EARLY ADOPTERS: REPORT

A PREPRINT

J. Emmanuel Johnson
Information and Signal Processing
Universitat de Valencia
Valencia, Spain
juan.johnson@uv.es

Maria Plies
Information and Signal Processing
Universitat de Valencia
Valencia, Spain
maria.plies@uv.es

Valero Laparra
Information and Signal Processing
Universitat de Valencia
Valencia, Spain
valero.laparra@uv.es

Jose Antonio-Padron
Information and Signal Processing
Universitat de Valencia
Valencia, Spain
j.antonio.padron@uv.es

Gustau Camps-Valls
Information and Signal Processing
Universitat de Valencia
Valencia, Spain
gcamps@uv.es

July 19, 2019

ABSTRACT

The Earth is a complex dynamic and networked system. For the past few decades, a large number of extreme weather and climate events in Europe and worldwide have occurred which resulted in infrastructure damages and many casualties. Agricultural drought and wildfires are examples of some of the most critical hazards in terms of frequency, severity and impact on livelihoods. Detecting such extremes and anomalies is of paramount relevance to mitigate impacts and incorporating prevention measures. However, the phenomena are difficult to predict as they are complex and depend on many factors. A vast suite of approaches have been developed to monitor and characterize e.g. agricultural drought, based on either climatic ground-based data, soil moisture data or a variety of remote-sensing drought proxies. A recently proposed Soil Moisture Agricultural Drought Index (SMADI) is a simple and intuitive index that determines agricultural drought events based on key remote sensing indicators: land surface temperature (LST), vegetation indices (e.g., the NDVI) and surface soil moisture (SSM). While indices like SMADI and other alternative hand-crafted indices have been widely adopted and used in real practice yielding good results, they often ignore the complex nonlinear and multidimensional variable relations in the problem. In recent years, statistical machine learning (ML) has played a role in Earth observation data problems with positive results in problems like classification and anomaly detection. Machine learning can actually cope with multivariate and multiple source data sources and allows one to automatically detect anomalies. There is a plethora of ML algorithms for anomaly detection (AD), ranging from simple histogram-based models to more advanced hierarchical density-based clustering algorithms. Each has its advantages and disadvantages, but often need the help of user expertise and process understanding to improve results. In this work we introduce the application of a hybrid approach based on state-of-the-art ML anomaly detection methods and standard drought indices like SMADI for drought detection. We will illustrate the performance in the Earth System Data Lab (ESDL), a light-weight platform for Earth observation data analysis in the cloud. The ESDL allows to evaluate algorithms in a wide range of harmonized products including more than 40 variables spanning more than 10 years. The included variables account for atmospheric conditions, climate states, and terrestrial biosphere. Extracting anomalous events is possible with the multivariate spatial-temporal information contained within the Earth datacubes using advanced hybrid ML modeling. Algorithms will be compared in terms of accuracy, robustness and computational efficiency in selected examples of droughts in Europe during the last decade.

***K*eywords** Drought Detection · Machine Learning · Anomaly Detection

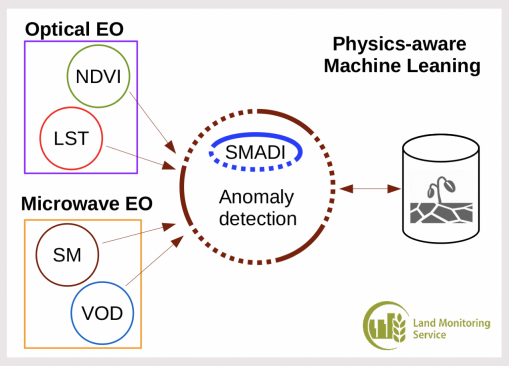


Figure 1: Physics-Aware Machine Learning

1 Introduction

According to climate projections, extreme events are likely to increase in frequency and intensity over the coming years [1]. In addition to frequency and intensity, some extreme events (such as droughts) are characterized by its area of impact as well as its effect on society. There are many studies that incorporate optical and microwave Earth observation (EO) data in drought detection for operation agricultural monitoring services [2]. Basic variables such as LST and NDVI derived from optical Earth observation and soil moisture (SM) and vegetation optical depth (VOD) [3] derived from microwave Earth observation are just a few of the many features that have been used in the context of drought detection. In addition, we have some newer indices such as the Soil Moisture Agricultural Drought Index (SMADI) [4] which is used as a global estimator specifically for agricultural drought. This is an index that uses the combination of NDVI, LST and SM that was assessed and validated versus known agricultural drought indices; e.g. the Soil Water Deficit Index (SWDI) [5] and the Crop Moisture Index (CMI) [6]. An intelligent combination of all of these variables (e.g. SMADI) provides a great opportunity to really capture the relationships and help the agricultural community with drought warning systems.

Like SMADI, often the inclusion of these variables was found to perform much better, was physically motivated and included intensity factors. However the calibration process can be lengthy and may not have captured all relationships between the data. With more data at higher spatial and temporal resolutions, this calibration process can hinder progress and may not be robust enough to the large amounts of data streams. With regards to the data accessibility, we have three important components in the modern age of 'Big Data': a) we have a huge volume of data which increases as much as terabytes and petabytes daily; b) there is a wide variety of data that is accumulated from many different sources like active sensing and ground measures at different spatial, spectral and temporal resolutions; and c) the speed at which we accumulate data is increasing and sometimes on the order of seconds requiring many fast and efficient pre-processing and storage mechanisms. The Earth Science Data Lab (ESDL) [1] is an upcoming platform that provides an opportunity for data centric processing methodologies as it is a central platform with easy access to multiple streams of data. That coupled with cloud computing facilities enables us to prototype many approaches with a central environment for processing and sharing of results. Our target application is drought detection whereby we use the previously mentioned variables. However, we would like to approach the problem from a different perspective, that of machine learning.

$$\mathcal{X} = \mathbf{x}(u, v, t, z)$$

where u is the longitude, v is the latitude, t is the time and z is the variable.

There are many applications nowadays that require data analysis to filter outliers ranging from credit fraud in banking to extreme events in Earth observation data. In machine learning, we define *anomalies* as observations which deviate sufficiently from the likely trend where we assume were generated by a different process than the norm. The observations themselves, we can define as *outliers* when they are sufficiently numerous than the standard distribution; e.g. 5%-10%. One has to be careful because in Machine learning (ML) algorithms as there is always some noise, ϵ associated with the learning procedure. So distinguishing between real data, noisy data, and outlier data can be a difficult task. Extreme events present a challenging task for ML as the amount of extreme labels present are often much less than the number of non-extreme labels which poses a difficult class imbalance problem [7]. Furthermore, the definition of extreme is rather

[1] `https://www.earthsystemdatalab.net/`

arbitrary and there exists no universal standard. In society, we typically measure an extreme event by the impact it has on society however physically this meaning can be different which poses a challenge for machine learning algorithms. On the other hand, the unsupervised family of algorithms make use of unlabeled data to assign a score to each sample and how it compares to the *normality* of the data distribution. We also specifically are looking for droughts which is one type of extreme in the hypothesis space of all extreme events. Therefore to obtain optimal results, it would be best to include known information (labels) about the drought events or a trusted proxy (SMADI) to help aid the results (figure 1). In this paper, we focus on utilizing unsupervised learning algorithms and comparing how the results are similar to the SMADI drought index. The SMADI index has been shown to be effective at characterizing droughts including their intensity. If there is indeed a link between the SMADI outcomes and the unsupervised learning methods, we hope to use the SMADI index as either an empirical feature or as a constraining factor to narrow the extreme detection scope to only droughts.

2 Methods

This section gives a general overview of the datasets, variables and propsoed methodologies used to answer the following research question: **can unsupervised learning ML methods capture drought events?**

2.1 Study Areas

We emulated the study area found in the original datasets for applying SMADI [8]. In this paper, we display preliminary results for two study areas: the contingious US area (CONUS) and the area of Russia for the time frame of 2010-2016. The CONUS area has many drought events of varying intensities where the most intense events occurring towards the years 2013-2016. The Russian drought has a major drought event that occurs for the year 2012 which was the target characterization; in addition there was a severe heatwave in 2010 [9].

2.2 Variables

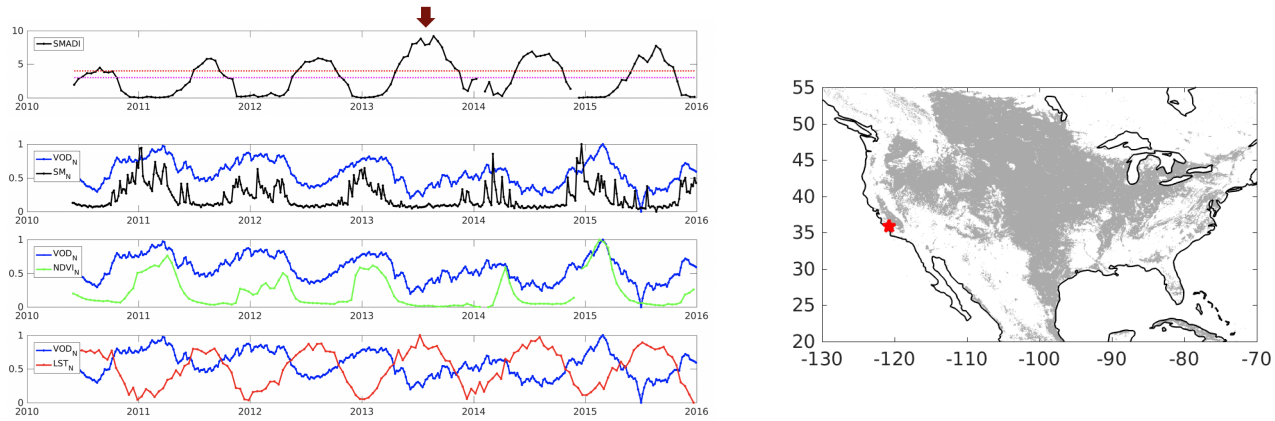
We selected four parameters related to the aspects of the drought: they include land surface temperature (LST), the normalized difference vegetation index (NDVI), the surface soil moisture (SM) and the Vegetation optical depth (VOD). The LST, NDVI and SM are variables that were used to effectively capture drought in the original study [4] and combined, these variables have been shown to effectively capture drought events. We use the Soil Moisture Agricultural Drought Index (SMADI) index as an assessment (proxy) variable to assess how well the machine learning algorithms performed. SMADI itself captures the lag between soil moisture conditions and plant response and it is scalable in space and time as it is a linear operation to calculate. It only utilizes the same three variables mentioned previously by empirically capturing the soil moisture deficit, the thermal stress and the unhealthy vegetation. As mentioned in the introduction, the previous study [4] found that it performed well compared to the Crop Moisture Index (CMI) and the Soil Water Deficit (SWD) [5] so it is trusted in the soil moisture community. The VOD comes from the L-Band microwave emissivity from the daily SMOS sensor [3, 10]. It measures attenuation due to canopy biomass and water content which offers the following advantages: a) it is distinct from greenness offering more information, b) it is not affected by atmospheric conditions/clouds, and c) it is sensitive to biomass and water-uptake dynamics. The sub-hypothesis is that **L-Band VOD captures drought-induced crop water stress and thus will aid in the detection of drought events for the unsupervised Machine Learning algorithms**.

We normalized all datasets (SM, LST, NDVI, and VOD) and then did a cubic interpolation for all of the signals to fill in the NAN values. As a simple preliminary experiment, we chose three time stamps as features (three 8-day cycles - 24 days in total). So every observation assumes we know the previous three observations in time (not space). We also used the entire time-frame as inputs (2011-2016). We applied a Savitzky-Golay filter with a time window length of 5 for the VOD dataset. We performed the ML algorithms with just the three original datasets (SM, LST, NDVI) and also performed the ML algorithms on the same three original variables with the inclusion of VOD (SM, LST, NDVI, VOD). The results are summarized and compared in the next section.

2.3 Machine Learning Integration

In this section, we elaborate on the tools used to expand upon the inclusion of just VOD and SMADI. We chose use two unsupervised clustering methods to perform anomolous detection for drought events.

The first algorithm is a simple Local Outlier Factor (LOF) [11] which is a well-known distance based approach. Given some data point $\mathbf{x}$, the LOF algorithm computes some outlier score $d(\mathbf{x})$ based on the Euclidean distance d between $\mathbf{x}$ and its k^{th} nearest neighbour. The scoring takes into account each of its n neighbours. Those points that have less



(a) (b)

Figure 2: California yearly drought for 2011 - 2016. (a) shows a group of averaged pixels for the red star located on (b).

amount of neighbours or a total community distance will be labeled as outliers. This method is known to perform well versus other known methods such as the Angle-Based Outlier Detection (ABOD) algorithm as well as the One-Class Support Vector Machine (OCSVM) algorithm when applied to real-world datasets. It is also relatively fast and only depends on the nearest neighbours component as a bottleneck (which exists many approximation techniques). The implementation was used from the general purpose outlier detection toolbox [12].

The second algorithm used was the Hierarchical Density-Based Spatial Clustering (HDBSCAN) [13]. This algorithm performs a hierarchical density based scheme which finds clusters of varying densities. It is very similar to the original Density-Based Spatial Clustering (DBSCAN) algorithm but with more robust parameter selection agenda (i.e. little or no parameter tuning). It's an ideal state-of-the-art (SOTA) algorithm for exploratory analysis. The implementation used can be found in the specific algorithm package [14].

2.4 ESDL Platform

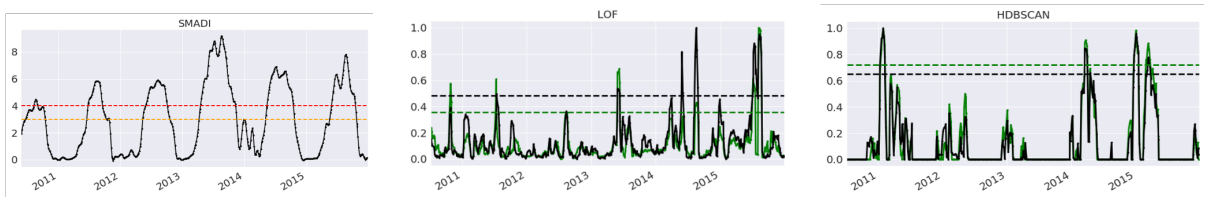
We chose to work with the ESDL platform because it would allow us to have access to over 40+ variables over the span of 10+ years. It is a convenient way to do exploratory analysis in an all python environment (as well as R and Julia) without the need to worrying about package management or cluster accessibility. The variables we intended to use were the land surface temperature (LST), the Normalized Difference Vegetation Index (NDVI), and the root soil moisture (SM). We also observed a section on the webpage which expressed interest in incorporating direct access to the EM-DAT; a database which documents areas where there are societal catastrophes. Unfortunately, the study time period was during 2010 until 2016 which is outside the range of a few variables. The reason is because that is when the SMADI indices were calibrated so we wanted to make a fair comparison between the two. Regardless, we made use of the platform for many preliminary studies and to prototype many approaches to dealing with the EO data.

3 Results

3.1 California Drought

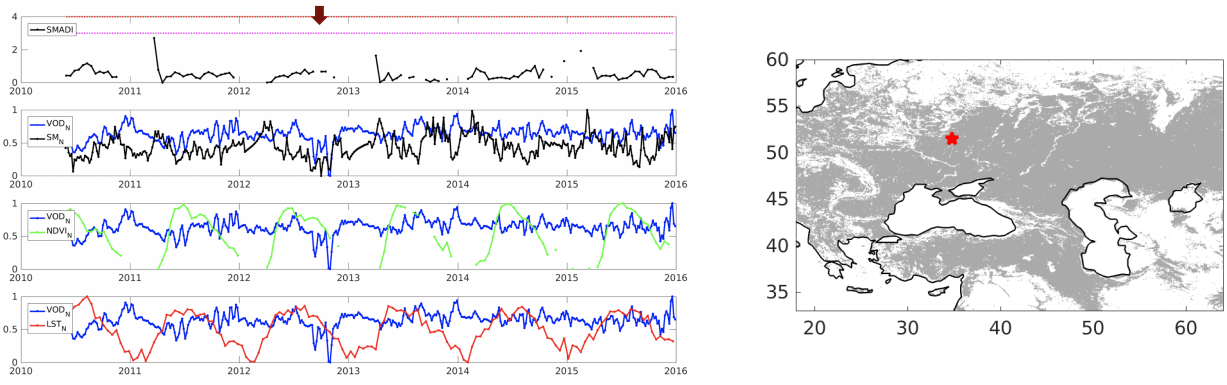
Here we see that the SMADI integration of SM, NDVI, and LST is key to detect droughts with drought severity categories which could possibly be better aligned with EM-DAT. Furthermore, the VOD time evolution reflects drought-induced crop water stress. We can see that the drought conditions are found each year but the severity is higher for the latter years.

We see that there is some correspondence between the SMADI indices and the unsupervised learning methods. We also see that the severity changes for 2014 with the addition of VOD for the LOF algorithm but not much for the HDBSCAN algorithm.



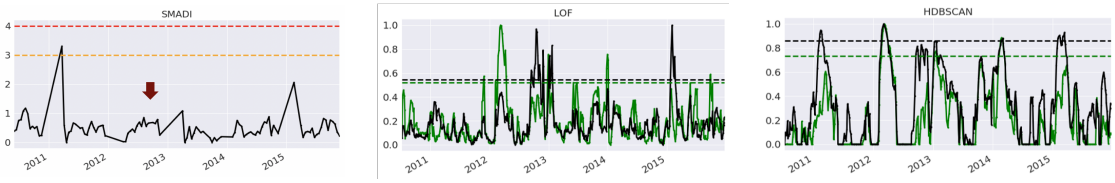
(a) (b) (c)

Figure 3: California yearly drought for 2011 - 2016. The (a) SMADI results compared to (b) simple LOF algorithm and (c) the HDBSCAN algorithm. The solid green line are the original variables (LST, SM and NDVI) and the solid black line includes the VOD variable. The dotted green line represents the top 5% least likely events (a.k.a. anomalies) for the combined variables without VOD and the dotted black line represents the top 5% least likely events with VOD.



(a) (b)

Figure 4: Russian drought in 2012 (June - September). (a) shows a group of averaged pixels for the red star located on (b).



(a) (b) (c)

Figure 5: Russian drought in 2012 (June - September). The (a) SMADI results compared to (b) simple LOF algorithm and (c) the HDBSCAN algorithm. The solid green line are the original variables (LST, SM and NDVI) and the solid black line includes the VOD variable. The dotted green line represents the top 5% least likely events (a.k.a. anomalies) for the combined variables without VOD and the dotted black line represents the top 5% least likely events with VOD.

3.2 Russian Drought

Here we see that the extreme drought of 2012 (EM-DAT) was undetected by SMADI. There was some correspondence between VOD and the other variables which give rise to the prediction of the droughts. This is evident in figure 5 where we see that there are some drought events that were detected by HDBSCAN and LOF. The LOF actually exhibited more changes with the addition of VOD than the HDBSCAN method overall.

4 Discussion

4.1 Preliminary Results

We found that the inclusion of VOD as a feature has observable indicators of drought events when compared to the traditional land surface temperature, NDVI and soil moisture features. There is substantial evidence that VOD can help the detection of drought occurrences and possibly other extreme events in nature; thereby justifying the use of it in modeling efforts.

We have also seen that there is indeed some correspondence between the unsupervised ML anomaly detection algorithms and the SMADI indices. We have shown that the ML algorithms (although unsupervised) were successfully able to detect drought events in the 2 scenarios of constant drought occurrences such as Sahel and more distinct drought occurrences such as CONUS. It should be noted that many ML algorithms have a trade off between simplicity and complexity so one needs to be careful when employing such algorithms blindly to applications. We have demonstrated that for two different approaches (LOF and HDBSCAN) with two different modeling philosophies, they were still able to detect extreme events. However, the LOF method saw more improvement (or at least correspondence to the SMADI indices and specific event of the Russian drought of 2012) than the HDBSCAN algorithm.

Our preliminary results were presented in the ESA Living Planet Seminary 2019 where there was great interest in the SMADI indices as well as the inclusion of VOD for extreme events. Due to the interest in the SMADI indices, the data was open-sourced which can be found here [2]. We welcome the collaboration opportunity to include SMADI as well as the VOD data into the ESDL if there is interest as we think it would be very valuable to the community and worth exploring further; especially for drought detection.

4.2 Future Work

The immediate next steps would be to add more features to the inputs to the learning algorithms. For this particular application and proof of concept, we only used single spatial location values with a time window of 3 cycles (24 days in the past). We know that time plays an important role in the detection of droughts but we would like to see if spatial pixel values impact the results so we would like to incorporate spatial attributes to the training data. We would like to extend the time window range as well as the spatial range and see what the trade-off is between the methods. It also does not hurt to add more variables that could potentially aid the detection of drouhgts.

We would also like to work with the EM-DAT database and use these as labels for training some ML algorithms. This would convert the unsupervised learning problem into a (semi-)supervised learning or physics-guided problem but it would allow us to constraint the results so that we can hopefully detect drought events more accurately. In doing so, we can do some more advanced and thorough statistical comparisons between the results of our learning algorithms and the EM-Dat (as well as SMADI).

We would like to add an additional algorithmic family based on advanced density estimation. This method works by transforming any high-dimensional, multi-variate distribution by way of the change of variables formula. In doing so, this allows us to calculate a complete distribution and therefore characterize the data. We can look at the information content of this density and try to find some correlation between the variables chosen and the droughts found. Furthermore, this is the perfect opportunity to experiment with this class of generative models (i.e. Normalizing Flows) which are very prevalent in the community but have rarely been explored with physical data for extreme events detection.

We plan to have more recent results at the upcoming Phi-Week event in September, 2019 and have submitted a draft to the Climate Informatics workshop, October 2019. We hope that these results will eventually culminate into a journal publication.

5 Assessment of Platform

5.1 Assessment of ESDL

Overall, I believe the ESDL platform is a bit step in the right direction. The fusion of AI and Climate change initiative is a huge effort (**citation**) that is taking place right now so common platforms such as this one offer a very important bridge between different communities (applied and ML). (See pangeo [3] as another example). As a practitioner of ML, it

[2] zenodo.org

[3] `https://pangeo.io/`

is very convenient for me to have all of the variables stored in a single location where I can access them at will. Python is also a very popular programming language with many different libraries to accomplish almost all of the necessary tasks possible.

There were a few difficulties we faced when doing our project but I attribute most of them to the fact that the platform is still young and undergoing development. The main problem was the data present. I was unaware of the data constraints so I chose a project where the data we wanted to compare to (SMADI and VOD) had been collected over the range of 2010 until 2016. This is very recent as was chosen because the satellite is fairly new and this is the only data they had available. There was little to no overlap between some of the variables such as LST and soil moisture. Furthermore, there were no common indices such as NDVI which I found surprising at it is a commonly used index in the remote sensing and Earth science community. Obviously one solution would allow everyone to have the ability to upload their own data but that could exponentially grow if not kept in check. Another possible solution is to develop the API to be able to call at will (under some reasonable constraints) certain data variables from the ESDL as needed to allow some uses to work locally.

5.2 Reproducibility

There were a few challenges we faced when using the ESDL system as highlighted above. Most of them were programmatic and we attribute them to the fact that we are still a fairly small community using the ESDL system. Furthermore, the xarray package is still fairly young and there aren't that many tutorials on how to use it effectively; especially in the machine learning context. One of the most important problems I encountered was the lack of changing the data from the lat-lon-time format to the samples-features and labels format which is needed form ML algorithms. It's also essential that this method be efficient as the ESDL is a large database. I also needed a fast and efficient function to transform the data to add spatial-temporal features. It works like collapsing a sequence of minicubes into samples in features of the large datacube. This is absolutely essential for methods that can take into account higher order spatial-temporal (and even spectral) respresentations of data (e.g. density estimation). Please see the github repository [4] for some of the helper functions I have created.

References

[1] Jakob Zscheischler, Miguel D. Mahecha, Stefan Harmeling, and Markus Reichstein. Detection and attribution of large spatiotemporal extreme events in Earth observation data. *Ecological Informatics*, 15:66–73, 2013.

[2] Iftikhar Ali, Felix Greifeneder, Jelena Stamenkovic, Maxim Neumann, and Claudia Notarnicola. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing*, 7(12):16398–16421, 2015.

[3] Alexandra G. Konings, Maria Piles, Narendra Das, and Dara Entekhabi. L-band vegetation optical depth and effective scattering albedo estimation from SMAP. *Remote Sensing of Environment*, 198:460–470, 2017.

[4] Nilda Sánchez, Ángel González-Zamora, María Piles, and José Martínez-Fernández. A new Soil Moisture Agricultural Drought Index (SMADI) integrating MODIS and SMOS products: A case of study over the Iberian Peninsula. *Remote Sensing*, 8(4), 2016.

[5] J. Martínez-Fernández, A. González-Zamora, N. Sánchez, and A. Gumuzzio. A soil water based index as a suitable agricultural drought indicator. *Journal of Hydrology*, 522:265–273, 2015.

[6] Rui Li, Atsushi Tsunekawa, and Mitsuru Tsubo. Index-based assessment of agricultural drought in a semi-arid region of Inner Mongolia, China. *Journal of Arid Land*, 6(1):3–15, 2014.

[7] Rémi Domingues, Maurizio Filippone, Pietro Michiardi, and Jihane Zouaoui. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74:406–421, 2018.

[8] Nilda Sánchez, Ángel González-Zamora, José Martínez-Fernández, María Piles, and Miriam Pablos. Integrated remote sensing approach to global agricultural drought monitoring. *Agricultural and Forest Meteorology*, 259(March):141–153, 2018.

[9] Milan Flach, Sebastian Sippel, Fabian Gans, Ana Bastos, Alexander Brenning, Markus Reichstein, and Miguel D. Mahecha. Contrasting biosphere responses to hydrometeorological extremes: Revisiting the 2010 western Russian heatwave. *Biogeosciences*, 15(20):6067–6085, 2018.

[10] David Chaparro, Maria Piles, Mercè Vall-llossera, Adriano Camps, Alexandra G. Konings, and Dara Entekhabi. L-band vegetation optical depth seasonal metrics for crop yield assessment. *Remote Sensing of Environment*, 212(April):249–259, 2018.

[4] `https://github.com/IPL-UV/esdc_tools`

[11] Markus M. Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: Identifying Density-Based Local Outliers. In *ACM sigmod record*, pages 93–104. ACM, 2000.

[12] Yue Zhao, Zain Nasrullah, and Zheng Li. PyOD: A Python Toolbox for Scalable Outlier Detection. pages 1–6, 2019.

[13] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017.

[14] Leland McInnes and John Healy. Accelerated Hierarchical Density Based Clustering. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017-Novem:33–42, 2017.