University of Colorado
Department of Civil, Environmental and Architectural Engineering
CVEN 5454 Quantitative Methods

R-Project

Due – by December 12th, 2025

The aims of this project are two-fold, (i) to develop your R-skills – with the syntax, commands, writing functions/codes in R and, (ii) to enhance your understanding of the statistical and probabilistic concepts and methods.

R-commands and functions should be clear and well documented for each problem. Also, the figures generated must be clear with captions and legends. 20% of points will be awarded for good presentation of R-commands and developing R-functions. I expect to see each of you develop your own library of functions in the process of completing this project.

All the data for the R-project can be obtained from here
(Monthly Lees Ferry flows are in AF (Acre-Feet). Divide the values by 10^6 to convert them to MAF (Million Acre-Feet))

1. Boxplot the monthly streamflow data of LeesFerry gauge on the Colorado River and overlay the monthly mean and median. (All the boxplots should be in one figure). What do you infer?
2. Calculate the mean, variance, skew and their robust counterparts of the Lees Ferry streamflow for each month. Plot them and comment on what you find (particularly, in conjunction with the boxplots from 1). What do you infer?
3. Do pair-wise scatterplot of each month – e.g., January vs February; February vs March, …, December vs January – along with a default smoother (smooth line through the scatter). For May flows as a function April and March flows do a 3-D scatterplot. Describe your findings.
4. Select three months – May, Aug and Dec – for these monthly flows fit a Normal, Gamma, Log Normal and Weibull distributions.
   (a) Test the goodness of fit of each of the fitted PDFs visually:
   Histogram with fitted PDF overlaid
   Q-Q plot
   And using the K-S test
   (b) Repeat (a) with a nonparametric (Kernel Density) PDF. Compare with the best parametric PDF you select from above and comment on your findings regarding best fit parametric and nonparametric PDFs
5. For the May streamflow at Lees Ferry
   - Generate 500 Monte Carlo samples from the best fit PDF (from above), each of same length as the observed data used in the PDF fit. Estimate the mean, standard deviation, skew and IQR for each Monte Carlo sample and boxplot them along with the corresponding statistics of the observed data shown as points.
   - Also, overlay the PDFs from the simulations in light grey with that of the historic data as solid line, to see how well the simulations capture the historic PDF.
6. Repeat 5 but using a nonparametric approach – 'smoothed bootstrap' or 'bootstrap'. Comment on the utility of the two Monte Carlo approaches in problems 5 and 6.
7. Fit a Copula to the April-May streamflow bivariate data
   (a) Show the bivariate PDF along with the data.
   (b) Generate 500 samples, each of the same length as the historic data. Boxplot the statistics and PDFs from the simulations along with that of the historic data for April and May flow (like problem 6).
8. Annual maximum streamflow at many sites across the country can be obtained from the USGS site
   Select a gauge that interests you. Preferably, gauge higher up in the basin (i.e. headwater with limited human influence) and with longer record

To this annual maximum flow compute 50-yr, 100-yr and 500-yr return period rainfall by fitting extreme value distributions to this data – (a) Gumbell EV-1, (b) Log-Pearson Type III, (c) Log Normal and (d) Generalized Extreme Value distribution. Use the 'fExtremes' or the 'Extremes' toolkit library in R. (d) Plot the histogram of the data and overlay the four fitted PDFs and develop four Q-Q plots - on the distribution.

9. Losses due to Hurricanes (Billion $) in the US for the period 1904-2017 is available here (See Weinkle et al. 2018 for more details). Use the data in the second sheet Convert the data into Billions by diving by 10^9, blanks mean $0 damages. Fit an appropriate model (i.e. GPD) to this data for damages exceeding $5B and plot the histogram of the data with the fitted PDF. Also compute the 100-year return period damage. Fit the model with thresholds of $2B and $6B and document their sensitivity and what might be an apt threshold?

10. Summer season (Jun-Sep) rainfall (cm) on a 0.25° x 0.25° grid over India is available for the period 1901 to 2016. For the full period (1901 – 2016) and recent (1980 – 2016)

(a) Compute the Mann-Kendall trend in rainfall at each grid and plot the trends (cm/yr) along with statistical significance at 95% confidence level.

(b) Correlate summer ENSO index with summer season rainfall at each grid and spatially map the correlation values.

(c) Compute All India rainfall (i.e. sum the rainfall over all the grids to get a single value for each year). Select the top 10% years with high rainfall and the bottom 10% with low rainfall. Average the rainfall over these years at each grid point and plot them as spatial maps for the high and low rainfall years separately. There are called 'composite' maps and will show the spatial pattern of rainfall during an All India wet or All India dry year. Spring season (Apr – Jul). Also correlate the All India Rainfall with the grid rainfall and plot the correlation map.

11. Streamflow at Lees Ferry is known to be driven by the preceding winter season climate phenomenon (i.e. predictors) – Pacific Decadal Oscillation, Atlantic Multi Decadal Oscillation and El Nino Southern Oscillation and April 1st Snow Water Equivalent (SWE). Flow from the preceding year's spring season is also a potential predictor that captures the year-to-year dependency.

The goal is to develop a robust functional relationship between flow and the best suite of predictors. To this end, the following steps need to be performed.

(a) Scatterplot spring flow with all the potential predictor variables and apply a default smoother in R to smooth the scatterplot and comment on the relationships you see (linear, nonlinear etc.).

(b) Using GCV or PRESS or AIC or BIC obtain the *best linear model* and perform model diagnostics including ANOVA. Also plot the historical and modeled values with 95% confidence interval along with 1:1 line for visual inspection of the model performance. What do you infer from the best model – i.e. the variables selected in the model.

(c) Estimate the skill of the model by repeatedly dropping 10% of points - drop 10% of points, fit the best model to the rest and predict the dropped points; compute skill measures, R^2, RMSE; repeat. Boxplot the skill measures.

(d) Bootstrap the data and obtain the estimates of the model parameters for each bootstrap sample. Repeat 500 times. Plot the parameter estimates as a histogram along with the estimate from the original sample.

12. Problem 7-48 from 'text' using R-JAGS - Bayesian library/package.

13. Problem 7-50 from 'text' using R-JAGS - Bayesian library/package.

For both of the above problems, plot the histogram of the posterior samples of the parameters and compute the mean and the 95% confidence interval.

14. Fit a Bayesian Regression for problem 11 using all three predictors. Plot the posterior distribution of the coefficients; the estimate of Spring streamflow along with 95% confidence intervals. Compare the posterior distributions with that from 11 (d)