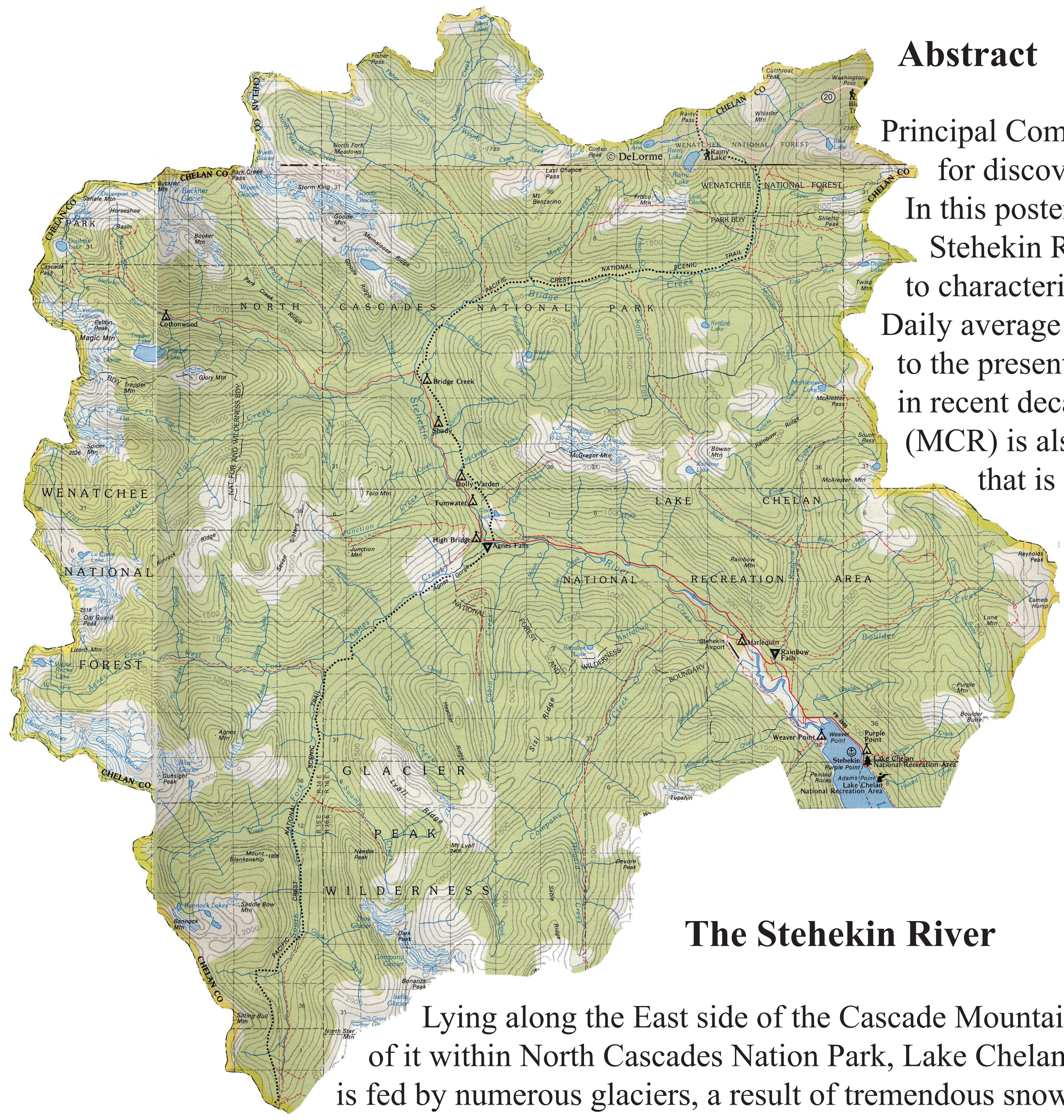
Analysis of Historical Stehekin River Flow Data with Principal Components Analysis and Multivariate Curve Resolution

Clare E. Wise and Barry M. Wise



Abstract

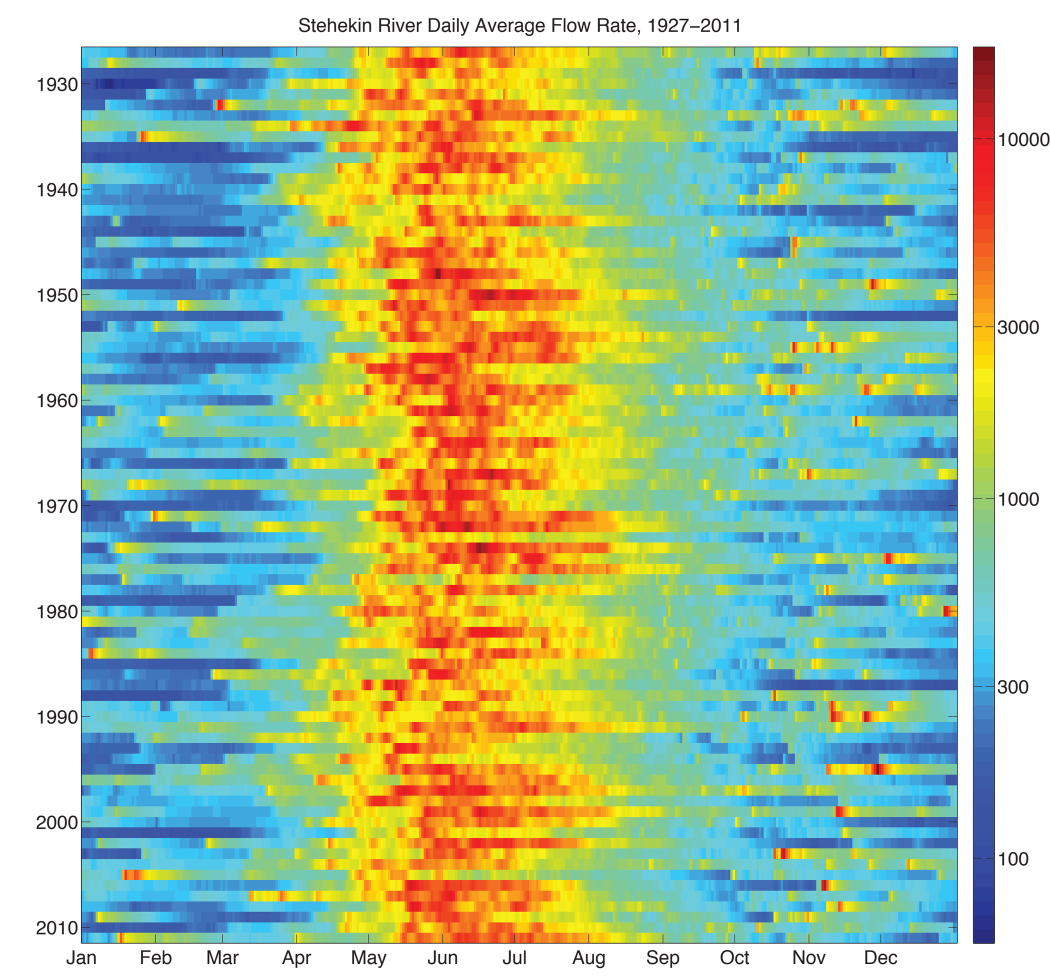
Principal Components Analysis (PCA) is a powerful tool for discovering and visualizing trends in multivariate data. In this poster, PCA is applied to the daily flow rate of the Stehekin River (Chelan County, Washington, USA) in order to characterize its natural variation over the course of a year. Daily average flows are available starting in 1927 and continuing to the present day. The variation over the years, and any trends in recent decades, is also of interest. Multivariate Curve Resolution (MCR) is also applied to the river data in order to get a decomposition that is non-negative.

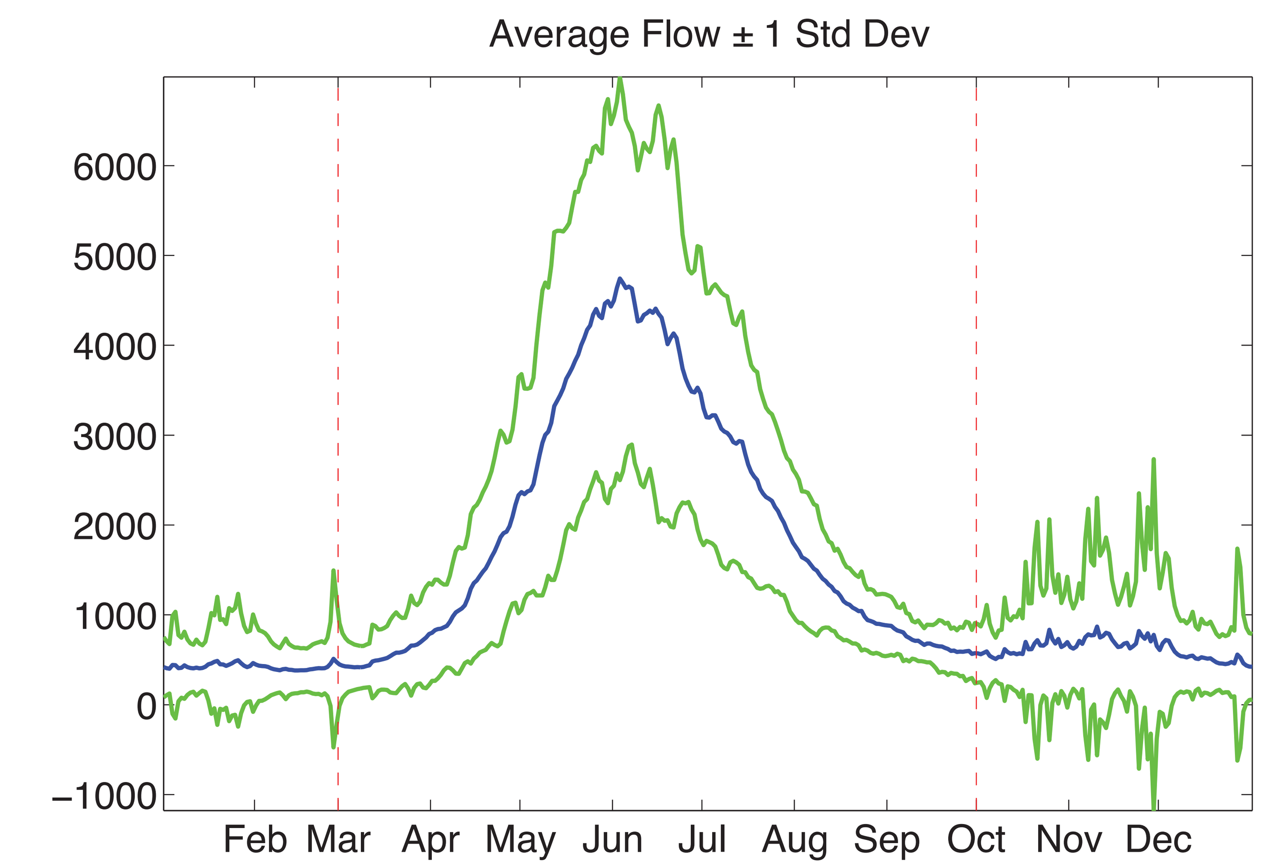
The Stehekin River

Lying along the East side of the Cascade Mountains, the Stehekin River watershed covers 321 square miles, much of it within North Cascades Nation Park, Lake Chelan National Recreation Area, and Glacier Peak Wilderness. The river is fed by numerous glaciers, a result of tremendous snowfalls typical of the Cascade Crest. It is one of the most wild and beautiful bodies of water in the 48 states. Map shows extent of river watershed. Photo shows lower Stehekin River river and head of Lake Chelan from Librarian Ridge on McGregor Mountain.

River Flow Data

Average daily flow rates are available for the Stehekin starting in late 1926. Flow rates are taken at a gauging station approximately 1 mile from the mouth of the river. As the figures below illustrate, the flow is quite variable. Flows typically range from 200 CFS to 10,000 CFS over the course of a year. Spring floods are common, however, a surprising number of Fall floods also occur, including the largest flow ever recorded of 25,000+ CFS on October 20 of 2003.

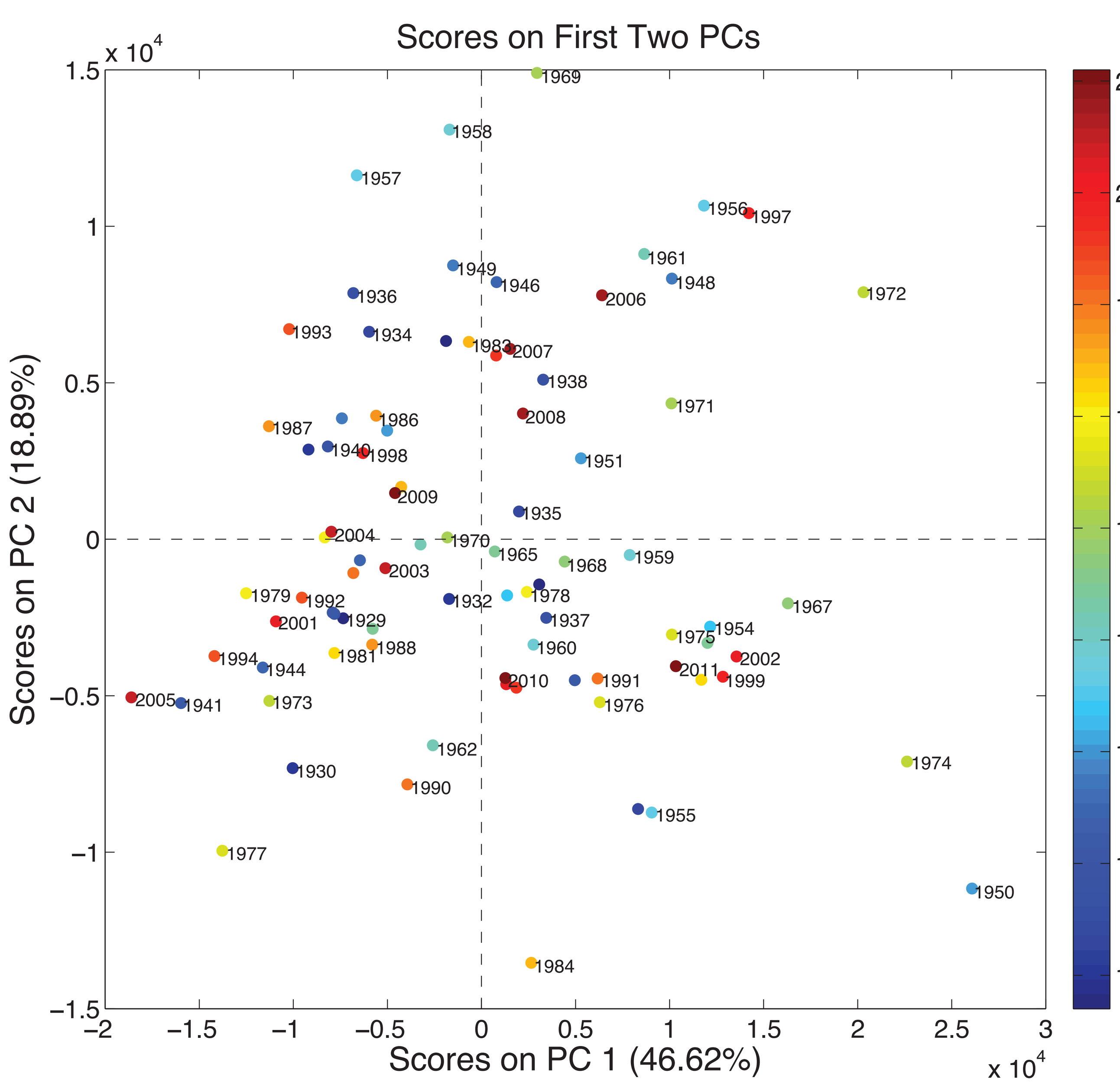


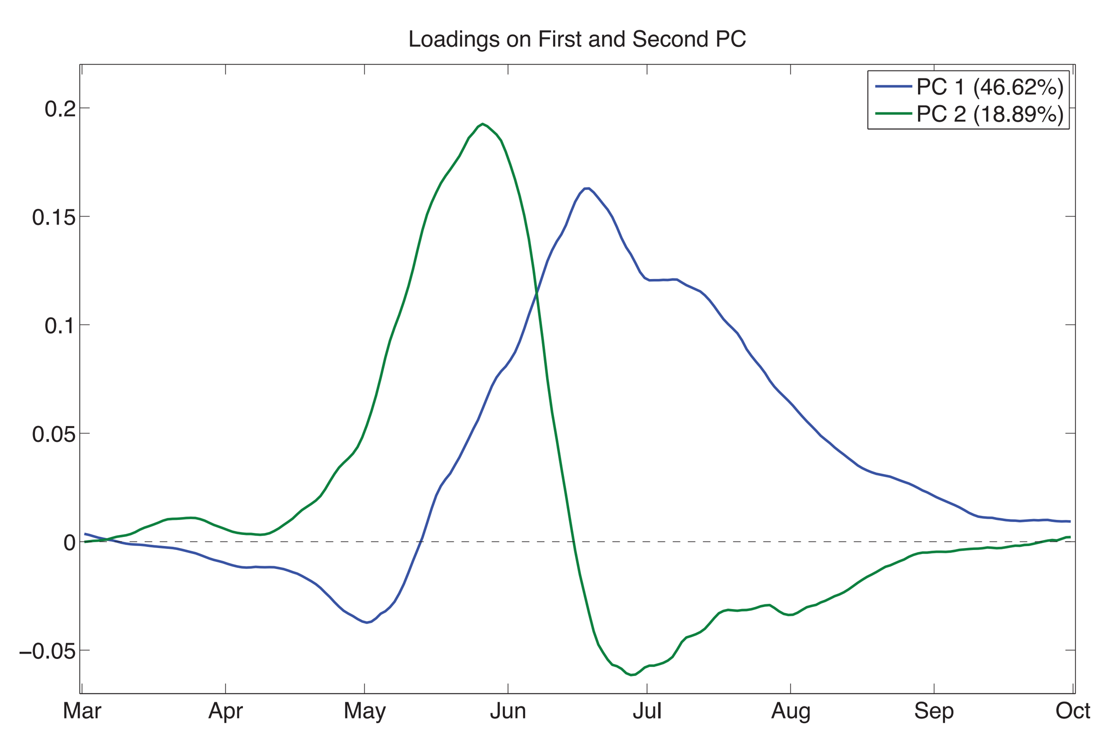


Average flow showing ± 1 Standard Deviation. Note large variation due to fall floods. Range of data used in this study is indicated with the dashed red lines.

PCA of March-October Flow Data

PCA can be used to produce plots that capture much of the variability in the river flow data. Scores and loadings from PCA of 15-point smoothed, mean-centered data are shown below. The loadings from the first PC can be interpreted as late Spring flow while the second PC indicates early Spring flow with some decrease in later flow. Interesting years can be easily found. For instance, the year 1950 had a very late Spring runoff. The year 1977 had the lowest average flow on record, and thus had low flows in both early and late spring. Years 1957 and 1958 had very early runoff.



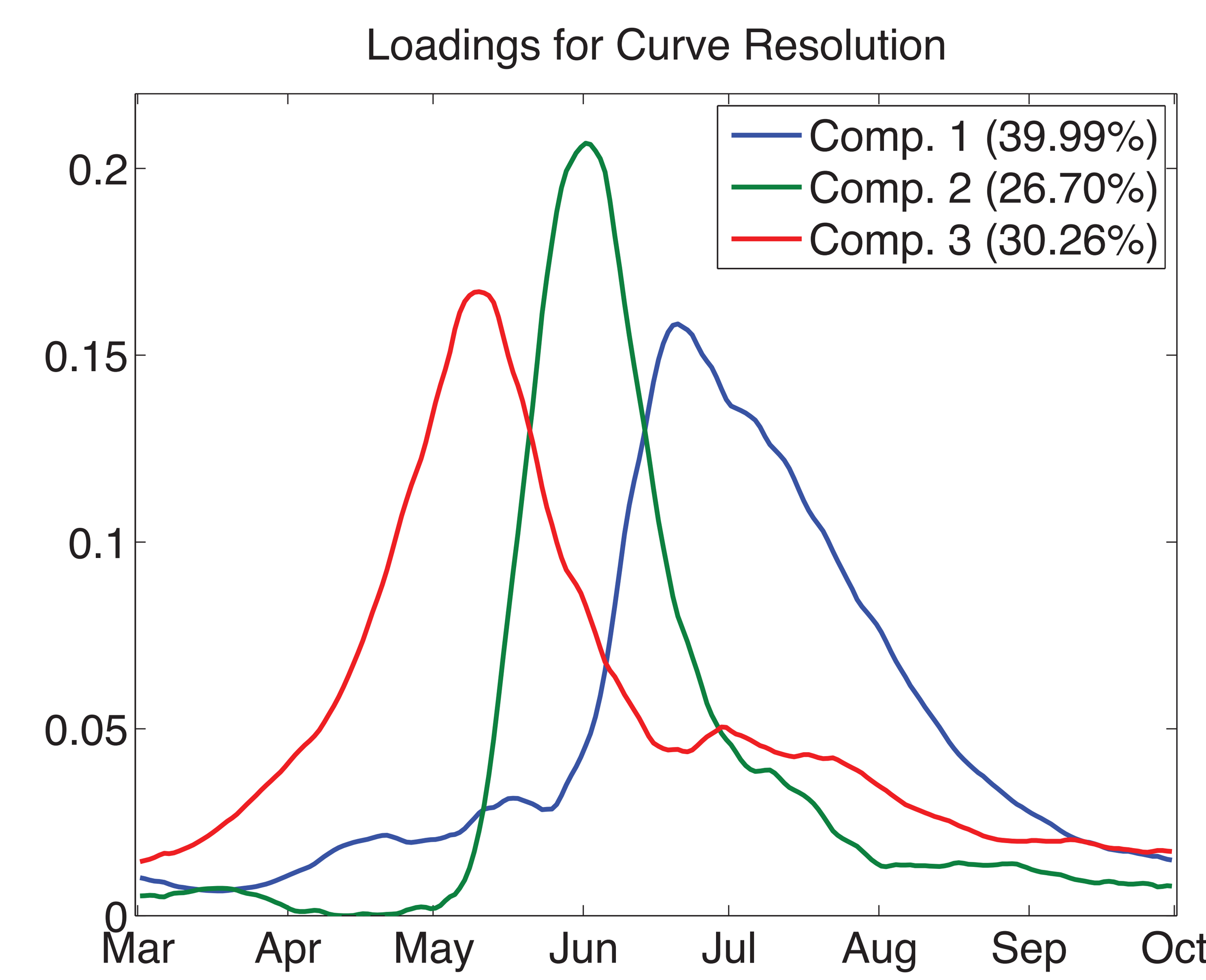


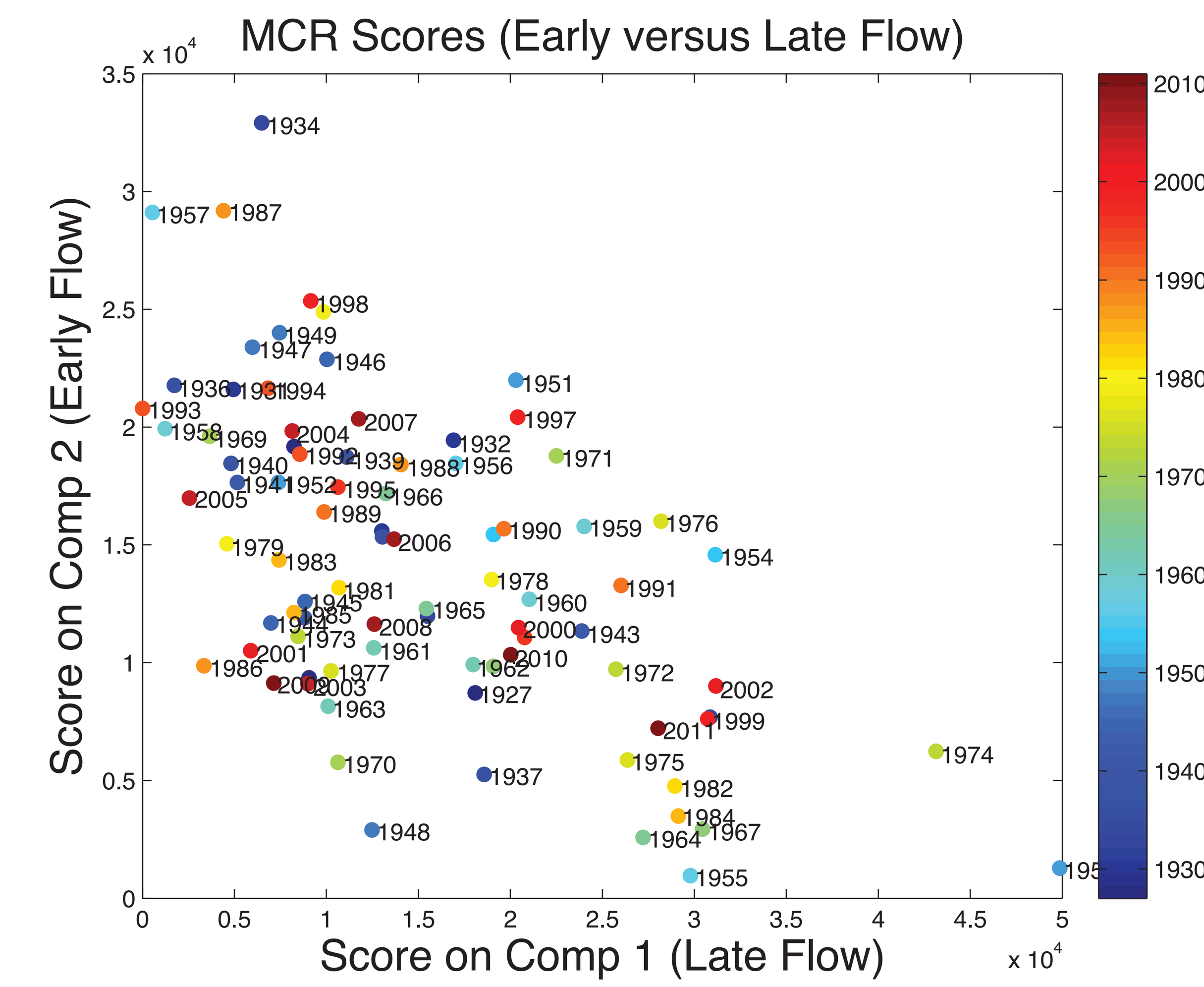
Average Scores of the Last Ten Years

The average scores for the last 10 years, 2002-2011, were calculated. The average scores of -0.0010e4 on PC 1 and 0.0013e4 on PC 2 put the point so close to the center of the plot that the plot symbol would cover the cross-hairs.

MCR of Flow Data

Multivariate Curve Resolution can be used to obtain a non-negative decomposition of the flow data. A three factor MCR model captured 97% of the total sum-of-squares of the 15-point smoothed March-October flow data. The MCR loadings show how the flow was broken into early, middle and late periods. A plot of the MCR scores for the early versus late factor show some negative correlation: when there is substantial early runoff, less snowpack is available to produce flow later.





Conclusions

PCA and MCR aid visualization of the flow data. The decompositions make it easy to identify years in terms of earliness/lateness of the flow and total flow. Interesting years are easily found. Data from the last decade is, on average, very similar to the entire 85 year record. No obvious trends with time were apparent.

EIGENVECTOR RESEARCH INCORPORATED