Machine Learning For Earth System Sciences
Assignment 2

Name: Manav Nitin Kapadnis
Roll No: 19EE30013

Problem Statement:

Consider a gridded global land surface temperature and rainfall all over the world (1deg-1deg resolution). Use any dataset from the internet.

1. At each location, consider annual maxima and annual minima of temperature and annual maxima of daily rainfall as block extremes.
2. For the period 1950-1999, use these block extremes to fit a GEV distribution at each grid location for all three (max temp, min temp, max rainfall).
3. For each of the years 2000 onwards, calculate the return periods of the annual extremes at each location, w.r.t. the fitted GEV distributions.
4. Illustrate on a map of the world (pixelated map is enough) at which locations extreme events with high return periods have become more frequent since 2000.
5. Using any clustering algorithm, cluster the locations on the basis of the 3 GEV parameters as estimated in (3).

Generalized Extreme Value Distribution

The GEV distribution is a family of continuous probability distributions developed within extreme value theory. Extreme value theory provides the statistical framework to make inferences about the probability of very rare or extreme events. The GEV distribution unites the Gumbel, Fréchet, and Weibull distributions into a single-family to allow a continuous range of possible shapes. These three distributions are also known as type I, II, and III extreme value distributions. The GEV distribution is parameterized with a shape parameter, location parameter, and scale parameter. The GEV Is equivalent to the type I, II, and III, respectively when a shape parameter is equal to 0, greater than 0, and lower than 0. Based on the extreme value theorem the GEV distribution is the limit distribution of properly normalized maxima of a sequence of independent

and identically distributed random variables. Thus, the GEV distribution is used as an approximation to model the maxima of long (finite) sequences of random variables.

The cumulative distribution function (CDF) of the GEV distribution is:

$$F(x;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

where three parameters, ξ, μ and σ represent a shape, location, and scale of the distribution function, respectively.
Note that σ and 1 + ξ(x-μ)/σ must be greater than zero. The shape and location parameter can take on any real value.

Dataset Description

A station observation-based global land monthly mean surface air temperature dataset at 0.5 × 0.5 latitude-longitude resolution for the period from 1948 to the present was developed recently at the Climate Prediction Center, National Centers for Environmental Prediction. Some of the unique features of this dataset that differs it from other similar datasets are:

1. It was created by integrating two huge individual data sets of station observations from the Global Historical Climatology Network version 2 and the Climate Anomaly Monitoring System (GHCN + CAMS), allowing it to be updated in near real-time.
2. Some unique interpolation methods, such as the anomaly interpolation approach with spatially-temporally varying temperature lapse rates derived from the observation-based Reanalysis for topographic adjustment.

Furthermore, when the dataset is compared with several existing observation-based land surface air temperature data sets, the preliminary results show that the quality of this new GHCN + CAMS land surface air temperature analysis is reasonably good and the new data set can capture the most common temporal-spatial features in the observed climatology and anomaly fields over both regional and global domains.
The study also demonstrates that there are significant differences in observed surface air temperature and existing Reanalysis data sets, and vary throughout space and seasons. As a result, the temperature data sets from the Reanalysis 2 m may not be suitable for model forcing and validation. The GHCN + CAMS data set will mainly be utilized as one of the land surface

meteorological forcing inputs for deriving other land surface variables such as soil moisture, evaporation, surface runoff, snow accumulation, snowmelt, etc. As a byproduct, this monthly mean surface air temperature data set can also be applied to monitor surface air temperature variations over global land routinely or to verify the performance of model simulation and prediction.

Solution Approach

Initially, we determine the maxima and minima of temperature values from the hourly values of the temperatures that the dataset provides us. A one-year timeframe was formed for this, and the maximum and minimum temperatures were recorded. The block extremes for each of the sites are these maxima and minima.

Note that the locations are real-world locations. Each pixel in the map corresponds to a 1^o x 1^o area. Since the original dataset provides information for the 0.50^o x 0.5^o region, the resolution was upsampled by taking the mean of the 4 0.5^o x 0.5^o to form the 1^o x 1^o region.

The GEV distribution would then be fitted to each of the locations. Each location has a time series of temperature associated with it. Now, in order to fit the GEV, Scipy's built-in function genextreme was utilized. It took roughly 2 hours to fit GEV distribution for each of the 64800 sites on the map (because the number of latitudes is 180 and the number of longitudes is 360).

Furthermore, return levels were estimated for return periods of 5, 10, 20, 50, and 100 starting in the year 2000. It was discovered that when the return duration lengthened, the temperature returned at higher levels. This could be due to the fact that the temperature distribution is monotonically growing, with higher temperatures having higher return times.

The next step is to illustrate which locations saw an increase in extreme events with slow return frequency after the year 2000. As a result, a manual selection of frequency 10 was picked to test this. This implies that if an extreme temperature happens more than 10 times in a specific region, that region will be displayed separately on the world map. Furthermore, the precise temperature value may not occur multiple times. To make the findings more realistic, a leave of 1K was allowed, which means that if the temperature is within 1K of the extreme value of the temperature, that region will also be segregated. The pixelated map is shown in Figure 1.

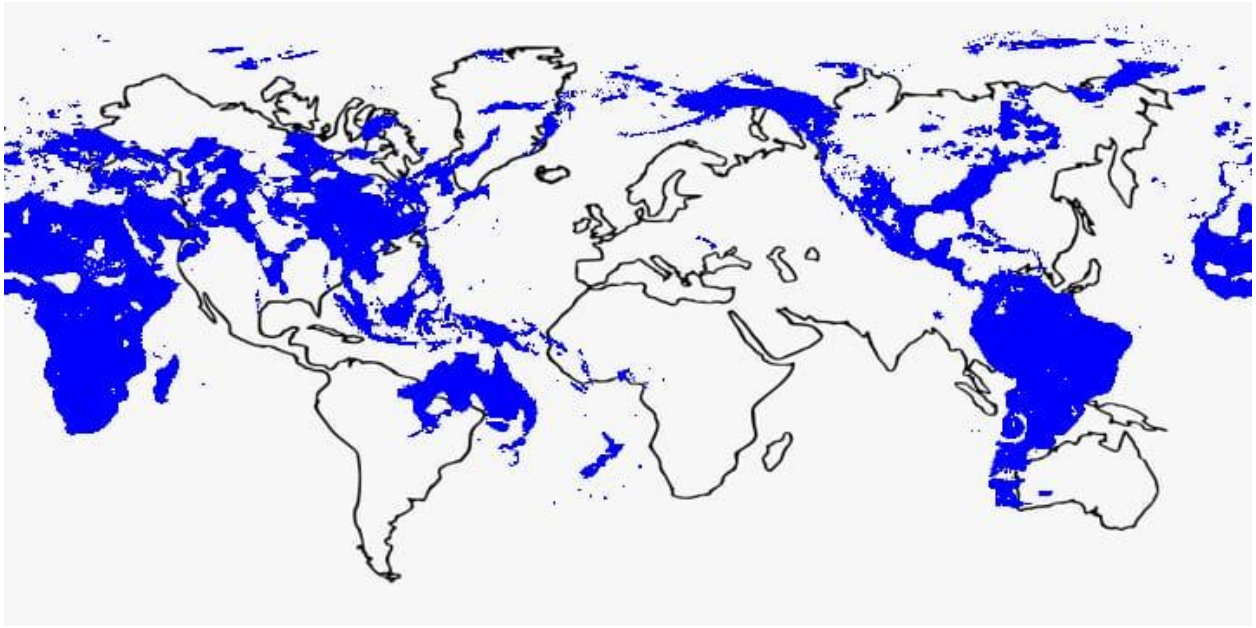

Figure 1: Map with locations having an extreme event with a frequent high return period.

The final step is to cluster the locations based on the parameters (shape, location and scale) of the GEV distribution. The K Means clustering technique was implemented with the feature vectors comprising of shape, location, and scale associated with each location. It is based on the assumption that locations belonging to the same cluster will have comparable values of shape, location, and scale. A total of 20 clusters were created. Because the silhouette score was not computed and verified for different numbers of clusters on different sites, this may not be the optimal number of clusters.
