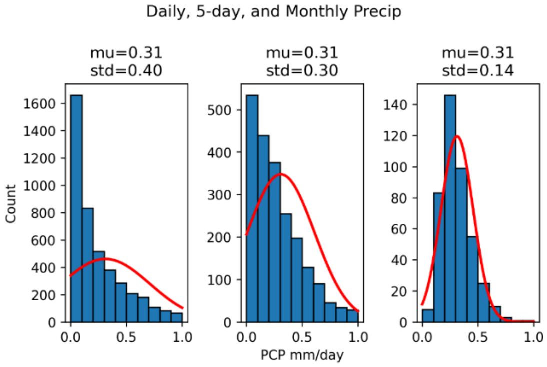
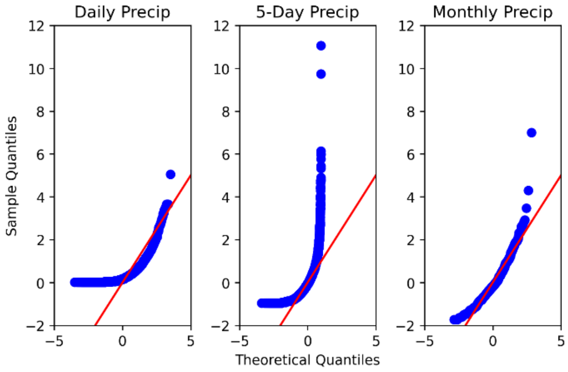
Homework 2
Eliott Foust
Code uploaded to https://github.com/wefoust/Meteo515_AtmosStats

Question 1:

1a) The three plot show histograms of daily, 5-day, and monthly precipitation in State College from 1950-2015. Trivially, the count reduces as averaging applies since the number of samples decreases. More importantly, the distribution changes as more averaging is applied. The distribution of daily precip is highly skewed to the right. The skew infers daily precip totals are more likely to be lower than the mean precipitation. When looking at the 5-day precip totals, the mean remains the same as the daily precipitation, but the distribution becomes less skewed. When viewing the monthly precipitation distribution, it resembles a normal distribution that is centered on .31mm/day. Additionally, the distribution begins to look symmetric, indicates there exists some scenarios where monthly precipitation is equally like to become wetter or drier than the mean. The transformation in shape is explained by the central limit theorem. Despite the volume of samples decreasing, larger samples are (taking the mean of more days) drive the distribution to appear normal.



1B) The figure shows a Q-Q plot of precip. The Q-Q plot reiterates how larger samples drive a distribution to appear normal. When the values are over the theoretical line, they are overestimating the frequency that would constitute a normal distribution. Conversely, they underestimate a normal distribution when values are lower. Hence, the last figure resembles as normal distribution because the red line (theoretical distribution) is a better fit for the data.



1C) This code tests the Goodness of fit for the Daily, 5-day, and monthly precip datasets to a gaussian fit. This code creates bins based on 0-20, 20-40, 40-60, 60-80, and 80-100 percentile of each dataset. It then calculates the Z score at each bin edge. The Z score is used to find the area under a standard normal curve. The area*count of each bin is the expectancy. The counts are then converted to frequencies. The frequencies and counts are then used as parameters for the chisquare function where it returns a multi-dimensional list containing Chi2 values and P values. If P value is less than alpha (.05 in this case) then we reject the null and conclude that the distribution is not normal.

The results of the chi squared test reveal test statistics 11689, 5272, and 755 for daily, 5-day, and monthly respectively. The P values for the test are approximately for all cases. This indicates the null should be rejected in all cases.

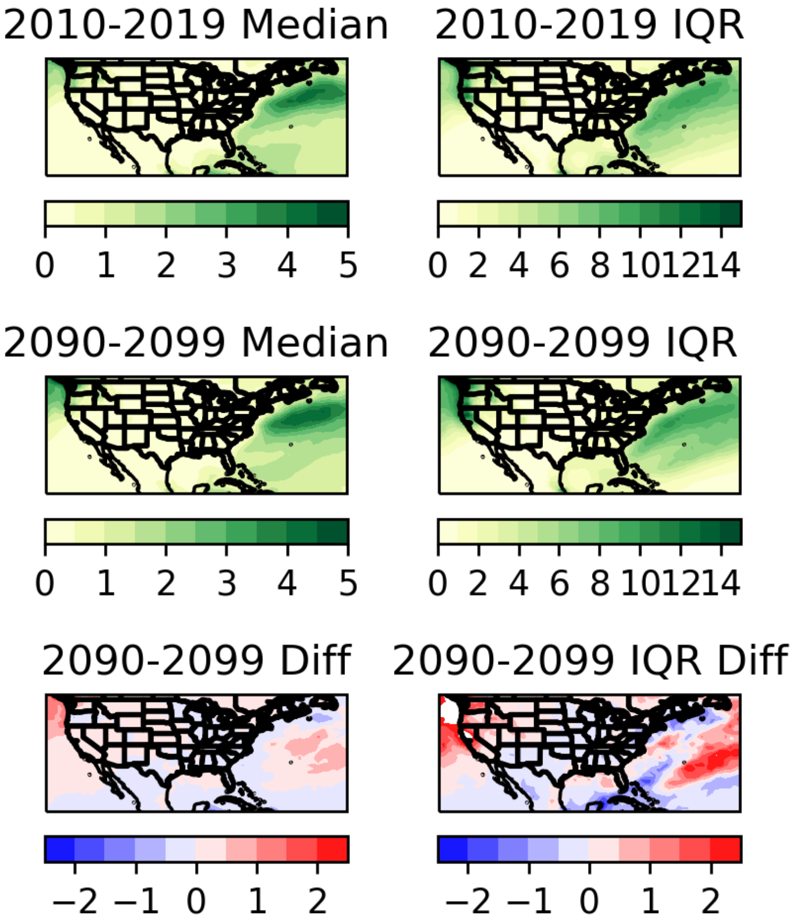
Question 2

A)
Significance Test – A permutation test can be conducted by grouping the September intensity indexs two respective categories. The categories should be combined and permuted. Then the means should be subtracted from each other. The "permute and subtract the means" should be repeated such that a normal distribution forms. Then, a z-test can be performed to test whether the difference in means (mean of cat 5's minus the means of other hurricanes) are significant are significant. A **Z** or t-test can't be performed initially because it is unlikely that the data is gaussian. Also, the small sample size discourages the use of a Z-test.
Test Statistic – Create permutation distribution and then use Z statistic
Null Hypothesis- The mean of the Cat 5' indexes is the same as the mean of the September indexes
Null Distribution – normal distribution

B)
Significance Test – A t-test would be appropriate in this situation. Here, the climatology/population mean is known and normally distributed if trends are removed. Since the sample is so small, a t-test would be more appropriate than a **Z**-test.
Test Statistic – t-statistic
Null Hypothesis- The average temperature on the vacation is equal to the climatology
Null Distribution – normal distribution

Question 3

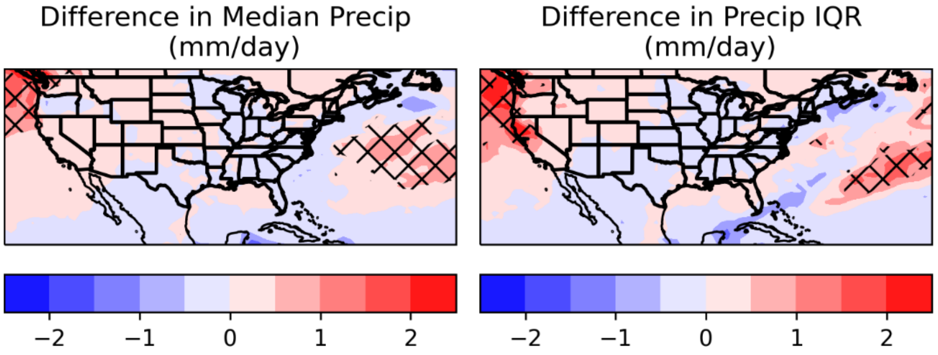
3B)



3C) A permutation test is viable in this situation because it is nonparametric and it does not require any information or assumptions on the sampling distribution. It may not be a good assumption to assume the data is gaussian at all locations so we cannot apply a z or t test directly. In this case, the null hypotheses imply the difference in median and IQR of precip is the same at the beginning and end of the century. We then create a permuted sample distribution and apply z or t test to see if we can reject the null hypothesis.

An added benefit of using permutation tests in this scenario is that the test allows for a wide range of statistics to be tested with the same methodology.

3D) This section calculates whether the differences in median and IQR are significant. This is done by concatenating the 2010's and 2090's datasets. The code then permutes the concatenated data into two equal datasets. The IQR and median of each dataset are calculated and the differences in IQRs and medians are appended to the list diffMedian and diffIQR. This process is repeated 1000 times. Afterwards the 97.5 percentile is calculated from the appended arrays. If the 2010 and 2090 difference in median and IQR are greater than the 97.5 percentiles, then the difference is significant.



Difference in Median Precip (mm/day)

Difference in Precip IQR (mm/day)