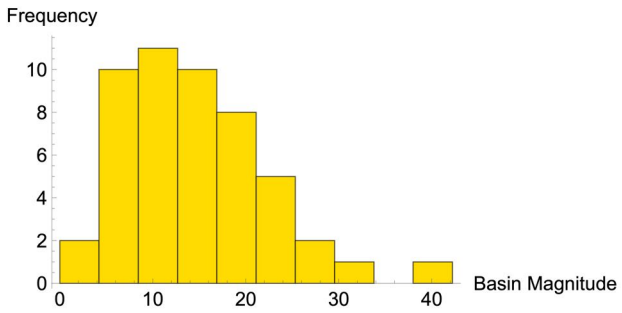
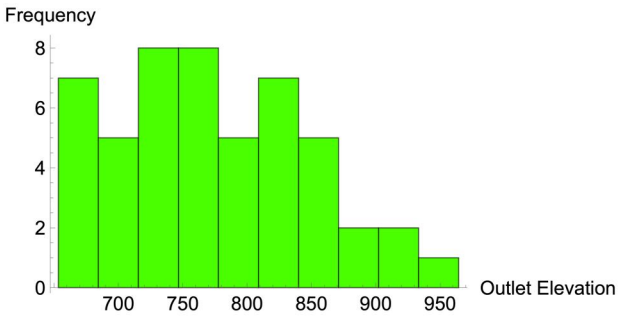
Data Analysis & Statistics for Earth Scientists
Nanjing University, Spring 2026

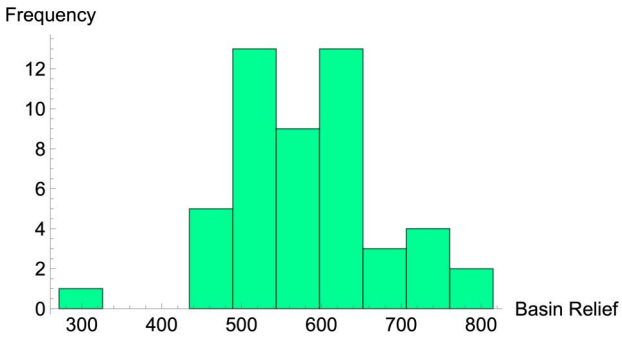
Lab 1 Assignment

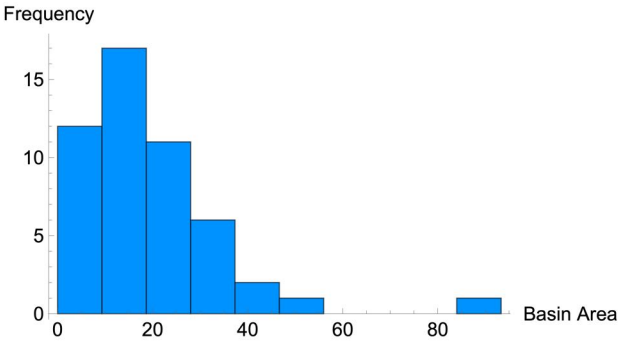
1. Open the Kentucky.dat datafile in the PAST program or the Kentucky.csv file in another mathematical application (e.g., MS-Excel, *Mathematica*™, MatLab™).

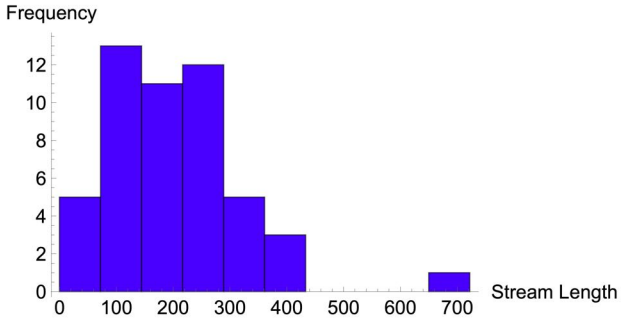
2. Create 10-bin histograms for all seven variables. (60 points)

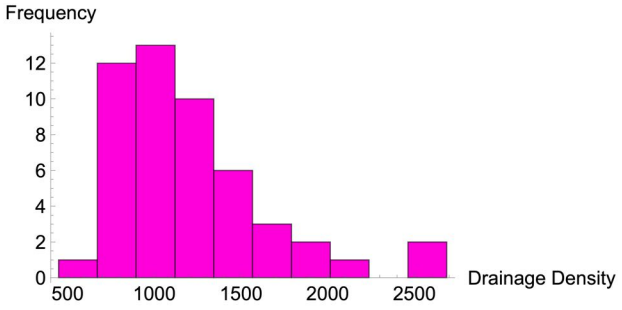


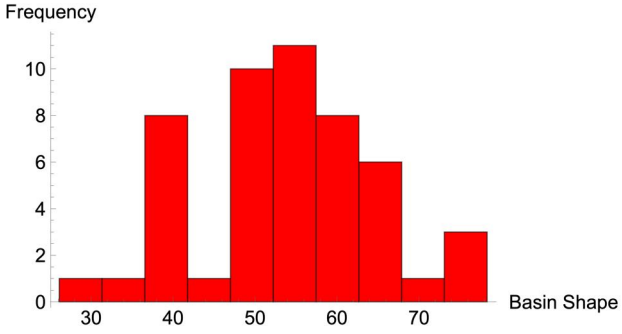




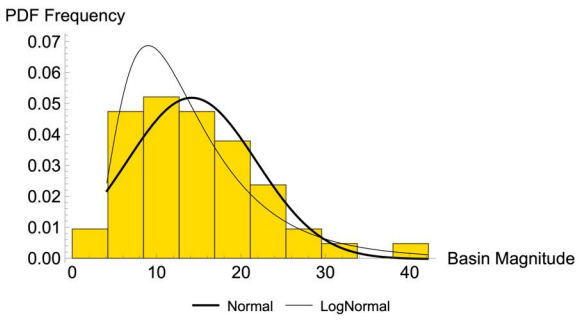


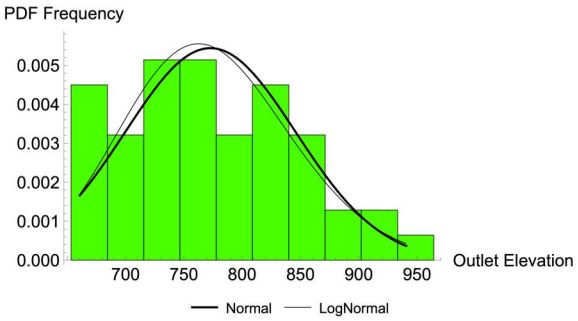


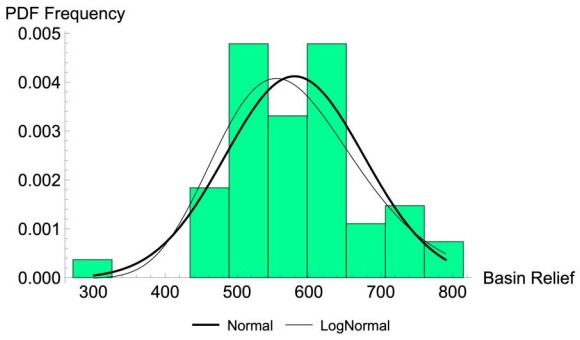


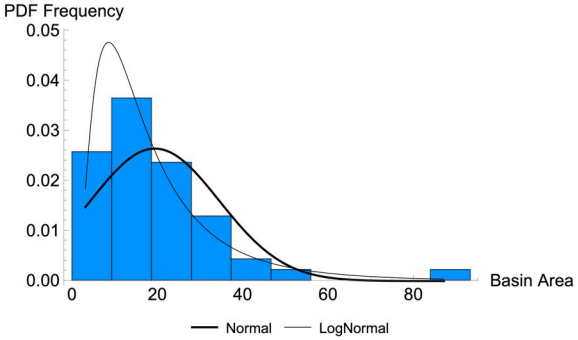


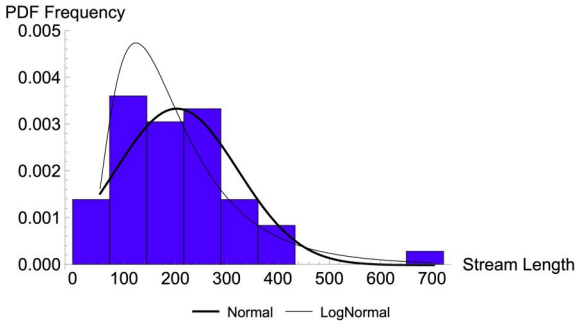
a. Identify the variable whose distribution most closely resembles a normal distribution. Be sure to justify your answer. (10 points)

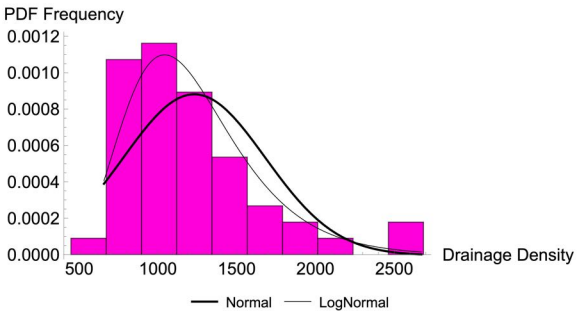


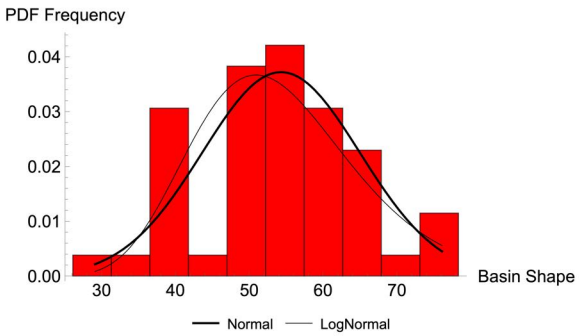










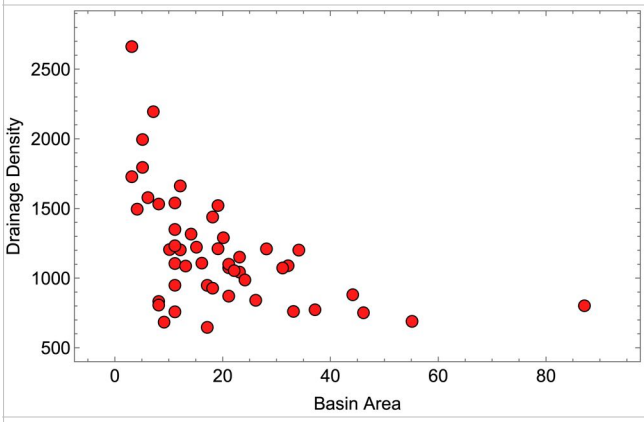


Normal Distribution Test Table

Variable	Mean	Std. Dev.	Anderson-Darling A^2	Crammér- von Mises ω^2	Pearson χ^2
Basin Magnitude	14.0600	7.6614	0.3886	0.4073	0.0757
Outlet Elevation	772.1000	73.0485	0.7038	0.6697	0.5749
Basin Relief	578.8600	96.4409	0.6577	0.6998	0.1719
Basin Area	19.2600	15.0716	0.0772	0.0971	0.0118
Stream Length	201.3600	119.2852	0.3385	0.4059	0.0669
Drainage Density	1225.8000	450.5827	0.1439	0.1798	0.0235
Basin Shape	54.2000	10.6846	0.8480	0.8148	0.7792

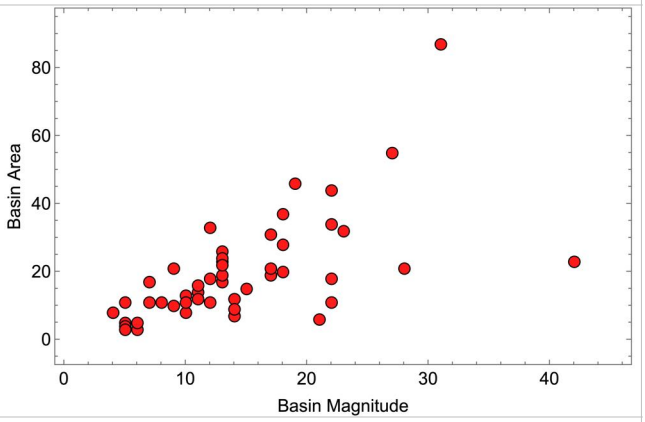
3. Create scatterplots for the following variable pairs. (40 points)
   - Basin Magnitude vs. Basin Relief
   - Basin Magnitude vs. Basin Area
   - Basin Relief vs. Stream Length
   - Basin Area vs. Drainage Density

Basin Area vs Stream Density



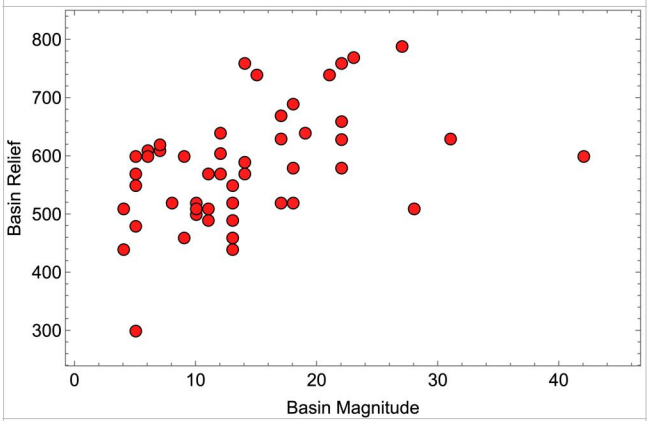
Basin Area: Low Variance, Reverse Heteroscedastic
Drainage Density: Very High Variance, Reverse Heteroscedastic

Basin Magnitude vs Basin Area



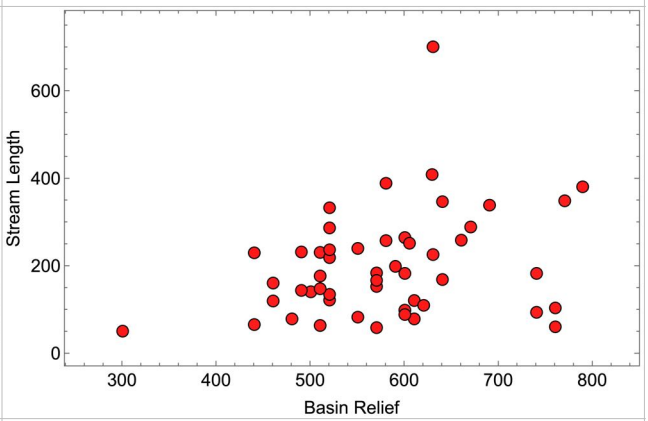
Basin Magnitude: Very Low Variance, Heteroscedastic
Basin Area: Low Variance, Heteroscedastic

Basin Magnitude vs Basin Relief



Basin Magnitude: Very Low Variance, Homoscedastic
Basin Relief: High Variance, Heteroscedastic

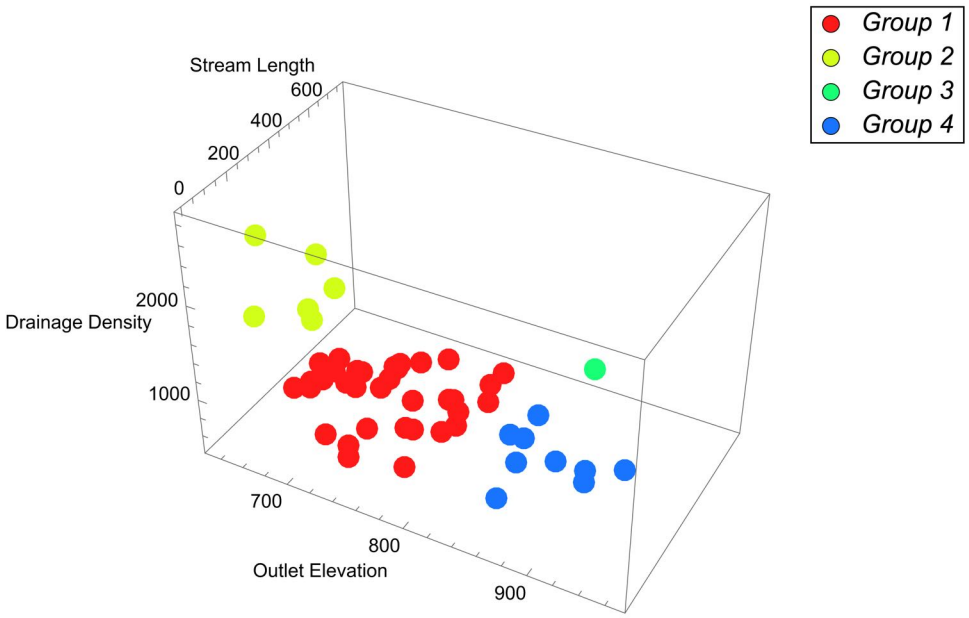
Basin Relief vs Stream Length



Basin Relief: High Variance, Homoscedastic
Stream Length: High Variance Homoscedastic

a. Characterize each scatterplot using the following categories: wide scatter, narrow scatter, boundary-constrained scatter, categorical scatter, homoscedastic scatter, heteroscedastic scatter. Remember each scatterplot involves two variables. (40 points)

4. Create a 3D scatterplot of the Outlet Elevation, Stream Length and Drainage Density variables. (10 points)



a. Describe the geometric relations among the basins. Are there any clusters? If so identify them. (30 points)

There are four clusters of basins in this plot, though one of the clusters might be more properly regarded as an outlier.

Group 1: 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 28, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 48, 49, 50
Group 2: 1, 2, 6, 10, 11, 12,
Group 3 (Outlier): 30
Group 4: 3, 4, 7, 23, 26, 27, 29, 31, 46

The fact that, within the groups, there exists a strong tendency toward runs of consecutive numbers suggests there may be a fairly strong tendency for adjacent drainage basins to have similar characteristics.

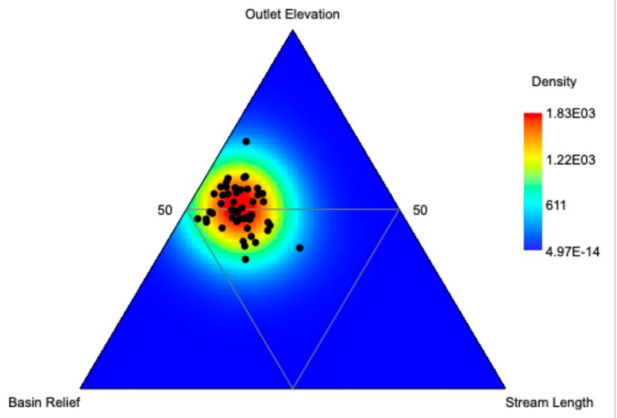
5. For each variable, list or calculate the following parameters and present these summaries in the form of a table: maximum value, arithmetic mean, geometric mean, median, minimum value, range of observations, variance, standard deviation, coefficient of variation. (20 points)

Descriptive Statistics Table

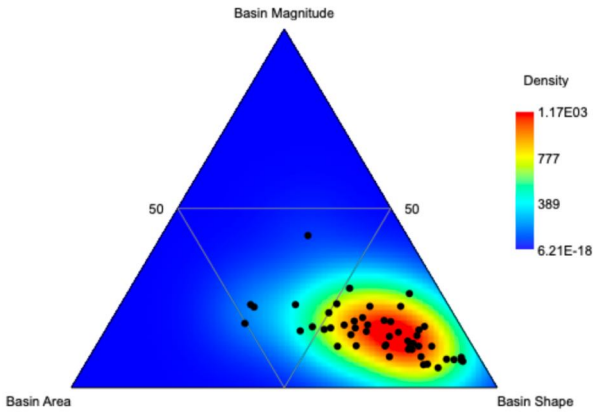
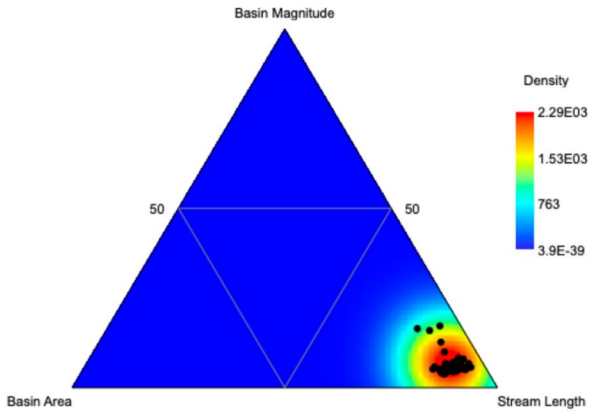
Statistic	Basin Magnitude	Outlet Elevation	Basin Relief	Basin Area	Stream Length	Drainage Density	Basin Shape
Maximim	42	940	789	87	702	2667	76
Minimum	4	660	300	3	52	652	29
Range	38	280	489	84	650	2015	47
Arithmetic Mean	14.06	772.10	578.86	19.26	201.36	1225.80	54.20
Median	13.00	755.00	575.00	16.50	181.00	1111.50	55.00
Mode	12.21	750.00	506.46	9.00	162.50	1223.39	55.39
Geometric Mean	12.13	768.71	570.52	14.77	170.92	1157.04	53.08
	1.93	3.39	8.34	4.49	30.44	68.76	1.12
Harmonic Mean	10.35	765.38	561.62	11.03	144.22	1099.08	51.89
Variance	59.89	5444.99	9490.65	231.79	14519.34	207168.16	116.49
Standard Deviation	7.74	73.79	97.42	15.22	120.50	455.16	10.79
Coefficient of Vairation	0.55	0.10	0.17	0.79	0.60	0.37	0.20

6. Using the table you created above, answer the following questions. (35 points)
   a. Variable with the greatest range? Drainage Density
   b. Variable with the smallest range? Basin Magnitude
   c. Variable with the greatest difference between its arithmetic and geometric mean? Drainage Density
   d. Variable with the greatest variance? Stream Length
   e. Variable with the least variance? Basin Magnitude
   f. Variable with the greatest variability? Basin Area
   g. Variable with the least variability? Outlet Elevation

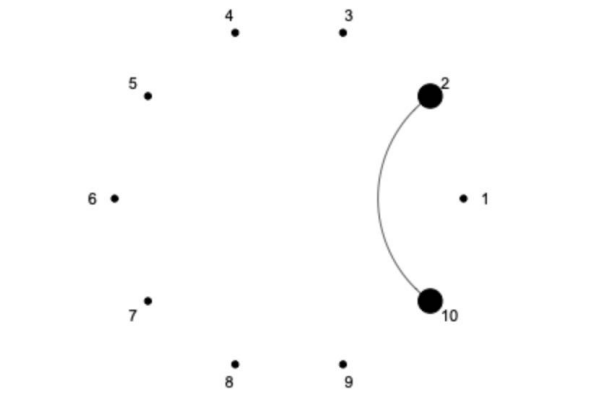
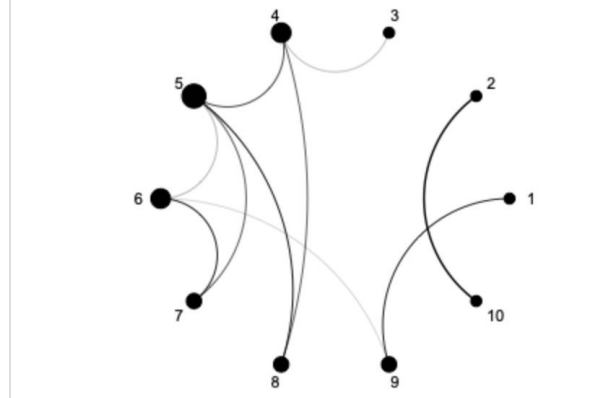
7. Create ternary density plots of the following variable triplets. (30 points)
   - Outlet Elevation, Basin Relief, Stream Length
   - Basin Magnitude, Basin Area, Basin Shape
   - Basin Magnitude, Basin Area, Stream Length
   a. Using the plots you created above, identify any evidence for data clustering and provide an interpretation for the patterns seen in the plots (in terms of the influence of different variables. (20 points)

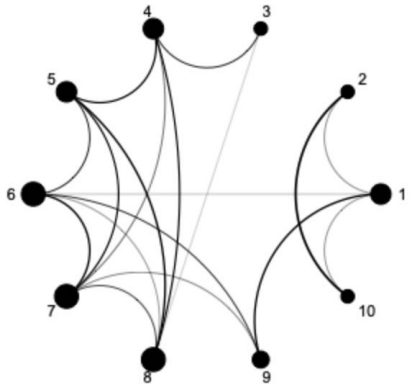
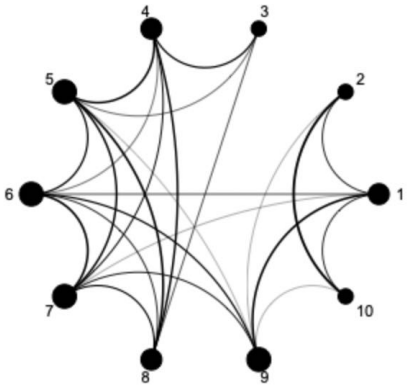
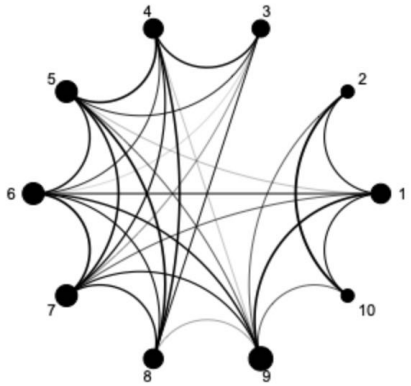
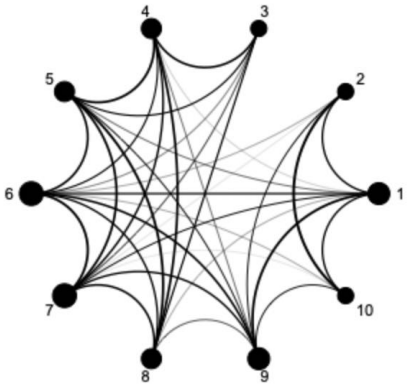
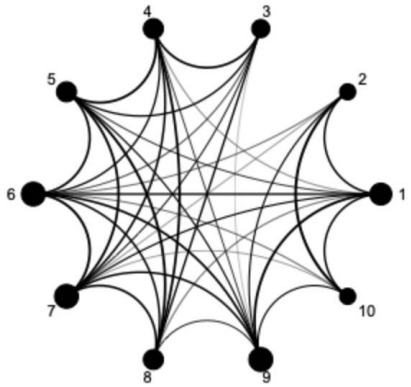
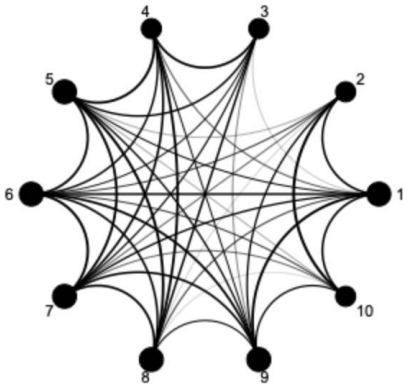


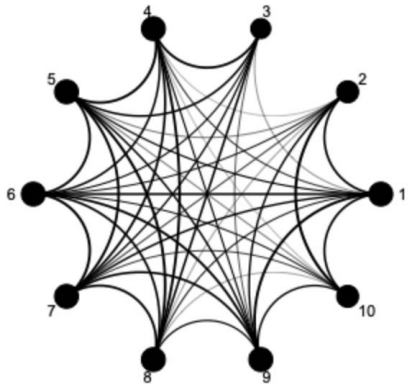
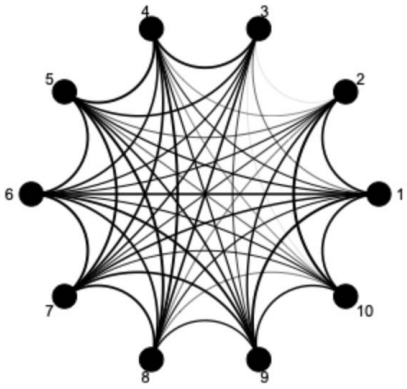
Two outliers (basins 30 and 34) but little evidence of strong clustering.

	One outlier (Basin 24) and one outlying cluster of three basins (basins 25,30 and 50).
	One distinct outlying cluster (basins 19, 20, 21) and two basins that exhibit intermediate characteristics (basins 23 and 24). The fact that these are all consecutively numbered basins suggest a strongly localized geographic distinction in the case of these five basins.

8. Make a network plot of the first ten basins using all variables. (10 points)
   Based on the plot you created above, along with the data table, identify and explain your reasoning for the following identifications. List all basins with equivalent characteristics. (30 points)

	
90	80

	
70	60
	
50	40
	
30	20

	
10	0

a. Most distinct basin. Basin 1 - Fewest number of high linkages
b. Least distinct basin. Basin 6 - Greatest number of low linkages

Basins	Strong links	Intermediate Links	Weak Links
4	7	2	-
5	7	2	-
8	7	2	-
7	7	2	-
6	6	3	-
9	8	1	-
1	6	3	-
3	5	2	2
2	3	3	3
10	4	4	1

c. Set of most similar basins. Basins (((4, 5, 8, 7), (6, 9)), ((1, 3), (2,10))) - Greatest similarity of linkages

Total = 335 points