THIS PAPER HAS BEEN JUDGED

"BEST POSTER PRESENTATION"

AT THE WESTERN SNOW CONFERENCE

RECORD ESTIMATION AND EXTENSION FOR PRECIPITATION DATA

by Stephanie E. Smith[1] and Eric Weiss[2]

ABSTRACT

The desire to use data from climate stations installed in the 1980's for streamflow volume regression equations required an estimating procedure to fill in missing data and extend the records back to 1961 to create a complete 30 year record. The past success of using organic regression MOVE 1 and MOVE 2 procedures in extending streamflow records (Hirsch, 1982; Alley and Burns, 1983) led to the development of a similar application for precipitation data. A MOVE 2 estimation procedure was developed for monthly precipitation data which uses maintenance of variance techniques to minimize variance bias, and preserves seasonal variation in precipitation. Results using this procedure to estimate precipitation for climate stations located in the Bridge River system in British Columbia are presented and compared to estimates made using simple linear regression.

INTRODUCTION

The development of statistical regression equations for streamflow volume runoff forecasting, like many applications of hydrometeorological data, requires a complete set of data for a relatively long duration for a number of stations within a localized area. In developing forecast equations for the Bridge River system in southwestern British Columbia (Smith and Weiss, 1996), we wanted to utilize the many data collection platforms (DCP) stations that were installed by B.C. Hydro inside the basin in the mid-1980's. It was our hope that these stations would improve our water supply forecasting capability since previous forecasts had used long term stations generally located outside the Bridge River basin. The period of record chosen for the regression analysis was 1961-1990, making it necessary to develop a procedure to both fill in any small gaps in the data records, and extend the DCP records back to 1961. The periods of record for each of the stations in the basin are shown in Figure 1. By using organic regression to compute the estimates, and through consideration of the non-normality and seasonality of precipitation data, we hoped to develop estimates which best preserved both the mean and the natural variability of the actual precipitation.

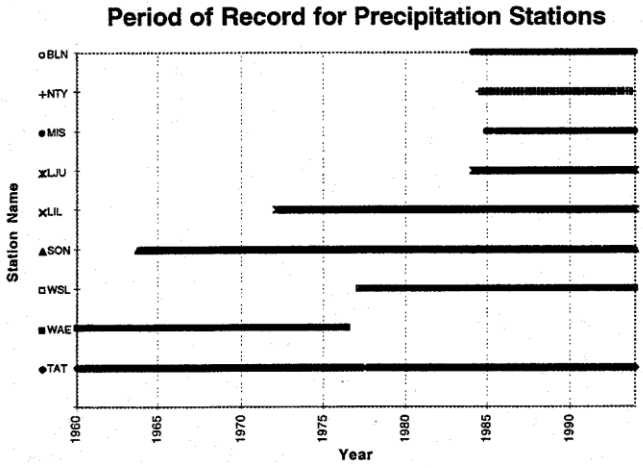


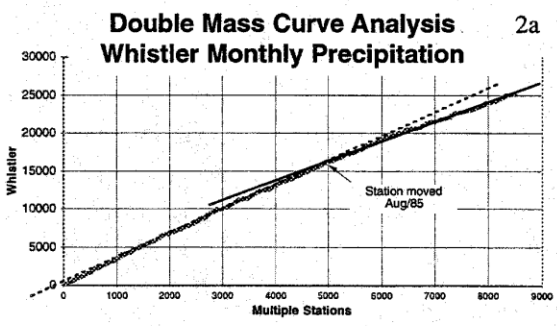
Figure 1

METHODOLOGY

Before attempting any estimates of missing monthly data, we took a closer look at the daily and even the hourly records in an effort to re-create the record at these time scales to fill in any gaps in the monthly data . Ideally, this would be done in real time, as the data are archived. Records for two stations at Whistler, WSL and WAE, were merged together to make one long record. To correct any changes in accumulation patterns caused by this merging of station records, or by other changes in station location over time, we performed a double mass curve analysis for all stations. The climate station at Whistler, for example, was moved again in 1985 which caused a

1 Hydrometeorologist, Power Supply Operations, BC Hydro

2 Senior Hydrology Engineer, Power Supply Operations, BC Hydro



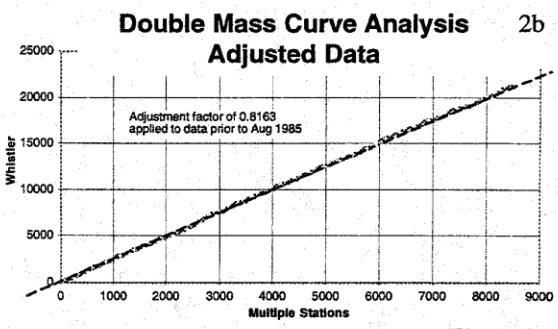


Figure 2

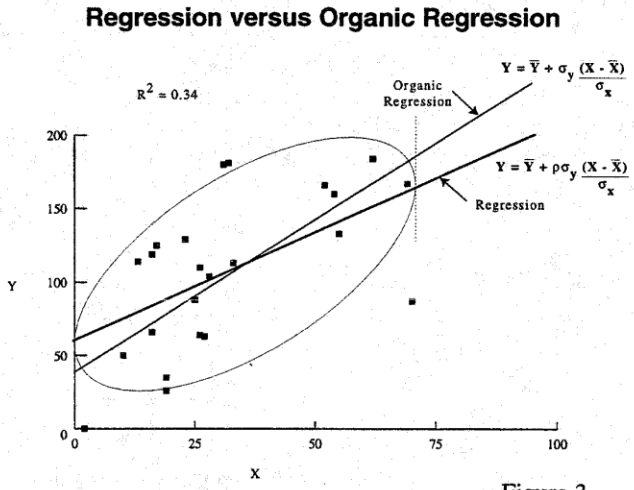


Figure 3

Statistics for the Standardized Cube Root of Monthly Precipitation Estimates for Whistler

	Mean	Standard Deviation
Observed data	1.00	1.00
MOVE 2	1.00	0.99
Linear Regression	1.00	0.77

Winter estimates based on Tatlayoko Lake, correlation coefficient = 0.75

Table 1

change in the accumulation pattern (Figure 2a). By applying a slope adjustment factor to the data prior to August 1985, the record is now consistent over time (Figure 2b).

Hydrometeorological data often demonstrates a cyclical or seasonal nature, which needs to be accounted for when attempting to estimate missing data. Estimates made without taking this seasonality into account will underestimate the natural variability at some times of the year and overestimate it at other times. After comparing the mean monthly precipitation at each station, we identified three distinct precipitation "seasons" in this region: Winter (October -February), Spring (March-May) and Summer (June through September). Estimates were made for each of the seasons independently.

Prior to applying any statistical estimation procedure, we first tested the normality of the precipitation data. The cube root transform was chosen to compensate for the positively skewed nature of precipitation data where required (Helsel and Hirsch, 1992).

Figure 3 demonstrates the theoretical difference between the statistics obtained from linear regression and organic regression. Table 1 shows the specific statistics for Whistler for both techniques. Linear regression preserves the mean of the independent variable, but the variance of the estimated data is lower than that of the observed data. Organic regression preserves not only the mean, but also the variability of

the observed data. The maintenance of variance is important, particularly for extending a record over a long period, to ensure that the estimated data mimics the actual variability of data that is received in real time.

RESULTS

Estimates were made for each of the eight stations using the best correlated station for each season using both regression estimation methods, beginning with the station with the least missing data and working through to the stations with shorter records. Figure 4 plots normalized observed and estimated data for Whistler. Upon close scrutiny, the figure shows that linear regression estimates tend to cluster around the mean, with relatively few estimates falling outside one standard deviation of the observed data. MOVE 2 estimates approximate the variability of the observed data much more closely.

An examination of the absolute error of each estimate made by both estimating procedures is displayed in Figure 5. Even though MOVE 2 maintains the variability of the whole data set, linear regression gives better individual estimates more frequently, except at the extreme high and low values.

CONCLUSION

We have adopted a standard procedure for estimating missing monthly precipitation data required to develop statistically-derived water supply forecast equations.

Precipitation at Whistler
Normalized Observed Data and Estimates

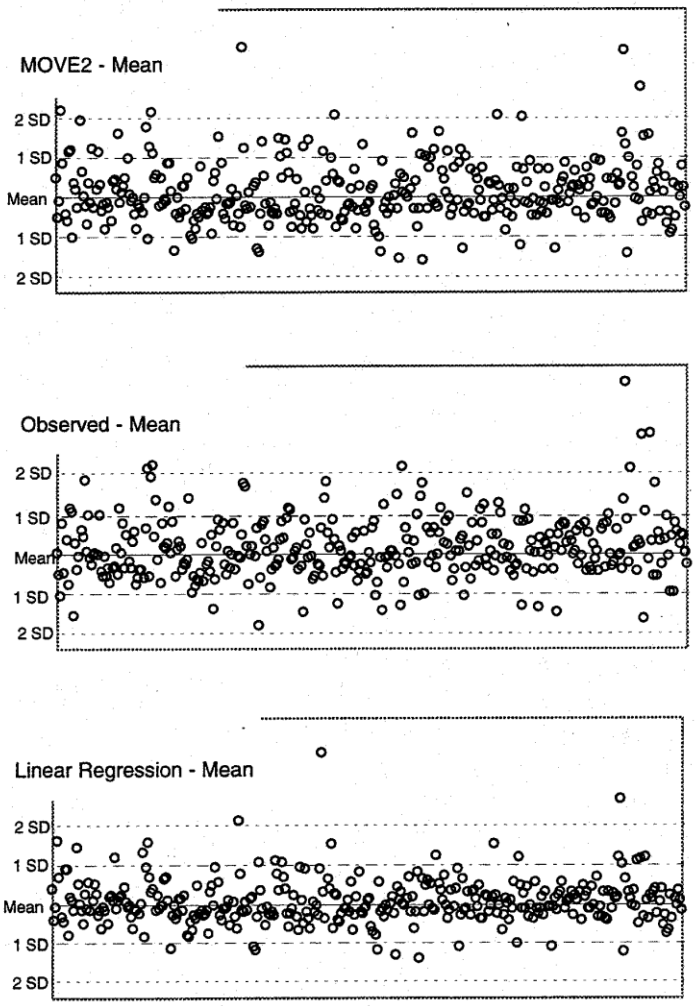


Figure 4

We adjust observed monthly precipitation data when a double mass curve analysis indicates that a change in station location has altered the catch characteristics of the precipitation gauge.

To increase the sample size of the data used in the estimation procedures, we would of course prefer to use all months of the year in a single analysis. However, temporal precipitation patterns in British Columbia vary significantly. Estimates made without taking seasonality into account will underestimate the natural variability at some times of the year and overestimate it at other times. Therefore, we group data into "seasons". The length of the season is chosen such that each month within the season has roughly the same mean monthly precipitation.

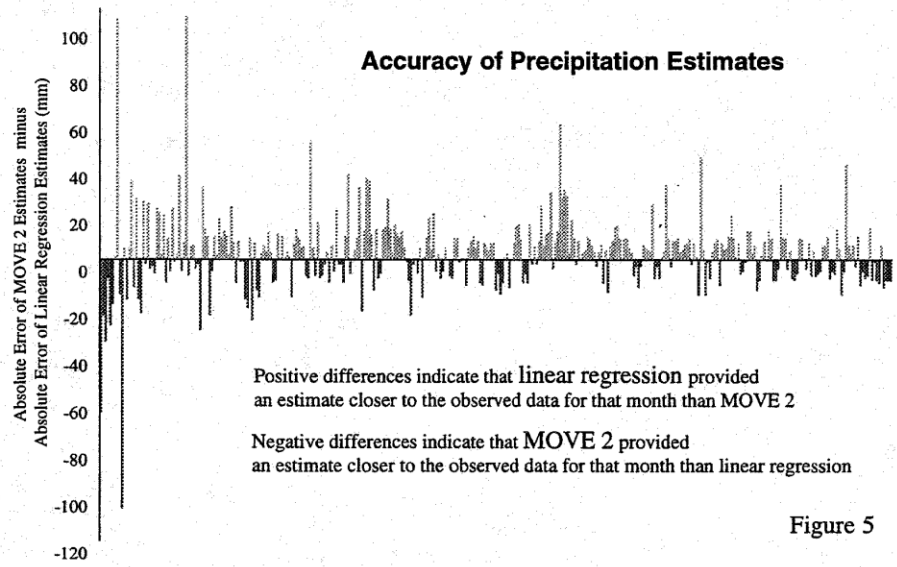


Figure 5

We check that the observed precipitation data within a season is normally distributed before making any estimates of missing data or extending the record. Occasionally, observed monthly data within the season are positively-skewed. If data are skewed, we have found that cube root transformations tend to be more normally distributed.

We believe that organic regression, rather than linear regression, should be used to extend a data record over a long period to maintain the natural variability of the precipitation. Estimates of missing data using organic regression reflect both the mean and variance of the observed data, whereas estimates using linear regression reflect the mean of the observed data only - their variance is less than that of the observed data.

Although organic regression is preferred for extending data records, linear regression also has its place in making estimates of missing data. Linear regression will provide estimates that are closer to individual observed data more frequently than estimates made using organic regression. Therefore, if you only need to make a few estimates of data that are near the mean, linear regression might be preferred. We try to keep in mind the purpose for making the estimates. For developing water supply forecast equations, we prefer organic regression because it ensures that the estimated data mimics the actual variability of data that is received in real time.

REFERENCES

Alley, W.M. and A.W. Burns. (1983). "Mixed-station extension of monthly streamflow records." *J. of Hydraulic Engineering*, 109(10), 1272-1284.

Helsel, D.R. and R.M. Hirsch (1992). Statistical Methods in Water Resources. Elsevier Science Publishers B.V. New York.

Hirsch, R.M. (1982). "A comparison of four streamflow record extension techniques." *Water Resources Research*, 18(4), 1081-1088.

Smith, Stephanie and Eric Weiss (1996). "Application of principal components regression to water supply forecasting." Proceedings of the 64th Western Snow Conference, Bend OR.