Multivariate Extreme Event Detection on ESDL Colombian data cube

Maximiliano Garavito Chtefan,* María Daniela Lizarazo Sandoval,† and Bjorn Reu

Escuela de Ingeniería de Sistemas e Informática,

Universidad Industrial de Santander

Javier Alejandro Acevedo Barroso‡

Departamento de Física, Universidad de los Andes

(Dated: July 25, 2019)

Abstract

It is known that climate change is making extreme weather more frequent and it is of utmost importance understanding the causes and consequences of those events, so that it becomes possible to predict when they are going to happen, how to mitigate them, how to minimize damage and how to help recover the ecosystem, specially in a country with such rich diversity as Colombia. For this first, it is necessary to know how to extract and define what an extreme event is in the data and understand to what weather conditions they are associated to in real life. In this study, we focus on the detection of extreme events, and the algorithms for the preprocessing of such complex data. In order to do so, we relied on the Earth System Data Lab Cube, which is a real multivariate Earth system data stream. Our workflow starts with seasonality reduction and standarization for feature extraction for variables separately, and concludes with the implementation of KDE for anomaly detection. This method is based on Multivariate anomaly detection for Earth observations: a comparison (M. Flach et al., 2017) and Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western Russian heatwave (M. Flach et al., 2018). Since the data provided comes from real sources there were extra considerations to take and application of additional techniques (Fast Fourier Transform) for cleaning and pre-processing, which also are included and explained.

I. INTRODUCTION

During 2010-2011 Colombia was one of the most affected countries by extreme weather due to the raining phenomen called La Niña, according to European NGO, Germanwatch [1], and the opposite drought event called El Niño. El Niño Southern Oscillation, or ENSO, involves large-scale patterns of pressure, temperature, precipitation and winds that affect weather and climate over both the tropics and, through teleconnections, regions outside of the tropics, while during La Niña events, the trade winds resume and intensify, which increases the equatorial upwelling and extends the colder water of the east Pacific westward. The resulting effects are essentially the reverse of El Niño: the western Pacific experiences higher than average precipitation, and the flooding that results can be severe if it follows extended drought brought on by El Niño [2]. Therefore, the importance of early detection and analysis of extreme events.

In order to understand how an extreme event works, it is fundamental to know the context or background where it occurs, and the system where it came from. All systems have

* hipermaximus@gmail.com

† danimari98@hotmail.com

‡ ja.acevedo12@uniandes.edu.co

subsystems, and our planet is not the exception. Each of these subsystems can be monitored and characterized by multiple variables [3]. Being the Earth such a complex entity, getting and organizing these variables is a difficult task; nevertheless, thanks to satellites and signal processing techniques, this data is nowadays avalible as a multidimensional structures called Data Cubes. The Earth System Data Lab is a multivariate dataset of essential Earth System variables on a common grid and sharing a common data model [3]. It provides access to a series of highly curated "data cubes", in particular, the Colombian Data Cube, which is the focus of our study. For this task, a workflow proposed by Flach *et al.* [3] in 2017, along with some methods originally used to study the Russian Heatwave in 2018 by Flach *et al.* [4]. We also adapted some of the work done by Mahecha *et al.* [5] detecting the impacts of extreme events. The final aim is to relate those extreme event detections to real life events as El Niño or La Niña. This report is organized as follows. In section II we discuss our initial exploration of the data and the preprocessing used. In section III we explain the workflow used to process the data and detect extreme events. Finally, we offer concluding remarks on IV.

II. DATA EXPLORATION AND PRE-PROCESSING

Initially, we worked with the global and colombian cubes on its early versions, then we moved to version 2.0.1 of the Colombian Data Cube with 0.083 degrees resolution. This cube has 95 variables; however, during the initial exploration of the data we found that most variables were riddled missing data and *NaNs*. Also, there was no data from 2015-12-31 onwards. To explore the features space, we wrote a null exploration function: *explore_nulls*, which for a certain area centered in a fixed latitude and longitude, returns the area to explore, along with the feature names of the varibles with less than threshold percentage of nulls.

Some of the viable features found during the initial exploration were: evapotranspiration, precipipation chirps, enhanced vegetation index terra, fapar, ground primary productivity, precipitation trrmm, directional hemispheric reflectance shortwave and infrared. A plot of some of them is available in Figure 1. We developed the workflow using mainly *gpp* and *evotranspiration*.

Since the number of variables was reduced, it was not necessary to apply PCA, instead an individual anomaly detection was proposed for each variable and then taking into account the correlation matrix, a look up for relation between extreme events in related variables.(Figure 3)

Furthermore, while plotting a dynamic video that allowed to see the variabilty of each variable over time, it was noticeable that even thought Colombia doesn't have delimited seasons like other countries a certain seasonality stood up, which needed to be reduced in order to do the feature extraction. In consideration of the above, it was stablished the workflow.

III. WORKFLOW

As it was mentioned it before, multiple functions were implemented to facilitate and reduce redundance in the workflow. Therefore, they were ensambled it in a Toolbox for pre-processing the ESDL cube data.(Figure 4)

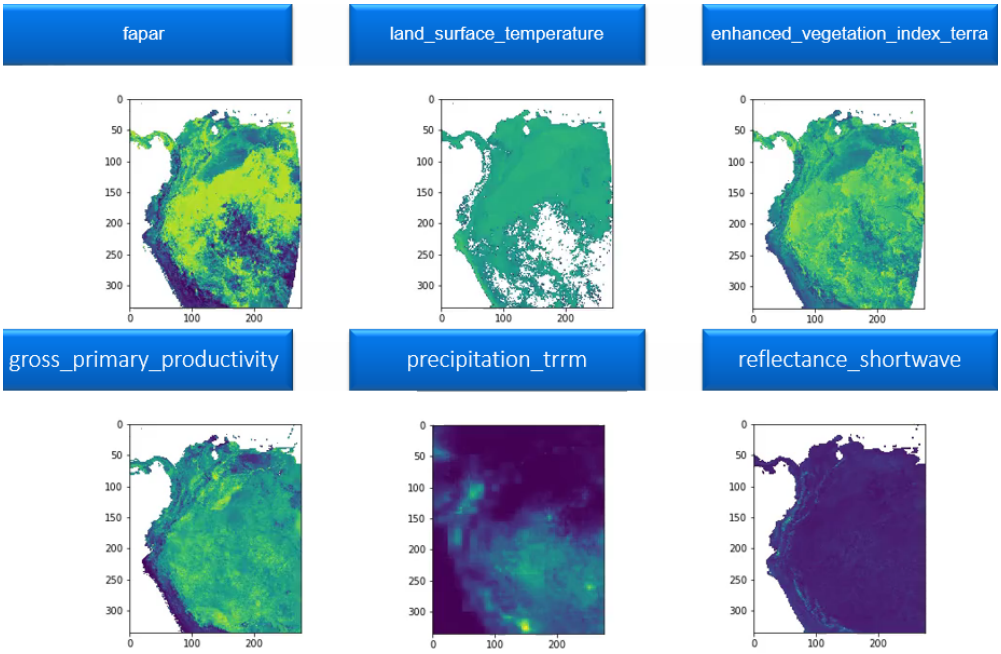

FIG. 1. Some of the viable features to used during the initial exploration.

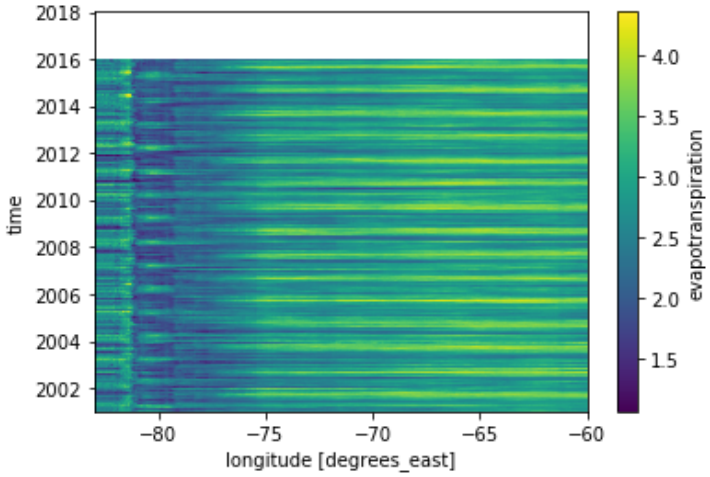

FIG. 2. Evapotranspiration plot over time at different longitudes. Note the lack of data from 2015-31-12 onwards, and the anual trend on the easternmost degrees.

A. Standarization and seasonality reduction

A set of functions were implemented for doing the Fast Fourier Transform over. These functions take the data given by a Cube and return another Cube with the time coordinate changed by frequency and values of the absolute values of the Fourier coeficients for each variable and for each longitude and latitude. To apply the FFT we first standardize the data to get rid of biases in the coefficients (e.g. the first coefficient would likely be the biggest if the mean is not 0 but this is not representative of seasonality but just the value of the mean

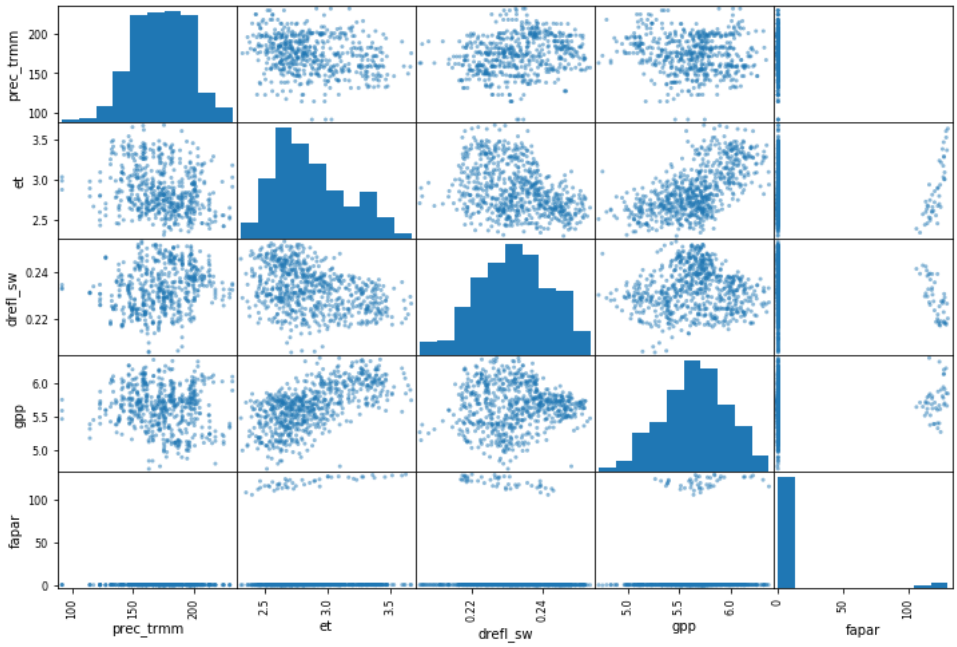

FIG. 3. Correlation matrix.

of the time series).

To eliminate the various frequencies that represent the seasonality it was necessesary to do tests on how to choose those biggest coefficients. The most naive approach was to select the biggest N coefficients but this would mean that variables with less seasonality would lose information that isn't associated to its seasonality, whilst at the same time variables with a lot of seasonality would not be properly cleaned of it.

Other approaches were tried by the idea to eliminate the first biggest coeffcients that contribute to a threshold percentage of the sum of the signal (e.g. the biggest N coefficients that contribute the T% of the sum of the time series); another equivalent approach was to choose the biggest coefficients by how many standard deviations they were from the mean.

It is necessary to take into consideration up until what point does it make sense to talk about seasonality, as the Nyquist frequency is just 4 days. So it might be necessary to exclude off some possible frequencies from the seasonality reduction if the frequency of the selected high coefficient is too high.

Other ways via machine learning could be advisable to implement. Since we are already using KDE to do the anomaly detection,it would be possible to repurpose it to also clean the time series of the seasonality.

B. KDE for anomaly detection

Currently, we are working on the implementation of this algorithm in Python, taking into account the already implemented functions in Julia on the cube [3]. However, if the results are not optimal, we rely on the interoperability between Python and Julia to apply the existing anomaly detection functions.

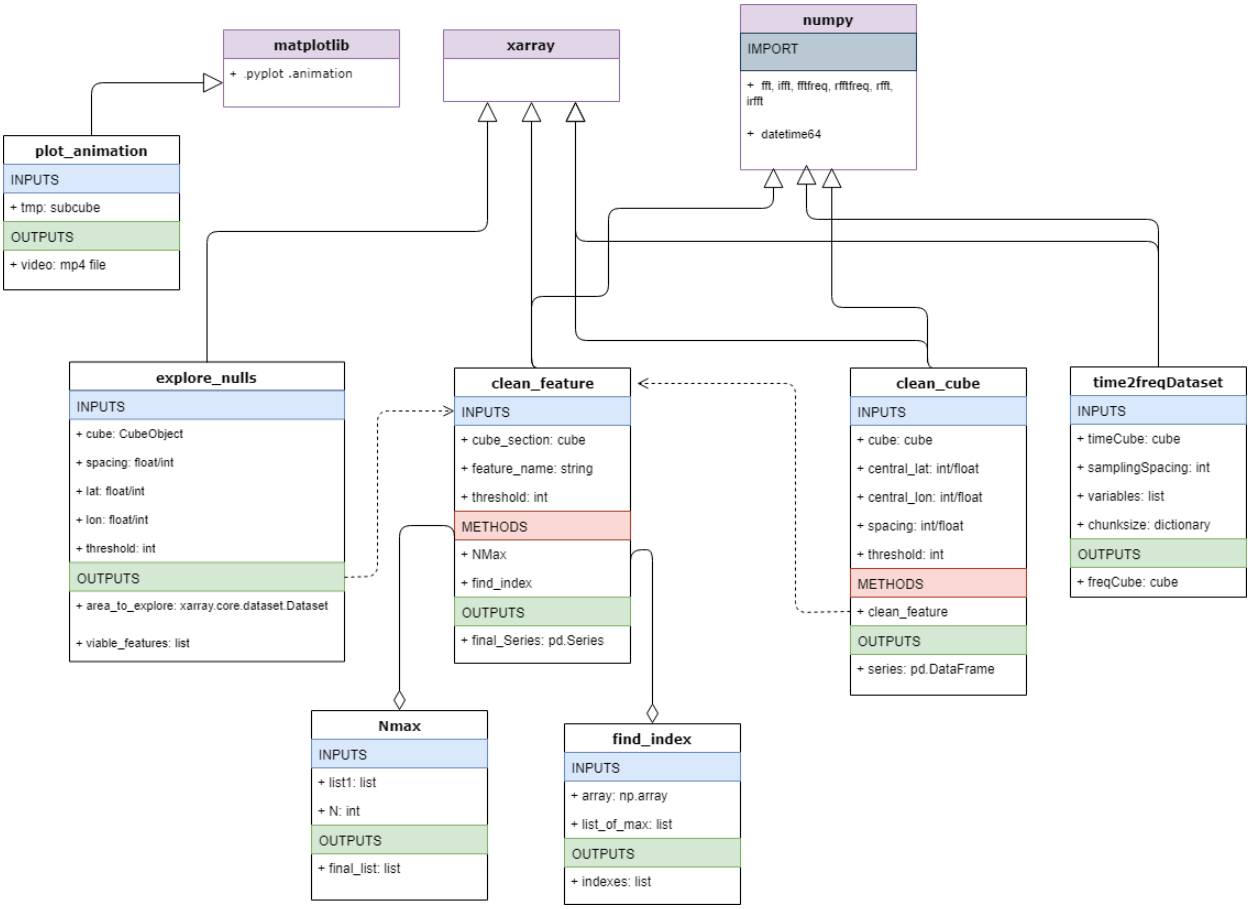

FIG. 4. "Toolbox class diagram."

IV. CONCLUDING REMARKS

The use of the Colombian ESDL Cube allowed to define in an abstract way which was the seasonality by zones regarding a certain geography and time. Thanks to this, the characterization of the seasonality was done not by one general periodicity but in terms of a number of periods and frequencies, being El Niño one of those. Therefore, it was possible to conclude that the behavior of seasonality found in Colombia and the ecuatiorian differs from the usual found in other countries, and this findings and its results are fundamental for the next step of anomaly detection.

A. Biggest struggles using the ESDL Cubes

The work was initially started in the Julia programming language but it was noticed that for a more high-level analysis and workflow it would be faster to work on Python.

After exploring the cube it became clear that there was a good chunk of values missing so we had to select those features and dates where the data had the best integrity. In particular the cube 'CUBE_V2.0.1_colombia_time_optimized_0.083deg' had good integrity between the beginnig date of 2001-01-05, until 2015-12-31, whilst the cube had dates up to 2017-12-31.

In python the main struggles were dealing with Xarray itself as grouping by multiple

coordinates is not fully supported and it would have been easier to group by latitude and longitud together.

---

[1] T. Hinchliffe, Colombia among top countries most affected by 'extreme' weather, Colombia Reports (2011).

[2] I. R. I. for Climateand Society, Why do we care about el niño and la niña, International Research Institute for Climate and Society.

[3] M. Flach, F. Gans, A. Brenning, J. Denzler, M. Reichstein, E. Rodner, S. Bathiany, P. Bodesheim, Y. Guanche, S. Sippel, and M. D. Mahecha, Multivariate anomaly detection for earth observations: a comparison of algorithms and feature extraction techniques, Earth System Dynamics **8**, 677 (2017).

[4] M. Flach, S. Sippel, F. Gans, A. Bastos, A. Brenning, M. Reichstein, and M. D. Mahecha, Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western russian heatwave, Biogeosciences **15**, 6067 (2018).

[5] M. D. Mahecha, S. Sippel, F. Gans, J. F. Donges, T. Kaminski, S. Metzger, M. D. Migliavacca, D. Papale, A. Rammig, and J. Zscheischler, Detecting impacts of extreme events with ecological in situ monitoring networks, Biogeosciences **14**, 4255–4277 (2017).