Geostatistics
Fall 2018
Homework 1: Probability and Statistics

Jonathan Frame
University of Alabama

October 8, 2018

Introduction: This document is intended to satisfy the second deliverable for homework 1. Below are my answers to the homework problems including text and sample figures – no more than 1 page per question, in PDF format. Additional figures are provided in Appendix A, or by running the corresponding Matlab script.

This assignment included a statistical analysis of 643 watersheds using 50 physical characteristics. Of those 50 characteristics 13 are hydrologic signatures (e.g, statistical metrics of hydrologic responses within each watershed), and 37 are watershed attributes (e.g., characteristics of the watershed independent of the hydrologic regime, but with influence on the hydrologic response).

Contents

List of Figures

List of Tables

1 Data Exploration and Visualization:

The three largest absolute-value correlations between individual catchment attributes and hydrologic signatures are shown in Table 1. These correlations seem reasonable. The highest correlation (precipitation mean vs. discharge mean) makes complete sense in terms of mass conservation. The more water that falls on a watershed, the more water will pass through the watershed, and much of that will be in the form of surface water discharge. It would have been surprising if this were not the highest. The second correlation (precipitation mean vs. the 95^{th} percentile discharge) also makes sense, because the watersheds with the highest precipitation mean should also have the higher soil moisture levels, which will cause more precipitation to runoff as surface discharge, rather than infiltrating into the soil. The low precipitation frequency correlation with the runoff ratio is also reasonable. With lower precipitation frequency the soils will be dryer, which will soak up precipitation and reducing the surface runoff in the watershed. The analysis will focus on these five variables from this point forward, specifically runoff ratio.

Table 1: Three largest absolute value correlations

Rank	Watershed Attribute	Hydrologic Signature	Absolute value correlation coefficient
1	Precipitation mean	Discharge mean	0.89
2	Precipitation mean	95^{th} percentile discharge	0.85
3	Low precipitation frequency	Runoff ratio	0.74

Table 2 includes the mean, standard deviation, skewness and kurtosis of all variables involved in the three correlations discussed above. These statistics can be used to represent the probability distributions of each variable. Figure 1 displays this distribution graphically for precipitation mean in the form of a histogram. The empirical data is shown in blue, and a hypothetical distribution using the statistics in Table 2 is displayed as in the background. Hypothetical distributions are often useful to compare with empirical data to get an qualitative idea of how well statistics represent the data.
Figure 2 shows the joint probability in the form of a contour map. Each contour represents an equal probability of the two variables occurring simultaneously. The maximum likelihood appears as yellow contour lines at about runoff ratio = 0.7 and low precipitation frequency is near 210. Similar plots are shown for each of the other high correlation pairs in Appendix A.

Table 2: Statistics for variables involved in high correlations

Variable	Mean	Standard deviation	Skewness	Kurtosis
Precipitation mean	3.2	1.4	1.1	5.3
Discharge mean	1.5	1.5	2.5	10.1
Low precipitation frequency	255.4	34.7	-0.2	2.7
Runoff Ratio	0.4	0.2	0.9	4.1
95^{th} percentile discharge	4.9	4.8	2.3	9.4

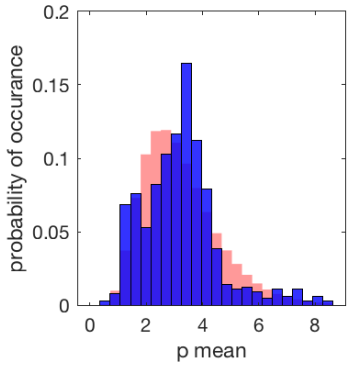

Figure 1: Histograms of high correlation variable

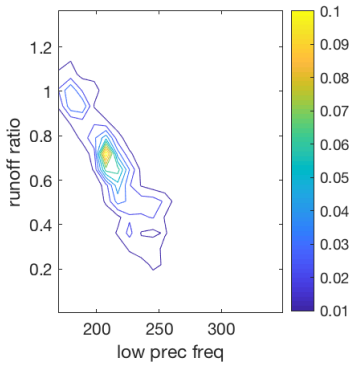

Figure 2: Contour plot of joint distribution

2 Sample Statistics:

I used a random sample of 50 catchments to calculate the sample mean and 90% confidence intervals of the runoff ratio. Table 3 shows the statistics of the random sample. Table 4 shows the 90% confidence interval of the mean. These statistics and confidence intervals will change depending on the samples taken from the full set of 643 catchments.

Table 3: Statistics of random sample

mean	St.Dev.	skewness	kurtosis
0.34	0.24	0.97	3.6

Table 4: 90% confidence interval of the mean

Rand. samples	Low bound	Upper bound
50	0.28	0.3931

The distribution of the sample mean is plotted in Figure 3 against the actual mean from all 643 catchments. This plot shows that the sample mean is slightly less than the actual mean, and has a slightly higher variance, hence the lower peak. Figure 4 shows the distributions of stratified sampling using the geologic and vegetation characteristics of the watersheds.

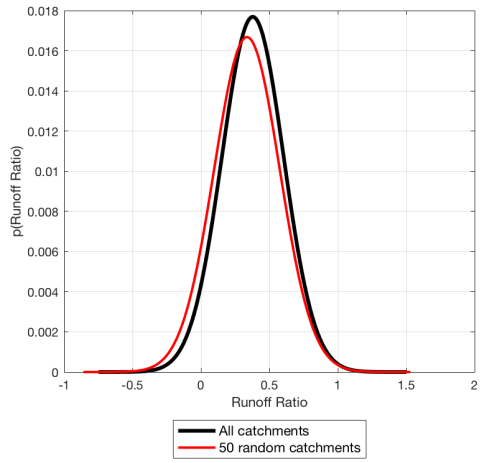

Figure 3: Distributions of sample mean

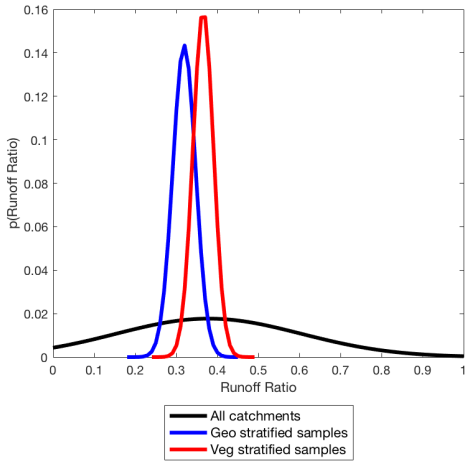

Figure 4: Distributions of sample mean with stratification

Stratified sampling requires some judgement decisions. I chose not to include any strata which had an $n_{optimal}$ value of less than 4. I did not replace those missing samples elsewhere, because I did not want to make any ad-hoc decisions on where to move samples. Thus, the total samples are less than 50 throughout the strata. Table 4 presents the results of the stratified sampling calculations. Again note that the statistics calculated from random sampling will not be repeated. The values in Table 4 correspond to the distributions shown in Figure 4, but are the result of just one run of the MatLab script included in this assignment.

Geological class types include in stratification sampling: 1:6.
Vegetation class types include in the stratification sampling: 1:3, 5, 10 & 11.

Table 5: 90% confidence interval of the mean using stratified sampling

Stratification variable	Total samples*	Estimate of the mean	Lower bound	Upper bound	Variance of the mean	Relative precision
Geology	47	0.32	0.31	0.33	8e-04	74
Vegetation	47	0.365	0.36	0.37	6e-04	91

*The total stratified samples were supposed to be 50, but because some strata contain too few values, and because of rounding within the equation to determine sample size within each stratum, the total samples are not consistent with the random sampling.

3 Testing for Normality:

I used the Kolmogorov–Smirnov test to determine whether the runoff ratio is normally distributed. The following hypotheses are proposed to test for normality of the runoff ratio values:

H_0: the runoff ratio is NOT normally distributed
H_1: the runoff ratio IS normally distributed

To test these hypotheses I used the kstest() function in MatLab. I tested the normality on the raw data, with box-cox and log transformations, and while standardizing each of these tests. To standardize I subtracted the data by the mean and divided the difference by the standard deviation:

$$[H, P] = kstest(\frac{(rr - mean(rr))}{std(rr)}) \tag{1}$$

where H is the result of the null hypothesis, P is the associated P-Value and rr is the runoff ratios with any indicated transformation.

Table 6: Kolmogorov-Smirnov test to determine whether the runoff ratio is normally distributed

Transformation	Standardization	H_1 Rejection status	P Value
None	No	Rejected	O^{-142}
None	Yes	Rejected	0.0004
Box-Cox	No	Rejected	O^{-169}
Box-Cox	Yes	Rejected	0.005
Log	No	Rejected	O^{-178}
Log	Yes	Rejected	O^{-15}
Log	Yes*	Not Rejected	0.058

*Includes a shift of 1.2 to runoff ratio values before log transformation.

Figure 5 shows the empirical and theoretical cumulative distribution function for the runoff ratios with no transformation. The black line shows that the raw values are not even close to a normal cumulative distribution. The blue line shows that with a standardization the cumulative distribution curve tends to match well, visually at least, with the theoretical curve. This plot is repeated in Appendix A to include the Box Cox transformation and a log transformation. As Table 6 indicates, it takes a shift made to the runoff rations, and a log transformation to pass the Kolmogorov–Smirnov test, rejecting the null hypothesis stated above.

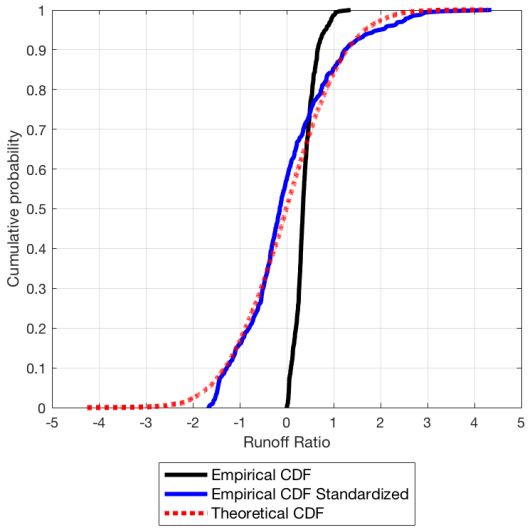

Figure 5: Cumulative distribution of Runoff Ratio

4 Hypothesis Testing:

I tested the runoff ratio values for a statistically significant difference forested sites vs. sites with other land cover types. Forested land cover has a vegetation class of 1, 2, 11 & 12. The runoff ratios were split between values corresponding to these land cover types (forested) and the other 10 types (not forested). MatLab includes a function to test two sets of data for inequality using the ttest2() command, which I used to test the following set of hypotheses.

H_0: there is NOT a significant difference between runoff ratios at forested sites vs. other land cover types
H_1: there IS a significant difference between runoff ratios at forested sites vs. sites with other land cover types

Student's T-test result: Forest Stratification hypothesis 1 is rejected, p value = 1e-41

This indicates that there is no statistical difference between forested and other land cover types. I would have expected a difference because the sites with a forested land cover will have more precipitation intercepted by the forest canopy. They will also take up more moisture from the soil than non-forest sites. These should change the quantity of overland flow reaching a drainage network.

I used a similar procedure to test for a statistically significant difference between runoff ratios at sites where the primary geological class is sedimentary vs. other classes. Sedimentary classes include 1, 4, 6, 7 & 8, with the rest of the geological classes representing other types of bedrock.

H_0: there is NOT a significant difference between runoff ratios at sedimentary vs. other geological classes
H_1: there IS a significant difference between runoff ratios at sedimentary vs. other geological classes

Geological stratification hypothesis 1 is rejected, p value = 7e-13

This indicates that there is no statistical difference between sedimentary and other geological types. I would have expected a difference because watersheds with sedimentary geological classification should have higher interception rates. Sedimentary rocks are softer, more porous and should allow for more vegetation growth.

I repeated the procedure once more for sites with unconsolidated sediments: geological class 7, vs. all other geology types.

H_0: there is NOT a significant difference between runoff ratios at sites with unconsolidated sedimentary vs. other geological classes
H_1: there IS a significant difference between runoff ratios at sites with unconsolidated sedimentary vs. other geological classes

Unconsolidated sedimentary hypothesis 1 is rejected, p value = 3e-17

This is the most surprising result. Unconsolidated sedimentary land cover should have a much higher hydraulic conductivity, which will allow a lot more precipitation to infiltrate into the ground, rather than runoff over the surface of the watershed.

A Appendix A: Bonus figures

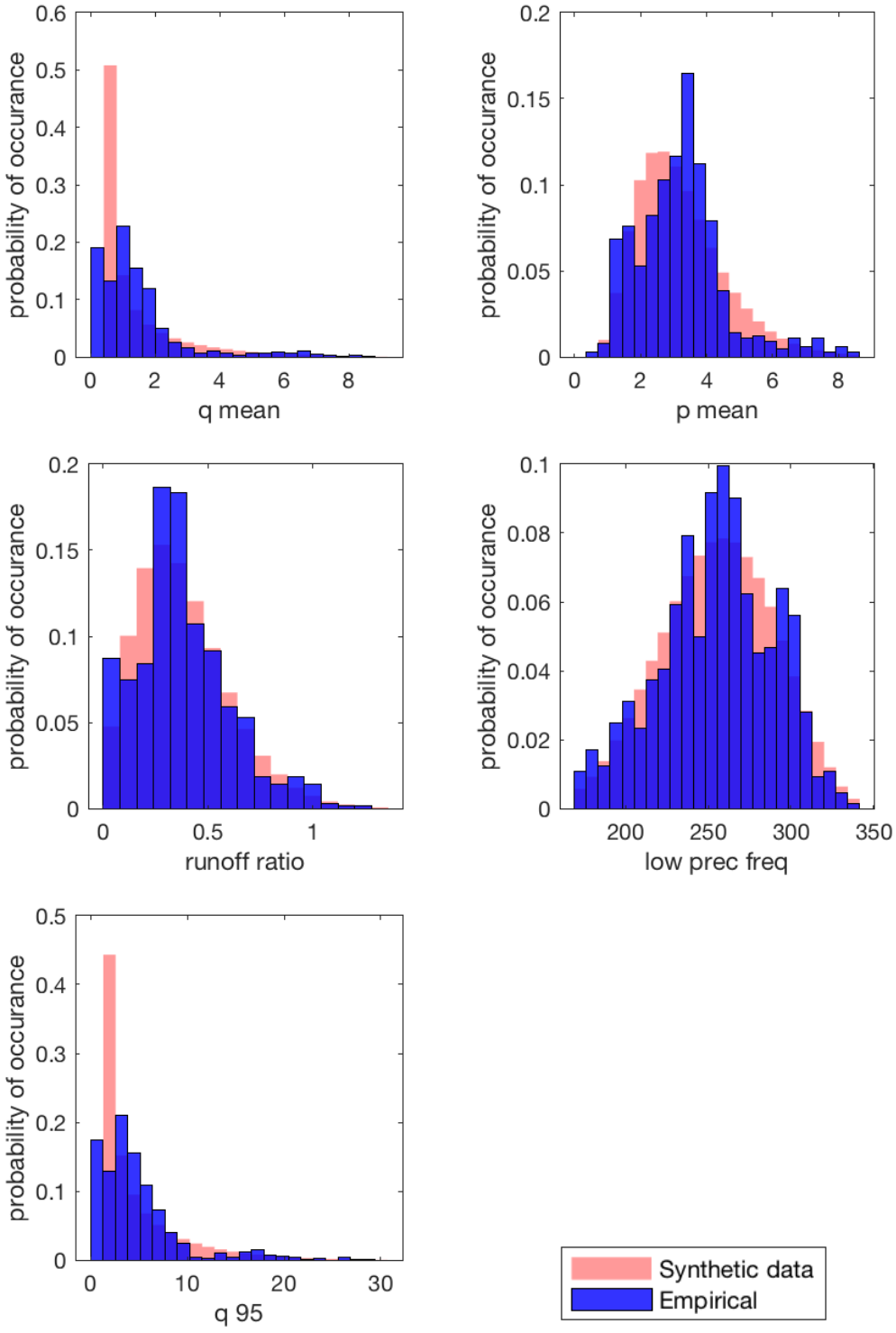

Figure 6: Histograms of high correlation variables

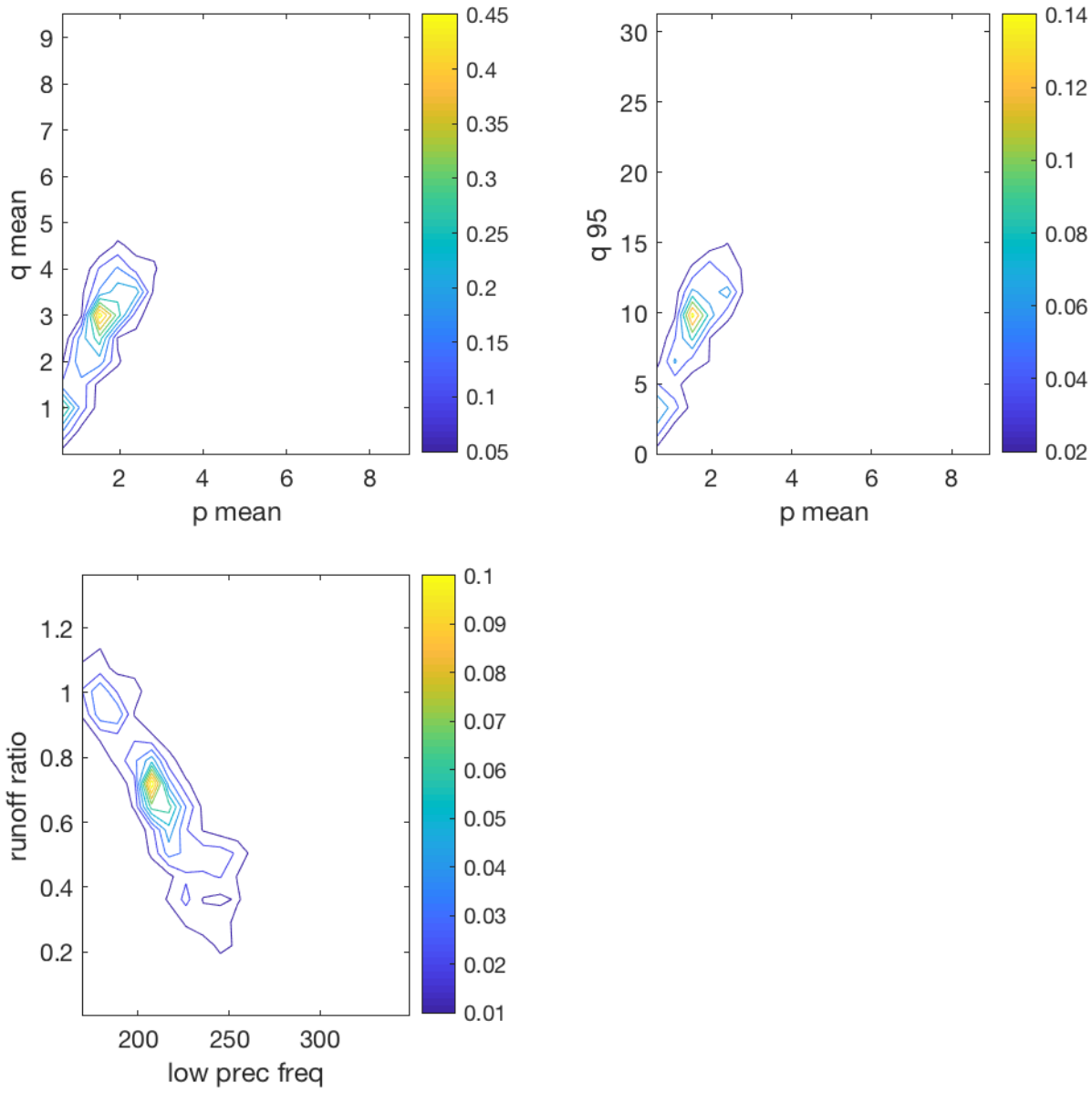

Figure 7: Contour plots of joint distributions for each of the three highest Attribute/Signature correlation pairs

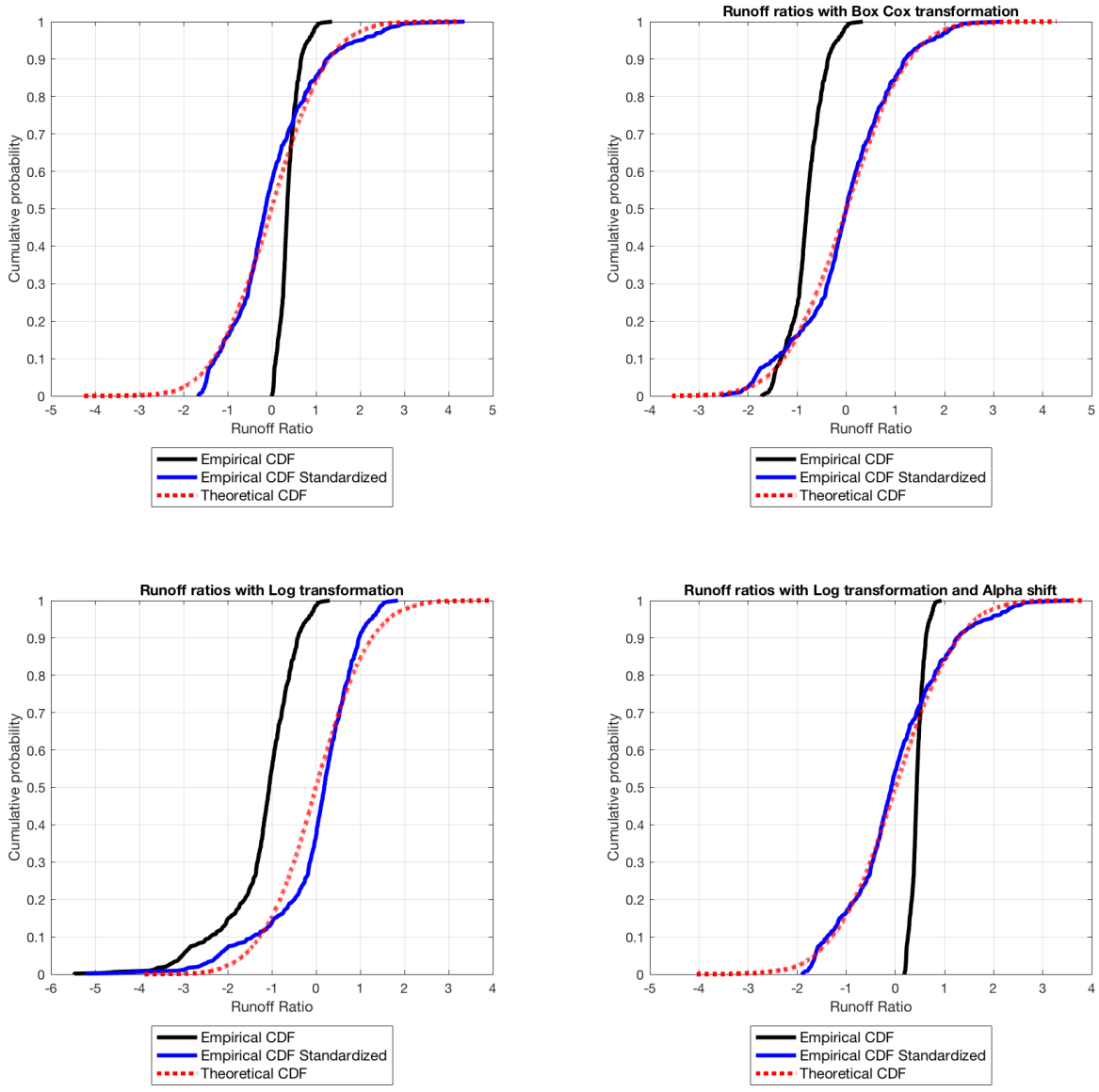

Figure 8: Cumulative distribution of runoff ratio, including transformations and standardizations