering this question we try to understand these anomalies, to explain why they happened. In a previous work, the authors presented a multivariate anomaly detection approach based on the combination of a vector autoregressive model and the Mahalanobis distance metric. In this paper, we present an approach for the attribution of the detected anomalous events based on the decomposition of the Mahalanobis distance. The decomposed form of this metric provides an answer to the question: how much does each variable contribute to this distance metric? The method is applied to the extreme events detected in the land-atmosphere exchange fluxes: Gross Primary Productivity, Latent Energy, Net Ecosystem Exchange, Sensible Heat and Terrestrial Ecosystem Respiration. The attribution results of the proposed method for different known historic events are presented and compared with the univariate Z-score attribution method.**

I. INTRODUCTION

The detection of multivariate extreme events is crucial to monitor the Earth system and to analyze their impacts on ecosystems and society. We expect that climate extremes such as droughts and heatwaves will increase as a consequence of climate change [1]. Hence, it is of a paramount importance to understand the drivers of such multivariate extreme events as well as the complex land-atmosphere-biosphere interactions, including those constellations that are not extreme for a single variable but are extreme for a combination of variables, also called compound event [1], [2]. In the last years several studies have gone in this direction, proposing different approaches: [2], [3], or [4] are just some examples. Unlike the different attribution methods proposed in the literature so far, the proposed method is suitable for data with low sampling rate where a pointwise detection and attribution can be applied.

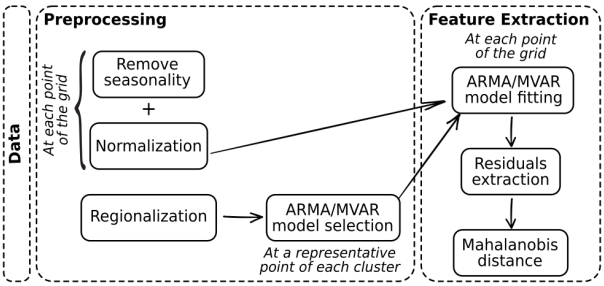

Fig. 1. Flowchart of the anomaly detection method using ARMA/VAR model(s) and Mahalanobis distance.

An abnormal event can be defined as those points within a time series that are not well represented by a previously fitted statistical model [5]. Following this intuitive concept, we have recently proposed a methodology based on linear regression models to detect extreme events in the biosphere [6],[7] (cf. Figure 1). More precisely, in [6] after preprocessing the data, we combine Autoregressive Moving Average Models (ARMA) with the Mahalanobis distance of the residuals between the models and the data to detect those points where the models and the data significantly differ and therefore can be considered as abnormal events. The method was further improved in [7] based on using a Vector Autoregressive (VAR) Model instead of multiple univariate ARMA models. The VAR model allows for presenting the variables with a model that takes into account their inter-dependencies and hence enables better whitening of the residuals and consequently better spatial and temporal detection accuracy of the anomalous events.

In this paper, we present an approach for the attribu-

Corresponding author: Y Guanche, yanira.guanchegarcia@dlr.de
[1] Institute of Data Science, German Aerospace Center, DLR, Jena, Germany [2]Computer Vision Group, Friedrich Schiller University, Jena, Germany [3] Michael Stifel Center for Data driven and Simulation Science, Jena, Germany [4]Max Planck Institute for Biogeochemistry, Jena, Germany

[1]https://www.ipcc.ch/

tion of multivariate anomalous events, where attribution here means to define the contribution of each of the variables involved to making the event an extreme one. The presented approach is based on the decomposition of the Mahalanobis distance of the residuals of the VAR model into components whereby each component presents the contribution of one of the used variables to the Mahalanobis distance. The decomposed form of this metric provides an answer to the question: how much does each variable contribute to this distance metric? The method is applied to the extreme events detected when using five land-atmosphere exchange fluxes (Gross Primary Productivity, Latent Energy, Net Ecosystem Exchange, Sensible Heat and Terrestrial Ecosystem Respiration). The attribution results of the proposed method for different known historical events are presented and compared with the univariate Z-score attribution method.

II. Anomaly Detection with vector autoregressive model and Mahalanobis distance

Let $x_i, i = 1, \cdots, N$ denotes 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 VAR model of the form

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

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

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

The steps of the anomaly detection method using the VAR(p) model in (1) are summarised in Figure I. After removing seasonality and normalizing the variables as two pre-processing steps, the data are clustered into climate regions according to the Koppen climate classification map [8]. Then for each climate region, a representative point that is geographically centered in the region has been selected. The VAR model order p was defined for every climate region, at each representative point, by means of a Bayesian Criterion [9]. Once the model order p is defined for each region, we proceed with the entire grid, fitting a VAR(p) model for each point in the grid. The residual vector $\mathbf{E}$ is calculated as the difference between the estimated VAR model output and the real value of the used variables. The Mahalanobis distance [10], [11] of the residual vector is then used as a measure of the deviation of the multivariate residuals at a certain time step from their joint distribution. The Mahalanobis distance is defined in the square unit as:

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

where $\bar{\mathbf{E}}$ and $\boldsymbol{\Sigma}$ 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.

When the value of the Mahalanobis distance of the residuals is large, it is assumed that something abnormal occurs in the system and the model is not able to correctly capture it. The easiest way to detect abnormal events is to set a fixed percentile threshold and look for the points with Mahalanobis distance surpassing this threshold. The reader is referred to [6], [7] for further details on different multivariate anomalous event detection methods.

III. Attribution Scheme based on Mahalanobis Distance Decomposition

Once an anomalous event is detected, the next natural question is: *which variables are causing this anomaly?* An intuitive approach to answer this question is to decompose the value of the Mahalanobis distance into components, whereby each component quantifies the contribution of one of the variables to the distance. Garthwaite and Koch [12] recently proposed the corr-max transformation for the decomposition of the Mahalanobis distance which can be easily implemented and provides helpful results from an attribution point of view. The decomposition has the form:

$$d_m(\mathbf{E}) = \mathbf{W}^T \mathbf{W}, \quad (4)$$

where $\mathbf{W} = (W_1, \cdots, W_N)^T$ is a vector with N elements, corresponding to the N variables contributing to the Mahalanobis distance $d_m(\mathbf{E})$, and is calculated by:

$$\mathbf{W} = (\mathbf{S}\Sigma\mathbf{S})^{-1/2} \mathbf{S}(\mathbf{E} - \bar{\mathbf{E}}), \quad (5)$$

where $\mathbf{S}$ denotes a diagonal matrix of the inverses of the standard deviations of the variables of $\mathbf{E}$. The

components of **W** should be uncorrelated, with the transformation chosen to maximize the sum of correlations between the corresponding elements of **S** and **W**.

IV. Experimental Results and Discussion

Data from the Earth System Data Cube (ESDC)[2] developed within the ESDL project has been used as the primary source of land-atmosphere exchange fluxes 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 five 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[3] initiative [13], [14]. The study area comprises Africa and Europe (see Figure 2). This area was defined as the main study area within the European project BACI: Towards a Biosphere Atmosphere Change Index[4] which is the framework of the current study.

The proposed method is applied for the attribution of two known historic events: the Russian Heatwave of July 2010 and the Drought that affected the Horn of Africa in November 2006. Attribution results of other historic events can be found in [7]. The definition of the temporal and spatial extension of these events was supported by socio-climate experts from the BACI project and is out of the scope of this study.

The results of the proposed method are compared with the univariate Z-score results [15] applied to the same historic extreme events. The Z-score is a measure that compares the distribution of a certain variable within the temporal extent of the detected anomalous event with the distribution of this variable in the entire time series. The Z-score quantifies the discrepancy between these two distributions. This is done separately for each variable. High positive or negative values of the Z-score indicate which variables are most different from their normal behaviour within the time duration of the event.

Figures 2 and 3 show the results obtained for the Mahalanobis distance decomposition and the Z-scores for the two known historic events. Each figure shows the spatial extension of the event (upper left subplot), the Z-scores (upper subplots) and the Mahalanobis distance decomposition (lower subplots). The Z-score subplots show the histograms of the five variables within the time window of the event (red) and the entire time series (grey) together with the value of the Z-score obtained from the comparison of both histograms. The lower subplots show the Mahalanobis distance intensity (map on the left) presenting the spatial extent of the detected anomalous event, in addition to five other maps, each one shows the contribution of one of the used variables to the detected events.

According to the Mahalanobis decomposition using the corr-max transformation, the Russian Heatwave (Figure 2) is most dominantly manifested in LE then in SH. These results of the Mahalanobis decomposition are coherent with the analysis proposed by other authors, [2], [4]. For the case of the Drought in the Horn of Africa, (Figure 3) GPP and NEE, are the most contributing ones. Validating the attribution results of a drought are not trivial since these are very long events where several factors are involved. The Z-score analysis sorts the contribution of the variables differently. These discrepancies between the Z-score approach and the proposed multivariate approach show the relevance of performing a multivariate analysis and the limitations of univariate approaches in such complex systems like biosphere or climate systems. Unfortunately, a detailed quantitative evaluation of the performance of the proposed method is not possible due to the lack of ground truths for the attribution of the extreme events considered in this study.

[2]//http://www.earthsystemdatalab.org/
[3]//http://www.fluxcom.org/
[4]//http://www.baci-h2020.eu

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

[1] M. Reichstein, M. Bahn, P. Ciais, D. Frank, M. D. Mahecha, S. I. Seneviratne, J. Zscheischler, C. Beer, N. Buchmann, D. C.

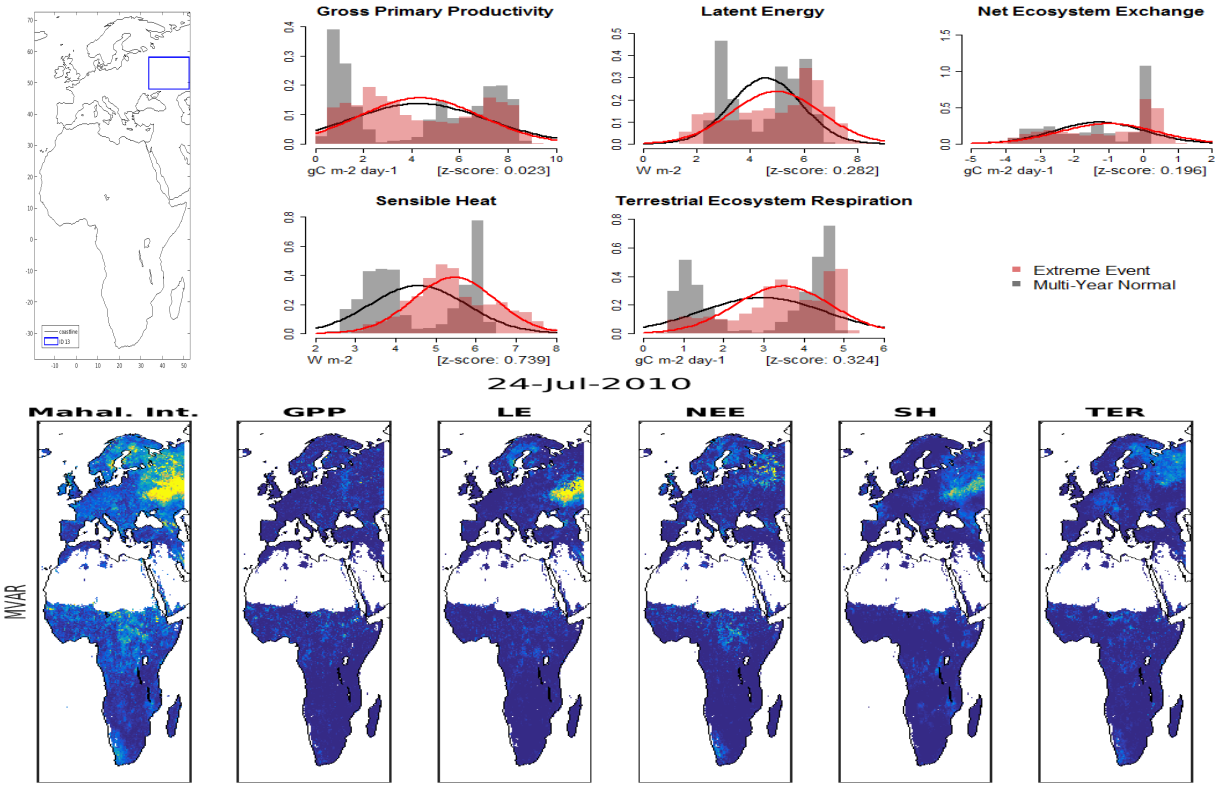

Fig. 2. Attribution scheme for the Russian Heatwave from July 2010. Upper left plot: spatial extent of the Heatwave (blue rectangle), upper right plots: Z-score for the five variables involved, lower plots: Mahalanobis intensity (left) and its decomposition into the five used variables .

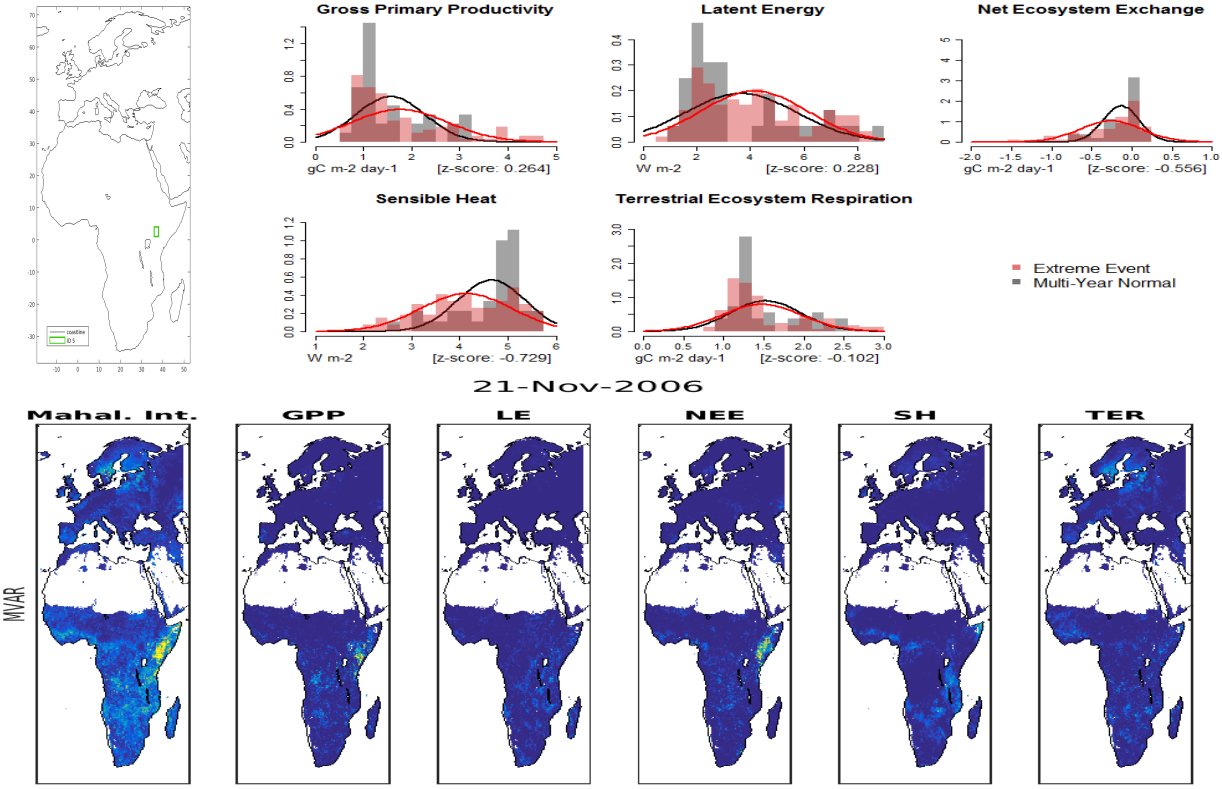

Fig. 3. Attribution scheme for the Drought in the Horn of Africa from November 2006. Upper left plot: spatial extent of the Drought (green rectangle), upper right plots: Z-score for the five variables involved, lower plots: Mahalanobis intensity (left) and its decomposition into the five used variables.

Frank, *et al.*, "Climate extremes and the carbon cycle," *Nature*, vol. 500, no. 7462, p. 287, 2013.

[2] D. G. Miralles, A. J. Teuling, C. C. Van Heerwaarden, and J. V.-G. De Arellano, "Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation," *Nature geoscience*, vol. 7, no. 5, p. 345, 2014.

[3] J. Zscheischler, M. Reichstein, S. Harmeling, A. Rammig, E. Tomelleri, and M. D. Mahecha, "Extreme events in gross primary production: a characterization across continents," *Biogeosciences*, vol. 11, no. 11, pp. 2909–2924, 2014.

[4] M. Flach, S. Sippel, F. Gans, A. Bastos, A. Brenning, M. Reichstein, and M. Mahecha, "Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western russian heatwave," *Biogeosciences*, vol. 16, pp. 6067–6085, 2018.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[6] Y. Guanche, M. Shadaydeh, M. Mahecha, and J. Denzler, "Extreme anomaly event detection in biosphere using linear regression and a spatiotemporal mrf model," *Natural Hazards*, pp. 1–19, 2018.

[7] M. Shadaydeh, Y. Guanche, M. Mahecha, and J. Denzler, "Baci deliverable 5.4: Methods for attribution scheme and near real-time baci." http://www.baci-h2020.eu/index.php/Outreach/Deliverables, 2018.

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

[9] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[10] P. Mahalanobis, "On the generalised distance in statistics (vol.2, pp.49–55)," *Proceedings National Institute of Science, India.*, 1936.

[11] H. Hotelling, "Multivariate quality control," *Techniques of statistical analysis*, 1947.

[12] P. H. Garthwaite and I. Koch, "Evaluating the contributions of individual variables to a quadratic form," *Australian & New Zealand journal of statistics*, vol. 58, no. 1, pp. 99–119, 2016.

[13] 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*, vol. 13, pp. 4291–4313, 2016.

[14] M. Jung, S. Koirala, U. Weber, K. Ichii, F. Gans, G. Camps-Valls, D. Papale, C. R. Schwalm, G. Tramontana, and M. Reichstein, "The fluxcom ensemble of global land-atmosphere energy fluxes," in *Scientific Data*, 2018.

[15] J. Reiche, J. Balling, M. Herold, M. Niedertscheider, K. Erb, M. Urban, and C. Schmullius, "Baci deliverable 6.2: Product comparison and validation report." http://www.baci-h2020.eu/index.php/Outreach/Deliverables, 2018.