Spatio-Temporal Analysis of Drought Patterns in Amhara, Ethiopia

Iman Anwarzai - ina2109

December 16, 2022

1 Abstract

This study provides a spatio-temporal analysis of drought patterns in the Amhara region of Ethiopia. In order to identify whether a specific village is experiencing drought, rainfall data can be aggregated over that village using satellite rainfall estimates from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) dataset. Since this data is noisy and imprecise, it may be more beneficial to aggregate over a larger area, but not too large, since this would overlook important variations in weather patterns between different places. Unsupervised learning can be used to determine which places have historically had similar precipitation patterns, while finding the optimal scale to cluster over. This study uses the X-means algorithm which builds upon K-means by finding the optimal setting of k over the data (Pelleg and Moore, 2002). In addition, I use the G-means algorithm which converges when the data is Gaussian relative to their centers (Zhao et al., 2009). It was found that X-means can result in many clusters when the dimensionality of the data is low, while G-means' number of clusters rapidly increases with dimensionality. Additionally, there exists a tradeoff between error and loss of variance due to PCA.

2 Introduction

The Amhara region of Ethiopia offers many interesting geographical features, including rivers, mountains, and lakes, which result in varying levels of precipitation throughout the area. This study aims to find the best way to divide areas within this region based on their precipitation history. These separated areas will be used for other analyses related to drought by the International Research Institute for Climate and Society (IRI) at Columbia. The goal of this paper is to find these regions via clustering algorithms. I use the X-means and G-means algorithms since they do not require an initial setting of k, the number of clusters. Instead, they iteratively find the optimal value for k based on different sets of criteria. The data that we cluster over is precipitation data from the CHIRPS dataset.

Climate science has long used unsupervised classification methods to identify spatio-temporal patterns of variability in weather data. The empirical orthogonal functions (EOF) family of methods, derived from principal component analysis (PCA), are the most commonly used for this purpose (Hannachi et al., 2007). In this case, we are specifically interested in identifying droughts at a scale which balances accuracy and precision, so a more tailored set of methods is needed.

K-means is a popular unsupervised learning algorithm, but it has downfalls that can be mitigated by X-means, which efficiently estimates the optimal number of clusters in a dataset (Pelleg and Moore, 2002). Tamene et al. (2022) used K-means and X-means to identify areas in Ethiopia with similar environmental and biophysical features called similar response units (SRUs). They

found that SRUs had higher similarity with lower levels of within-cluster of variability than the former classification of the region into agro-ecological zones (AEZs). Kalliolevo et al. (2022) used X-means to find the number of clusters within Finland's biodiversity data.

Dinku et al. (2018) and Funk et al. (2015) both conduct a ground-truth validation of the CHIRPS dataset against weather station data, establishing that it is a reliable measure of precipitation across Eastern Africa. They find that error in the dataset over northern Ethiopia appears to be stochastic and decreasing over spatial and temporal scale. Similarly, Black et al. (2015) find that the coherence of satellite rainfall estimates improves with aggregation over space and/or time.

3 Methods

The data used in this study was collected by the CHIRPS satellite, which has rainfall data dating back 40 years in the form of daily data and pentads. The actual satellite collects the data in pentads, while the daily data is interpolated, so I use the pentad data to reduce noise. The rainfall data exists for each of 48x49 equally sized areas (or "pixels") over the Amhara region.

Rather than use the data as actual rainfall amounts, I converted into another format that can help deduce the rainfall compared to the "usual" precipitation for that time over that area. This format is called the Standardized Precipitation Index (SPI) and it is the Z-score of the precipitation for a certain date relative to the precipitation amounts for days around the same date for all years. Literature suggests using SPI with around 1 month around the date for meteorological drought (Homdee et al., 2016; Tirivarombo et al., 2018). In this study, I used an aggregation window with a radius of 4 pentads, so 9 pentads including the current date which results in 45 days for an individual year. Thus, for any one precipitation reading, I calculate the Z-score of that precipitation relative to 9 pentads*40 years = 360 other points to determine how typical this precipitation amount is. I also omitted the non-rainy season (October-January) from the dataset prior to doing any analysis to further reduce noise.

Principal component analysis (PCA) is a technique used to reduce the dimensionality of data to a lesser number of dimensions, which I will call m. It does this by transforming the data into a new coordinate system, where the new axes are called principal components. The principal components are chosen to preserve the maximum amount of variation in the data, and they are orthogonal to eachother (independent and not correlated). I used PCA for data compression, reducing the number of features from 2006 in the original data set (the number of features after omitting the non-rainy season). Clustering algorithms suffer from a phenomenon known as the curse of dimensionality when working with high-dimensional data. With more dimensions, the data is often sparse and the distances between points become less meaningful, which can make it difficult to analyze and model the data. This is because as the number of dimensions increases, the volume of the space increases exponentially, while the amount of data available to fill that space remains constant. This can lead to overfitting and other problems. Since we are mainly interested in unusual historical patterns, such as drought, it makes sense to use PCA here for dimensionality reduction and to preserve the features with the highest variance.

K-means is a widely used clustering algorithm which is used to divide data into a specified number of clusters. The K-means++ algorithm is used to randomly generate initial centers that are far apart, in order to prevent suboptimal solutions. Using these initial centers, the K-means algorithm assigns data points to its closest center, and then the centers for each cluster is recalculated to be the mean of its assigned points. This process continues until the cluster assignments no longer change, at which point the algorithm has converged and the final cluster assignments are returned. The biggest drawback of using K-means is that we must set the value of k prior to running the algorithm. The clustering can change drastically based on the value set for k. It is the goal of this study to find k such that it is not too small or too large, and so the X-means and G-means algorithm can be utilized to find this optimal setting.

The X-means algorithm is another unsupervised learning method for clustering data into groups. It is an iterative algorithm that starts with a pre-defined number of clusters and tries to improve the divisions based on the Bayesian Information Criterion (BIC). Equation (1) describes how the BIC is calculated, where k is the number of cluster centers, d is the number of dimensions, and n is the number of points in the dataset. I used the K-means++ algorithm to initialize the first two centers for the dataset. X-means was also run with a tolerance of 0.0001 for convergence, meaning that the algorithm will stop when the maximum value of the change of cluster centers is less than this tolerance amount.

$$BIC(C|X) = L(X|C) - \frac{k(d+1)}{2}\log(n) \tag{1}$$

The G-means algorithm iteratively grows the number of clusters by splitting clusters that have data points that are not Gaussian relative to their centers, which is determined by the Anderson-Darling statistic. Between iterations, K-means with K-means++ initialization is run to refine the solution. I ran G-means clustering starting with a single cluster and with tolerance once again set to 0.0001. I also set the repeat value of the G-means algorithm to 5. This means that the K-means between rounds is run five times before selecting the optimal clusters for that iteration. Repeat is used to prevent the algorithm from getting stuck at a local optimum when K-means is run.

I ran PCA on the dataset with the number of components set to 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, and 1000. For each set of transformed data as well as the original dataset, I ran the X-means and G-means algorithms to find the optimal number of clusters determined by these algorithms. When the algorithms converged, I calculated the sum of squared errors (SSE) of the clustering using Equation (2).

$$error = \sum_{i=0}^{N} ||x_i - center(x_i)||^2 \tag{2}$$

4 Results

I initially ran PCA with $m = 4$ components on the data, and the results are shown in Figure 1. It can be seen that in the first component, pixels on the west side of the region vary together while pixels in the east also vary together, but in the opposite direction. This component has the highest explained variance of 31%. For the second component, the region is similarly divided into north and south. For the third and fourth components, pixels in opposite corners vary together while the middle section of the map varies together in the opposite direction. In total, PCA with four components only accounts for around 58% of the explained variance. Past literature suggests PCA should result in at least 60% of the explained variance for it to be meaningful, so I conduct PCA with many values greater than four later in this study.

Next, I ran K-means with $k = 4$ on the original dataset (without PCA) and the results can be shown in Figure 2. It was interesting to see how K-means resulted in pixels that were geographically close ending up in the same cluster, even though the geographical location of features wasn't taken into account as a feature. The data was solely clustered based on the precipitation history. At this point, since PCA hasn't been conducted on the dataset, the data had a lot of noise due to the fact that there were over 2000 features. The error was calculated using Equation (2).

For the main study of the paper, I ran PCA on the data followed by the X-means and G-means algorithms in order to see how the algorithms converged when preserving different amounts of variance. Both algorithms were run 16 times for the following settings of m in PCA: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 1000, and 2006 (original dataset). The graph of m vs. variance explained can be seen in Figure 3. For different values of m, the X-means algorithm converges in the range $k = 140$ to $k = 464$. The G-means algorithm converges in

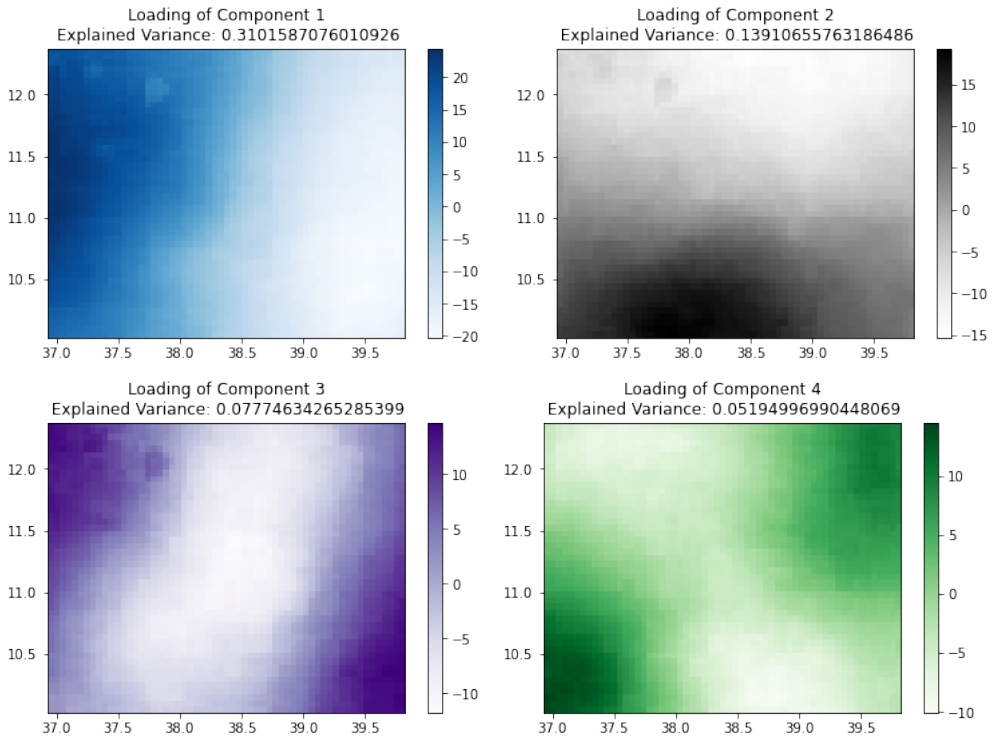

Figure 1: PCA with 4 components

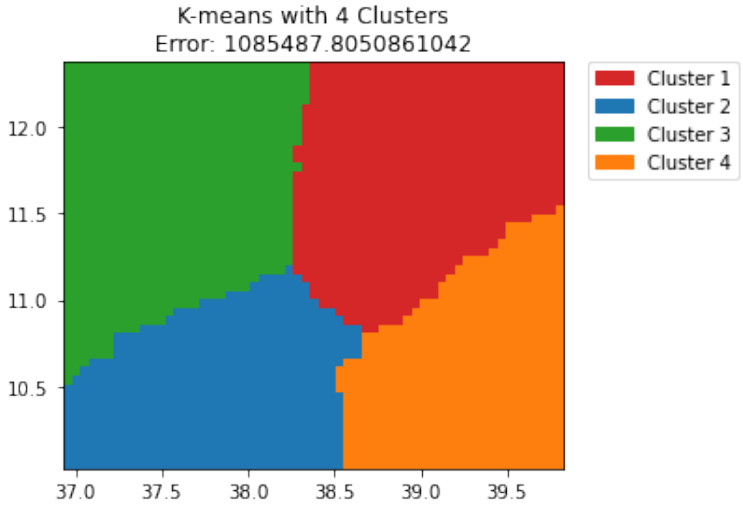

Figure 2: K-Means with 4 clusters

the range $k = 96$ to $k = 714$. The results of m vs. k for each algorithm can be seen in Figure 4.

Figure 5 displays the clustering assignments by the X-means and G-means algorithms for the following number of components from PCA: 10, 100, 2006. It can be seen again that geographically close pixels tend to be in the same cluster, even though the clustering is entirely based upon the precipitation history and does not take into account location.

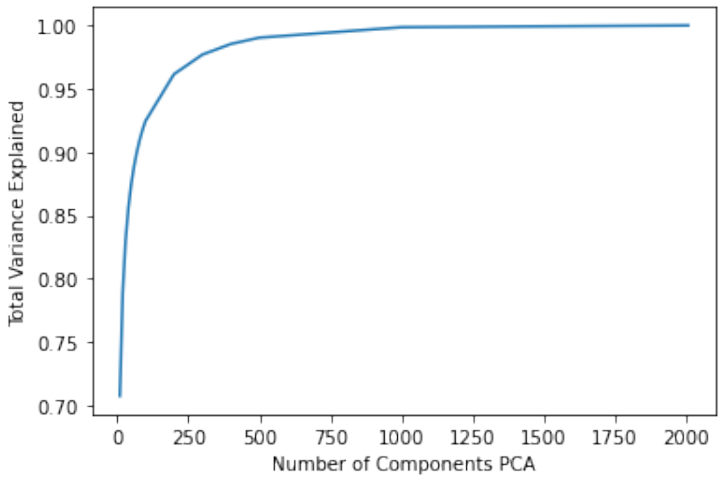

Figure 3: Number of components in PCA vs. total variance explained by the data

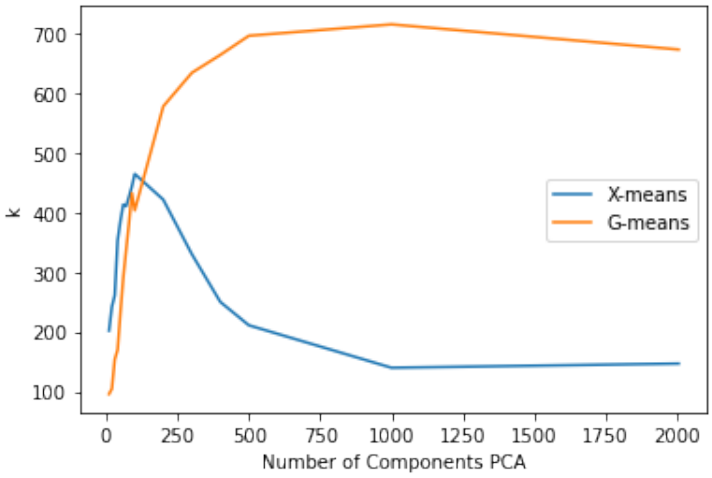

Figure 4: Number of components in PCA vs. the number of clusters for the X-means and G-means algorithms

5 Discussion

The trend shown by the X-means algorithm is that as the number of components m increases, the number of clusters spikes at first then gradually decreases. This could be due to the X-means algorithm trying to maximize the BIC (Equation (1)) which is negatively correlated with a high k as well as a high number of dimensions. Thus, for high dimensions, X-means converges faster. The G-means algorithm is shown to increase quickly with higher dimensions then gradually decrease after $k = 1000$. The G-means algorithm converges when the data is Gaussian relative to their centers, and this algorithm is shown to result in a very high k value. Since there are only 2832 pixels in the dataset, when k is large (>1000) then there are only a handful of pixels in each cluster, so it is less likely to classify a cluster's distribution as Gaussian, especially when there are many dimensions. At this point, the algorithm stops splitting because it cannot reject the null hypothesis of the points being not normally distributed. For both algorithms, k noticeably flattens after 1000 components where over 99.86% of the variance is explained.

Figure 6 shows how the error of each clustering increases with PCA. This result is rather intuitive,

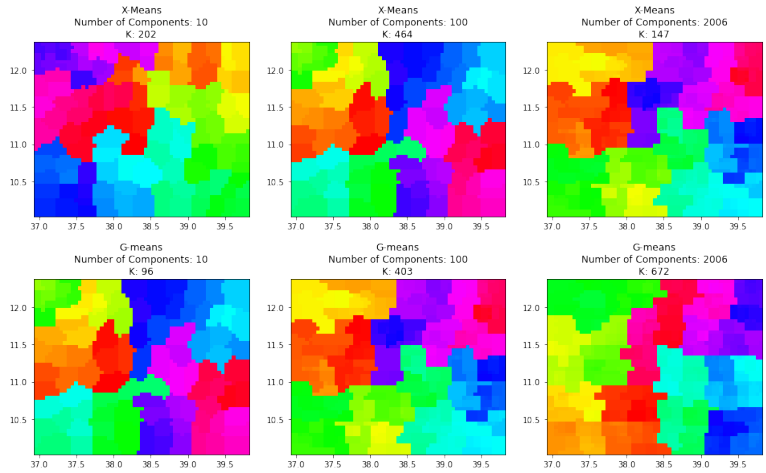

Figure 5: X-means and G-means clustering assignments for various values of PCA number of components

due to the curse of dimensionality making errors larger in higher dimensions. As a result, I decided to also graph the error vs. the number of clusters k for all runs of the X-means and G-means algorithms above. The results are shown in Figure 7 and it reflects the trends shown in Figures 4 and 6. In the X-means algorithm, with a small number of components, there is relatively small error and a small k. When the number of components is larger than approximately 200 (when k is approx. 400), then the optimal k decreases and the error increases due to the addition of more complexity that cannot be explained. To ultimately select the optimal value for k, there would have to be a tradeoff between error and loss of explained variance. If there is a cap on the number of clusters, then it would generally be better to cluster with fewer dimensions using the G-means algorithm or with more dimensions using the X-means algorithm.

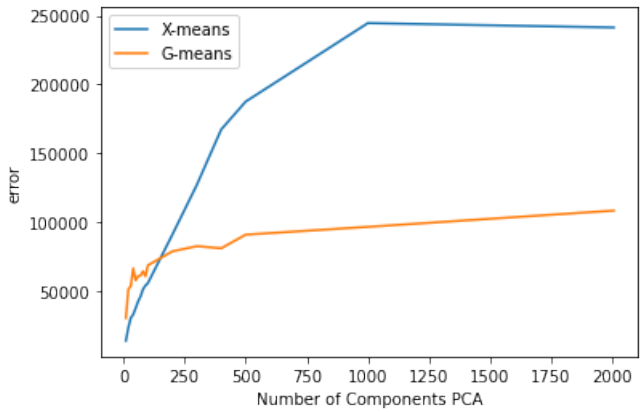

Figure 6: The error of the clustering vs. the number of components after PCA

6 Conclusion

To further refine our findings, it would be interesting to change some parameters in our algorithms, such as the radius of the SPI data. I used a radius of 4 pentads since it was suggested by other literature, but it may be beneficial to aggregate over a smaller window. In addition, there is a lake present in the upper left corner of the map, and so it may be helpful to run the clustering algorithm over a smaller area that avoids this. Additionally, the G-means and X-means algorithms

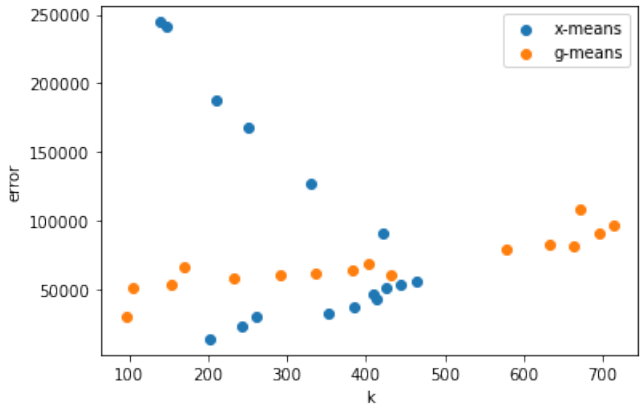

Figure 7: The error of the clustering vs. the number of clusters for all runs of the X-means and G-means algorithms in this paper

result in different solutions depending on how K-means++ selects the initial centers. The figures generated showed the same trends every run, but it may of note to study how this affects k.

One method for further validation of the optimal clusters is to calculate the percentage of the total variance explained rather than the total error. This could be something like the sum of squares between clusters divided by the total error. This way, the error is scaled by the variance of the data overall which is correlated with the number of dimensions in the data.

In order to validate the data, we have farmer polling data collected by the International Research Institute for Climate and Society (IRI) at Columbia in which farmers in various villages in Amhara, Ethiopia ranked their worst years in terms of drought. The borders of these villages are outlined in red in Figure 8. We could match the villages to the clusters found by our algorithms and compare the worst years as ranked by the farmers with the worst years according to the rainfall data within each cluster. This can be used to verify how similar precipitation patterns are in terms of extreme drought within each cluster. Overall, I aim for these findings to be a helpful starting point for clustering the Amhara region based on historical rainfall patterns to allow for further analyses related to climate science to be conducted for this region.

References

Black, E., Tarnavsky, E., Greatrex, H., Maidment, R., Mookerjee, A., Quaife, T., and Price, J. (2015). Exploiting satellite-based rainfall for weather index insurance: The challenges of spatial and temporal aggregation.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., and Ceccato, P. (2018). Validation of the chirps satellite rainfall estimates over eastern africa. *Quarterly Journal of the Royal Meteorological Society*, 144(S1):292–312.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1):150066.

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, 27(9):1119–1152.

Homdee, T., Pongput, K., and Kanae, S. (2016). A comparative performance analysis of three

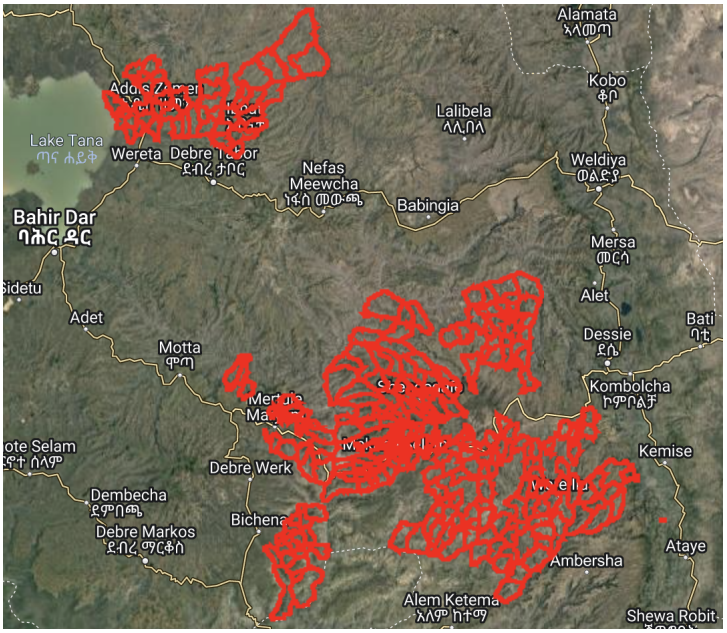

Figure 8: The Amhara region with individual villages outlined in red

standardized climatic drought indices in the chi river basin, thailand. *Agriculture and Natural Resources*, 50(3):211–219.

Kalliolevo, H., Salo, M., Hiedanpää, J., Jounela, P., Saario, T., and Vuorisalo, T. (2022). Considerable qualitative variability in local-level biodiversity surveys in finland: A challenge for biodiversity offsetting. *Journal for Nature Conservation*, 68:126194.

Pelleg, D. and Moore, A. (2002). X-means: Extending k-means with efficient estimation of the number of clusters. *Machine Learning, p*.

Tamene, L., Abera, W., Bendito, E., Erkossa, T., Tariku, M., Gelagay, H. S., Degefie, D. T., Sied, J., Feyisa, G. L., Wondie, M., and Tesfaye, K. (2022). Data-driven similar response units for agricultural technology targeting: An example from ethiopia. *Experimental Agriculture*, 58.

Tirivarombo, S., Osupile, D., and Eliasson, P. (2018). Drought monitoring and analysis: Standardised precipitation evapotranspiration index (spei) and standardised precipitation index (spi). *Physics and Chemistry of the Earth, Parts A/B/C*, 106:1–10.

Zhao, Z., Guo, S., Xu, Q., and Ban, T. (2009). G-means: A clustering algorithm for intrusion detection. In Köppen, M., Kasabov, N., and Coghill, G., editors, *Advances in Neuro-Information Processing*, pages 563–570, Berlin, Heidelberg. Springer Berlin Heidelberg.