Multisite stochastic weather generation for the San Juan River Basin

Nina Caraway
CVEN 6833: Advanced Data Analysis

December 19, 2011

1 Introduction

Generation of weather variables has been a topic of great interest in recent decades. Referred to as "weather generators", these models typically use precipitation as a driving variable and can be extended to simulating other weather variables such as temperature or solar radiation. Traditionally these models are parametric and precipitation generation relies on Markov chain or Poisson process framework (Katz, 1977; Foufoula-Georgiou and Georgakakos, 1991; Woolhiser, 1992). Fitted probability density functions (Gamma, Log-Normal, etc) then produce precipitation amounts, where models such as AR-1 can be implemented for temperature generation (Richardson, 1981). However, extending these models to multiple sites proved to not be a trivial task. One such attempt (Wilks, 1998) involved creating a mixed exponential distribution for precipitation with transformed spatially correlated normal variables. More recent work (Kleiber et al., 2011) utilizes a latent Gaussian process to drive occurrence and a transformed Gaussian process to estimate intensity, with locally estimated parameters interpolated across a grid. Nonparametric weather generators are an attractive alternative. Being data-driven, they can capture deviations from theoretical probability distributions, as well as nonlinearities between variables. Past methods include kernel density estimators (Rajagopalan et al., 1997) and resampling (Lall and Sharma, 1996). The model used in this work is based on (Apipattanavis et al., 2007), where precipitation occurrence is modeled with a two-state Markov Chain and weather variables are selected using k-nearest neighbor (k-NN) resampling algorithm.

This k-NN weather generator will be used to generate daily weather variables for the purpose of more robust streamflow forecasting. Currently the Colorado Basin River Forecasting Center (CBRFC) and the Natural Resource Conservation Service (NRCS) work together to predict streamflows at two time scales. At the seasonal time scale, CBRFC implements Statistical Water Supply (SWS), a regression-based method that relates observed data with future streamflow, and Ensemble Streamflow Prediction (ESP). ESP is based on historical daily weather sequences and a physically based watershed model (such as the Sacramento Soil Moisture model, SAC-SMA). NRCS uses a principle components regression technique. Forecasts from these models are qualitatively combined to issue a single "coordinated" forecast for the U.S. Bureau of Reclamation (USBR) and others. The ESP methodology has shortcomings in that the ensembles created are hindered by limited historical data, which becomes even more limited with the addition of climatological forecasts. With this weather generator, an infinite variety of weather scenarios can be generated, thus allowing for improved probabilistic forecasts.
However, k-NN resampling can only be performed on a single timeseries, which proves troublesome for multisite generation. Previous multisite work included areas with spatial variability, but not necessarily complex terrain (Buishand and Brandsma; Apipattanavis et al., 2007; Kleiber et al., 2011). The Colorado River Basin includes terrain well over 12,000 ft and ranges from desert to forest. The San Juan River Basin was chosen for testing as it contains such diversity within its own boundaries. The objective then became finding a methodology for weather generation that would capture spatial dependencies and correlations. K-means cluster analysis was then chosen as a technique that would organize available weather time series into correlated groupings. Clustering was performed on seasonal precipitation totals, where interestingly elevation was identified as a defining feature as well. Each cluster was combined by an elevation-weighted average to produce a synthetic time series for the k-NN weather generator. The results from this schema were then compared to those from a standard domain-aggregate method.

2 Data

The San Juan River Basin has a drainage area roughly the size of West Virginia and is divided into 24 sub-basins. Each sub-basin is divided into zones – either lower and upper, or lower, middle and upper. Each zone has historical observations of daily precipitation and temperature. Overall, this watershed has sixty-six timeseries available. Performing sixty-six simulations runs the risk of losing spatial correlations and is a time-intensive process. Previous works involved domain aggregation, which is essentially a spatial average. With this technique, however, one runs the risk of washing out individual site characteristics. Discussed in the methodology, k-means clustering finds a middle ground between the two methods.
Specify $N$ simulations, $D$ days, & whether conditional or not

Load weather states $x$

Fit Markov chain based on wet or dry occurrences

Montecarlo determines transitions, $S$, (dd, dw, ww, wd) for $D$ days in each simulation

$n = 1$

$d = 1$

$n = n + 1$

$d = d + 1$

Select all 7 day windows in history centered on current day, $d$

K-Nearest Neighbor: simulating from conditional PDF $f(x_d | x_{d-1}, S_d, S_{d-1})$

Sample historical day, $t$, using weight function $K[j(i)] = \frac{1/j}{\sum_{j=1}^{1/j}}$

Set $t+1$ weather states as simulated values

$D$ days?

yes

no

$N$ simulations?

yes

Save weather scenarios

Stop

no

Figure 2: Weather generator flowchart
3 Methodology

Steps of the k-NN weather generator are outlined in Figure 2. Apipattanavis et al. (2007) outlines the process in more explicit detail. The main difference (aside from conditional simulation, which is not discussed in this paper), is that here a two-state, instead of three-state, Markov chain is implemented. Instead of dry, wet, and extreme wet, it is only dry and wet. This wet/dry chain is fitted for each month to capture seasonality, then the k-NN resampling is conditioned upon a simulated precipitation state.

The cluster analysis chosen was an iterative refinement technique known as the k-means algorithm, or Lloyd’s hill climbing algorithm. The algorithm alternates between an assignment step and an update step given an initial set of $k$ means $\mu_1, \ldots, \mu_k$. Each observation is assigned to its closest centroid and the means are recalculated. The algorithm is finished when the two steps converge (MacKay, 2003). This process minimizes the within-cluster scatter. As it is a machine learning algorithm, different centroid values may be found with each iteration of $k$.

Streamflow forecasts are generally issued on a seasonal scale. Given precipitation as the driving variable for weather generation, seasonal precipitation totals were chosen for the k-means clustering. Seasons were defined as DJF, MAM, JJA, and SON. Figure 3 shows the diagnostics involved in picking $k$ clusters. Selection is based on a $k$ where within sum of squares does not significantly change with each increase in cluster size. For all four seasons, three clusters were determined as an appropriate selection. However, five or six clusters would have had to been chosen for JJA to match the within sum of squares for three clusters in DJF. This most likely occurs because of the more sporadic nature of summer convective precipitation versus the widespread nature of winter frontal precipitation. For ease of analysis, a consistent $k$ was chosen for all seasons. Each weather timeseries is then assigned to its assigned cluster. Each cluster is combined using an elevation-weighted average to produce a synthetic timeseries, which is then fed into the weather generator.

![Clustering of DJF Seasonal Precipitation with 50 reps at each K](image1)

![Clustering of JJA Seasonal Precipitation with 50 reps at each K](image2)

(a) Winter precipitation clustering  
(b) Summer precipitation clustering

Figure 3: Seasonal normalized within sum of squares for each kth cluster
4 Results

Boxplots in Figure 4 depict elevations of weather stations organized by clustering. Though clustering was performed on precipitation, this shows there is a heavy correlation with elevation as well. CBRFC designated zones are included as well to emphasize that classification of zones vary depending on the sub-basin. Assignment of 1, 2, or 3 is arbitrary and will change with each analysis due to the heuristic nature of the algorithm. Winter is more organized by elevation while summer most likely has more spatial correlations. The latter would be better depicted with a spatial map to complement the boxplots. Figures 5 and 6 compare pairwise correlations of precipitation intensities of a domain aggregated method with clustered pairwise correlations in Figures 7 and 8. With the domain aggregated method, weather is simulated simultaneously across the watershed. The idea behind clustering is to create distinct homogeneous regions, within which then weather is then simultaneously simulated. Figures 5 and 6 are actually quite skillful in capturing the range of correlations from low to high. With clustering, the large-scale forcing in creating similar behavior is diminished. To clarify, it may rain on one day in one cluster, and may not in another.

![Figure 4: Elevations of clusters with CBRFC zone designations overlaid.](image)

Figures 9 to 17 compare single site weather from the domain aggregate method to that from an elevation average from clustering. Through visual inspection, it is hard to determine if any improvement occurred for winter precipitation. However, daily maximum temperature experienced improvement, especially in the averages. For summer precipitation, July and August had an appreciable difference in capturing historical statistics. Conversely, maximum temperature appeared nearly-identical.
Figure 5: Domain aggregated winter pairwise correlations between stations for precipitation intensity

Figure 6: Domain aggregated summer pairwise correlations between stations for precipitation intensity

Figure 7: Within cluster winter pairwise correlations between stations for precipitation intensity

Figure 8: Within cluster summer pairwise correlations between stations for precipitation intensity
Figure 9: Upper Durango winter precipitation from domain aggregation

Figure 10: Upper Durango maximum temperature from domain aggregation

Figure 11: Upper Durango winter precipitation from clustering

Figure 12: Upper Durango maximum temperature from clustering
Figure 13: Upper Durango summer precipitation from domain aggregation

Figure 14: Upper Durango maximum temperature from domain aggregation

Figure 15: Upper Durango summer precipitation from clustering

Figure 16: Upper Durango maximum temperature from clustering
5 Conclusions

Overall, k-means clustering was tested and implemented as an alternative to simple domain aggregation for weather generation. Though precipitation correlations are well captured for a range of values in the simpler method, this proposed method strives to minimize forcing interactions between sites with low correlations in the first place. Initial analysis of distributional statistics on monthly values did not show a significant improvement in capturing historical values. However, this was a simple comparison and more robust methods can be implemented to determine skill. Also only one site was selected for comparison. Sites from other regions of the basin will be assessed as well, and the remaining two seasons not discussed will be scrutinized as well. Future analysis will involve a means of assessing basin-wide skill. Future work will also include testing for extreme behavior. Weather generators are notorious for under-simulating wet extremes, such as spells, so this new methodology will investigate possibilities of improvement. Furthermore, statistics will be performed on flow values after the weather ensembles are driven through a physical hydrologic model.

References


Buishand, T., and T. Brandsma ( ), Multisit simulation of daily precipitation and temperature in the rhine basin by nearest-neighbor resampling), pages = 2761–2776, volume = 37(11), year = 2001,, Water Resources Research.


