Homework 1 - Problem 3

This dataset contains data retrieved from the National Water Model (NWM) Forecast Viewer, a Tethys App created at Brigham Young University for retrieving NWM forecast results. Onion creek (COMID = 5781421) was selected for this particular statistical analysis, in order to better understand the information in this particular future snapshot. That said, forecasts are always changing, and thus this particular analysis will be irrelevant after the final forecasted day has passed.

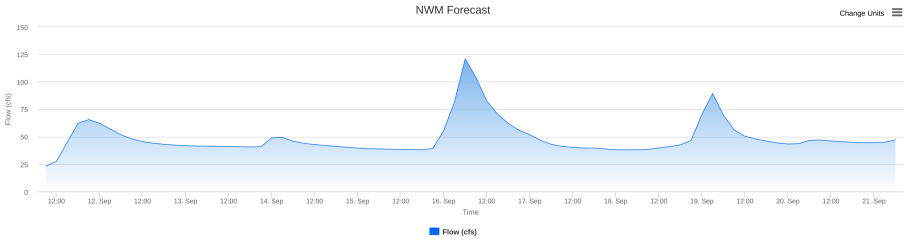

Figure 1: National Water Model Forecast Viewer - Screenshot of Results

The reason for my particular interest in this dataset is its relevance to the final stages of my newly-defined thesis topic. I am still in the very earliest stages of this project, but the gist of my thesis will be to attempt to find statistically significant correlations between each of the rating curves for the 6,600 United States Geological Survey (USGS) stream gages in the lower 48 United States and other nationally-available hydrological and geological datasets. For example, the first step will likely be to retrieve National Land Cover Dataset (NLCD) data and attempt to correlate it to these 6,600 rating curves. I will then continue this analysis for multiple other datasets, with an ultimate goal of determining a relationship between nationally available datasets and stream rating curves. Once I have validated the relationship, I will finally use it to programmatically determine the rating curves for all other 2.7 million stream reaches in the lower 48 United States, ultimately resulting in superior flood extent mapping (and, subsequently, superior flood predictions and warnings)! This is where the NWM data comes into play: once all of the analysis has been completed, the NWM streamflow forecasts will be used to inform flood extents using the rating curves that I will determine.

It would have been nice to do some statistical analysis on a rating curve, but the particular kind of analysis that we are currently working on seemed to lend itself more to the streamflow forecasts.

This particular dataset contains 80 streamflows, one for every 3 hours (8 per day) over a 10-day time period. From a quick glance at Figure 1, it is clear that there will be many large outliers, particularly on September 16th and 19th. This can be verified in the boxplot in Figure 2e. Because of these extreme outliers, we also expect to see a fairly large variance, which can be noted below in the

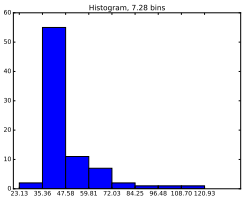

Figure 2: Sturges

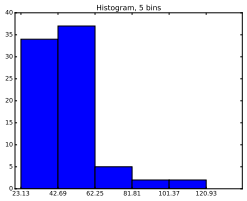

Figure 3: 5 bins

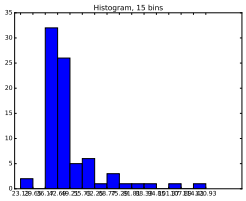

Figure 4: 15 bins

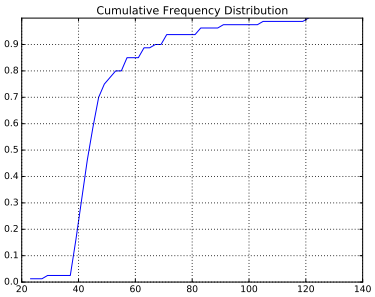

Figure 5: CDF

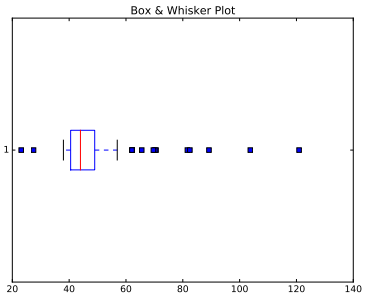

Figure 6: Boxplot

reported statistics. The interquartile range and skewness values are also fairly high, as would be expected, though the quartile skewness is lower than I would have initially guessed.

Here are the descriptive statistics:

mean, 48.46451125
median, 43.9737
var, 234.320703254
q25, 40.5241
q50, 43.9737
q75, 48.994625
iqr, 8.470525
skew, 2.49816738098
qskew, 0.18550503068

It is not at all surprising that the streamflow values would have such a high variance, given how common flooding is at Onion Creek. I also glanced at a few other nearby NWM forecasted streamflows, and many had similar peaks on September 16th, showing that current forecasts are predicting heavy rainfall and insufficient percolation to make up for it (which is also relatively characteristic of the East Texas landscape near Onion Creek). Notably, though, the medium-

range forecasts (which are used in this analysis) are the least accurate of the three forecasts available: short-range forecasts are accurate only a few hours in advance, and consist of only 9 hours of data; whereas long-range forecasts, though more precise and continuing out 30 days, would have been inconvenient for this study seeing as there are 16 of them rather than just one.

Overall, this analysis has not been particularly surprising, but it has not been particularly informing either. This dataset is not the best for this sort of analysis. That said, because the analysis has been set up programmatically, it can very simply be repeated in the future on new datasets as they become available. The ideal statistical analysis for this study would be to analyze two datasets and determine their statistically significant correlation (such as how NLCD and USGS rating curves correlate).

Here is the python code used to run the statistical analysis and create the plots. Note that statswrap.py is imported from my home directory; this module is included with the remainder of my homework 1 submission.

Listing 1: Using statswrap.py module created for HW1 to quickly draw statistical plots and compute statistical parameters.

```
import statswrap

stats = statswrap.Statistics('nwm_med_onionck_091116.csv', 'flow_cfs', 'stats/')
# stats = statswrap.Statistics('rcurve_onionck.csv', 'dep', 'stats/')

# Problem 1a
sturges = stats.get_sturges()
stats.plot_hist('sturges')
stats.plot_hist(5)
stats.plot_hist(15)

# Problem 1b
stats.cumul_freq_dist()

# Problem 1c
stats.write_stats()

# Problem 1d
stats.boxplot()

# Problem 2
stats.norm_dist(0,1,-3,3,100)
stats.lognorm_dist(0,1,0,10,100)
stats.gamma_dist([0.9,2],0.5,0,30,30)
```