Description of Kirsch-Nowak Streamflow Generator

Julianne Quinn, Matteo Giuliani and Jon Herman

August 23, 2017

1 Synthetic Streamflow Generation

This repository contains code for generating correlated synthetic daily streamflow time series at multiple sites assuming stationary hydrology. Monthly flows are generated using Cholesky decomposition (see *Kirsch et al.* (2013)) and then disaggregated to daily flows by proportionally scaling daily flows from a randomly selected historical month +/- 7 days as in *Nowak et al.* (2010). The monthly generation is described in detail in Section 1.1, while the daily generation is described in Section 1.2. Statistical validation using an example dataset from the Susquehanna River Basin is then provided in Section 2 (for a description of the system, see *Giuliani et al.* (2014)).

1.1 Monthly Streamflow Generation

For a given site, we denote the set of historical streamflows as $\mathbf{Q_H} \in \mathbb{R}^{N_H \times T}$ and the set of synthetic streamflows as $\mathbf{Q_S} \in \mathbb{R}^{N_S \times T}$, where N_H and N_S are the number of years in the historical and synthetic records, respectively, and T is the number of time steps per year. Here T=12 for 12 months. For the synthetic generation, the streamflows in $\mathbf{Q_H}$ are log-transformed to yield the matrix $Y_{H_{i,j}} = \ln(Q_{H_{i,j}})$, where i and j are the year and month of the historical record, respectively. The streamflows in $\mathbf{Y_H}$ are then standardized to form the matrix $\mathbf{Z_H} \in \mathbb{R}^{N_H \times T}$ according to equation 1:

$$Z_{H_{i,j}} = \frac{Y_{H_{i,j}} - \hat{\mu}_j}{\hat{\sigma}_j} \tag{1}$$

where $\hat{\mu_j}$ and $\hat{\sigma_j}$ are the sample mean and sample standard deviation of the j-th month's log-transformed streamflows, respectively. These variables follow a standard normal distribution: $Z_{H_{i,j}} \sim \mathcal{N}(0,1)$.

For each site, we generate standard normal synthetic streamflows that reproduce the statistics of $\mathbf{Z_H}$ by first creating a matrix $\mathbf{C} \in \mathbb{R}^{N_S \times T}$ of randomly sampled standard normal streamflows from $\mathbf{Z_H}$. This is done by formulating a random matrix $\mathbf{M} \in \mathbb{R}^{N_S \times T}$ whose elements are independently sampled integers from $(1, 2, ..., N_H)$. Each element of $\mathbf{C}$ is then assigned the value $C_{i,j} = Z_{H_{(M_{i,j}),j}}$, i.e. the elements in each column of $\mathbf{C}$ are randomly sampled standard normal streamflows from the same column (month) of $\mathbf{Z_H}$. In order to preserve the historical cross-site correlation, the same matrix $\mathbf{M}$ is used to generate $\mathbf{C}$ for each site.

Because of the random sampling used to populate $\mathbf{C}$, an additional step is needed to generate auto-correlated standard normal synthetic streamflows, $\mathbf{Z_S}$. Denoting the historical autocorrelation $\mathbf{P_H}$=corr($\mathbf{Z_H}$), where corr($\mathbf{Z_H}$) is the historical correlation between standardized streamflows in months i and j (columns of $\mathbf{Z_H}$), an upper right triangular matrix, $\mathbf{U}$, can be found using Cholesky decomposition such that $\mathbf{P_H} = \mathbf{U}^\intercal\mathbf{U}$. $\mathbf{Z_S}$ is then generated as $\mathbf{Z_S} = \mathbf{CU}$. Finally, for each site, the auto-correlated synthetic standard normal streamflows $\mathbf{Z_S}$ are converted back to log-space streamflows $\mathbf{Y_S}$ according to $Y_{S_{i,j}} = \hat{\mu_j} + Z_{S_{i,j}}\hat{\sigma_j}$. These are then transformed back to real-space streamflows $\mathbf{Q_S}$ according to $Q_{S_{i,j}} = \exp(Y_{S_{i,j}})$.

While this method reproduces the within-year log-space autocorrelation, it does not preserve year to-year correlation, i.e. concatenating rows of $\mathbf{Q_S}$ to yield a vector of length $N_S \times T$ will yield discontinuities in the autocorrelation from month 12 of one year to month 1 of the next. To resolve this issue, *Kirsch et al.* (2013) repeat the method described above with a historical matrix $\mathbf{Q'_H} \in \mathbb{R}^{N_{H-1} \times T}$, where each row i of $\mathbf{Q'_H}$ contains historical data from month 7 of year i to month 6 of year *i+1*, removing the first and last 6 months of streamflows from the historical record. $\mathbf{U}'$ is then generated from $\mathbf{Q'_H}$ in the same way as $\mathbf{U}$ is generated from $\mathbf{Q_H}$, while $\mathbf{C}'$ is generated from $\mathbf{C}$ in the same was as $\mathbf{Q'_H}$ is generated from $\mathbf{Q_H}$. As before, $\mathbf{Z'_S}$ is then calculated as $\mathbf{Z'_S} = \mathbf{C'U'}$. Concatenating the last 6 columns of $\mathbf{Z'_S}$ (months 1-6) beginning from row 1 and the last 6 columns of $\mathbf{Z_S}$ (months 7-12) beginning from row 2 yields a set of synthetic standard normal streamflows that preserve correlation between the last month of the year and the first month of the following year. As before, these are then de-standardized and

back-transformed to real space.

1.2 Daily Streamflow Generation

After generating monthly streamflows as described in Section 1.1, a nearest-neighbor approach described by *Nowak et al.* (2010) is used to disaggregate these streamflows to daily values. The first step in this method is to calculate the k nearest neighbors from the set of historical monthly streamflows for each synthetically-generated month. Nearness is determined by the real-space Euclidean distance, d, across M sites (equation 2):

$$d = \left[\sum_{m=1}^{M} \left((q_S)_m - (q_H)_m \right)^2 \right]^{1/2} \tag{2}$$

where $(q_S)_m$ is the real-space synthetic monthly hydrologic variable generated at site m and $(q_H)_m$ is the real-space historical monthly hydrologic variable at site m. For each synthetically-generated hydrologic variable in month j, d is calculated with respect to the all historical values of the streamflows in month j +/- 7 days. That is, rather than only considering historical January flows as neighbors to the synthetic January flows, for example, total flows over 31 consecutive days within the period from the last week of December to the first week of February are considered. The k-nearest neighbors are then sorted from i=1 for the closest to $i = k$ for the furthest, and probabilistically selected for proportionally scaling streamflows in disaggregation. We use the Kernel estimator given by *Lall and Sharma* (1996) to assign the probability p_n of selecting neighbor n (equation 3):

$$p_n = \frac{\frac{1}{n}}{\sum_{i=1}^{k} \frac{1}{i}} \tag{3}$$

Following *Lall and Sharma* (1996) and *Nowak et al.* (2010), we use $k = \left\lfloor N_H^{1/2} \right\rfloor$. After a neighbor is selected, the final step in disaggregation is to proportionally scale all of the historical daily streamflows at site m from the selected neighbor so that they sum to the synthetically generated monthly total at site m. For example, if the first day of the month of the selected historical neighbor represented 5% of that month's historical flow, the first day of the month of the synthetic series would represent 5% of that month's synthetically-generated flow.

2 Verification of Synthetic Hydrologic Statistics

This section provides a statistical validation of the synthetic generator using data from the Susquehanna River Basin. The Kirsch-Nowak generator was used in this system to generate synthetic streamflows at the Marietta gauging station upstream of Conowingo Dam (USGS station 01576000), as well as inflows to Muddy Run Reservoir, lateral inflows between Marietta and the Conowingo Dam, and evaporation rates over the Conowingo and Muddy Run dams simulated from an OASIS system model. Since the evaporation rates at the two dams from the OASIS model were identical, only one set was included in the stochastic generation. Additionally, since the evaporation rates are more normally distributed than log-normally distributed like the streamflows, they were first transformed with an exponential transformation before applying the monthly generation method described in Section 1.1. They were then back-transformed with a log-transformation before applying the disaggregation procedure described in Section 1.2.

As stated in Section 1, the goal of the synthetic generator is to produce a time series of synthetic hydrologic variables that expand upon those in the historical record while reproducing their statistics. Historical and synthetic probability of exceedance curves of daily hydrologic variables in the Susquehanna River Basin (Figure 1) indicate that the first is true, as the synthetic hydrologic variables include more extreme high and low values. The synthetic hydrologic variables also appear unbiased, as this expansion is relatively equal in both directions. Finally, the synthetic probability of exceedance curves also follow the same shape as the historical, indicating that they reproduce the within-year distribution of daily values. Probability of exceedance curves in Figure 1 were generated from 1000 years of synthetic hydrologic variables.

To more formally confirm that the synthetic hydrologic variables are unbiased and follow the same distribution as the historical, we test whether or not the synthetic median and variance of real-space monthly values are statistically different from the historical. The results of these tests are shown in Figure 2 for Marietta, which provides most of the system flow. This figure was generated from a 100-member ensemble of synthetic series of length 100 years, and a bootstrapped ensemble of historical years of the same size and length. Panel a shows boxplots of the real-space historical and synthetic

monthly flows, while panels b and c show boxplots of their means and standard deviations, respectively. Because the real-space flows are not normally distributed, the non-parametric Wilcoxon rank-sum test and Levene's test were used to test whether or not the synthetic monthly medians and variances were statistically different from the historical. The p-values associated with these tests are shown in Figures 2d and 2e, respectively. None of the synthetic medians or variances are statistically different from the historical at a significance level of 0.05.

In addition to verifying that the synthetic generator reproduces the first two moments of the historical monthly hydrologic variables, we also verify that it reproduces both the historical autocorrelation and cross-site correlation at monthly and daily time steps. The results of this analysis are shown in Figures 3 and 4. Figures 3a and 3b show the autocorrelation function of historical and synthetic real-space flows at Marietta for up to 12 lags of monthly flows (panel a) and 30 lags of daily flows (panel b). Also shown are 95% confidence intervals on the historical autocorrelations at each lag. The range of autocorrelations generated by the synthetic series expands upon that observed in the historical while remaining within the 95% confidence intervals for all months, suggesting that the historical monthly autocorrelation is well-preserved. On a daily time step, most simulated autocorrelations fall within the 95% confidence intervals for lags up to 10 days, and those falling outside do not represent significant biases.

Figures 4a and 4b show boxplots of the cross-site correlation in monthly (panel a) and daily (panel b) real-space hydrologic variables for all pairwise combinations of sites. The synthetic generator greatly expands upon the range of cross-site correlations observed in the historical record, both above and below. Table 1 lists which sites are included in each numbered pair of Figure 4. Wilcoxon rank sum tests (panels c and d) for differences in median monthly and daily correlations indicate that pairwise correlations are statistically different ($\alpha = 0.5$) between the synthetic and historical series at a monthly time step for site pairs 1, 2, 5 and 6, and at a daily time step for site pairs 1 and 2. However, biases for these site pairs appear small in panels a and b. In summary, Figures 1-4 indicate that the streamflow generator is reasonably reproducing historical statistics, while also expanding on the observed record.

References

Giuliani, M., J. Herman, A. Castelletti, and P. Reed (2014), Many-objective reservoir policy identification and refinement to reduce policy inertia and myopia in water management, *Water Resources Research*, *50*(4), 3355–3377.

Kirsch, B. R., G. W. Characklis, and H. B. Zeff (2013), Evaluating the impact of alternative hydro-climate scenarios on transfer agreements: A practical improvement for generating synthetic streamflows, *Journal of Water Resources Planning and Management*, *139*(4), 396–406, doi: 10.1061/(ASCE)WR.1943-5452.0000287.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resources Research*, *32*(3), 679–693.

Nowak, K., J. Prairie, B. Rajagopalan, and U. Lall (2010), A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow, *Water Resources Research*, *46*, W08529, doi:10.1029/2009WR008530.

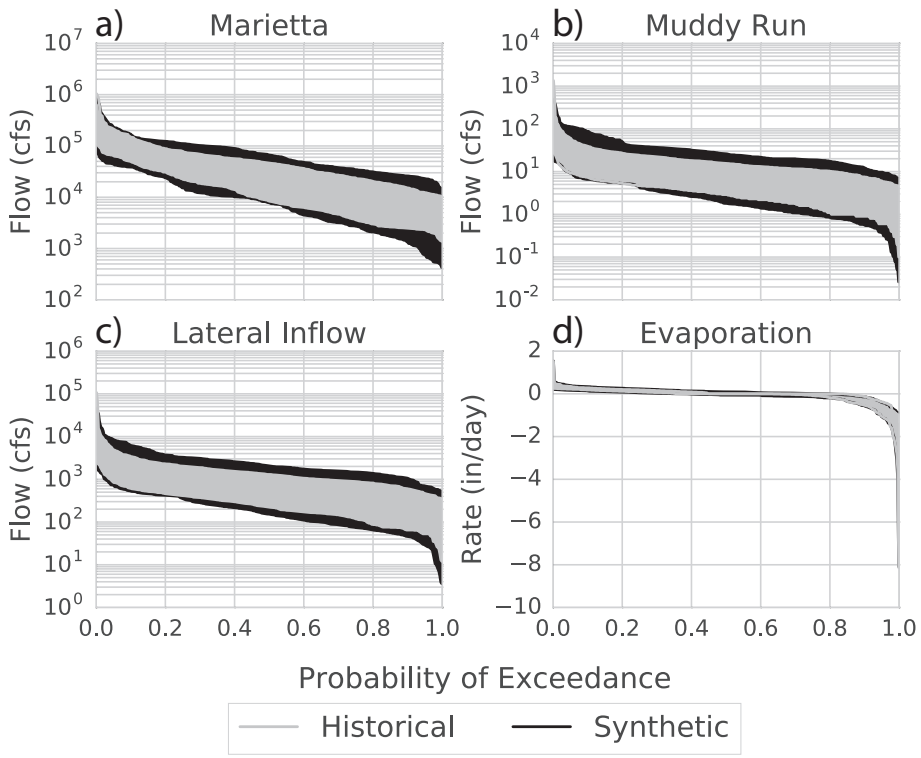

Figure 1: Probability of exceedance curves of the historical (gray) and synthetic (black) streamflows in the Lower Susquehanna River Basin. The synthetic streamflows increase the range of values over which the reservoir operating policies are optimized.

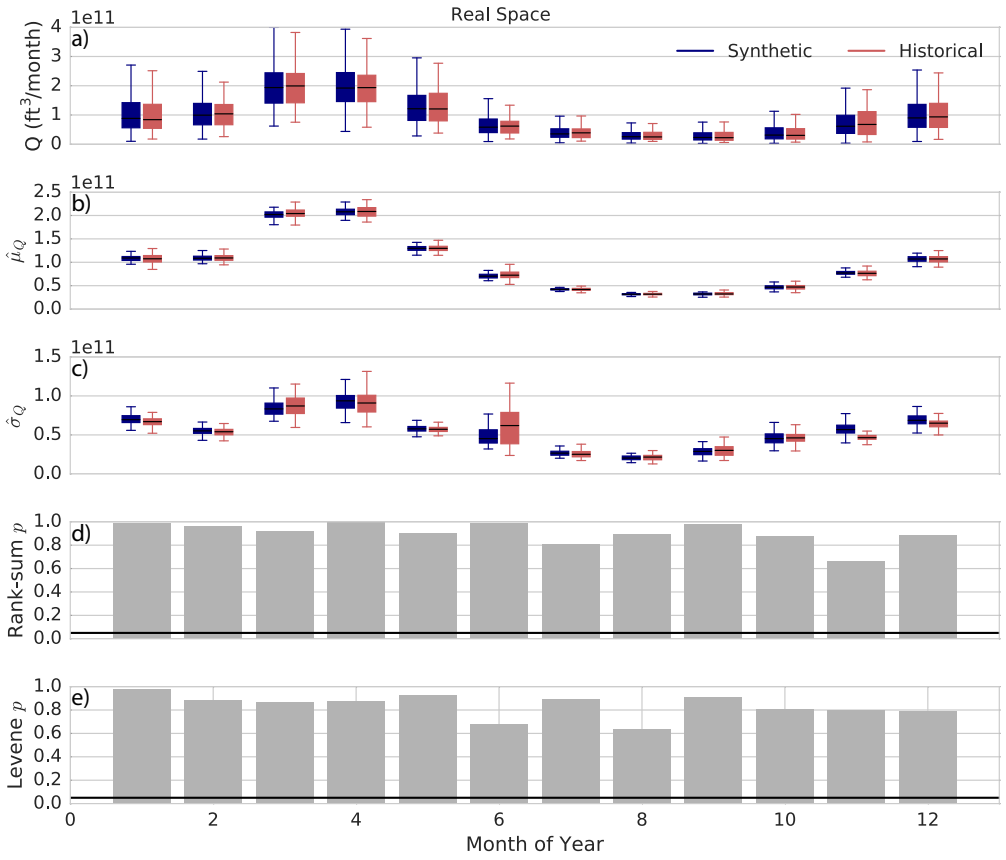

Figure 2: Boxplots of the historical (pink) and synthetic (blue) total monthly flows (panel a), mean monthly flows (panel b) and standard deviation of monthly flows (panel c) as well as p-values for differences in median (panel d) and variance (panel e) of monthly flows at Marietta. In the boxplots, a black line is drawn at the median, while the box edges extend to the quartiles and the whiskers to 1.5 times the interquartile range beyond the quartiles. p-values for differences in median were determined by a rank sum test, while those for differences in variance were determined by Levene's test.

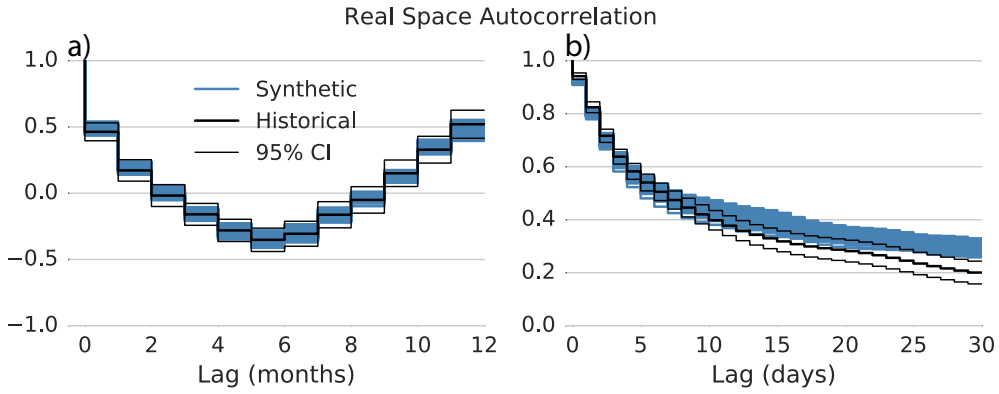

Figure 3: Historical (black) and synthetic (blue) monthly (panel a) and daily (panel b) autocorrelation functions for streamflow time series at Marietta. Black lines in panels a and b show both the mean and 95% confidence interval bounds on the historical autocorrelation.

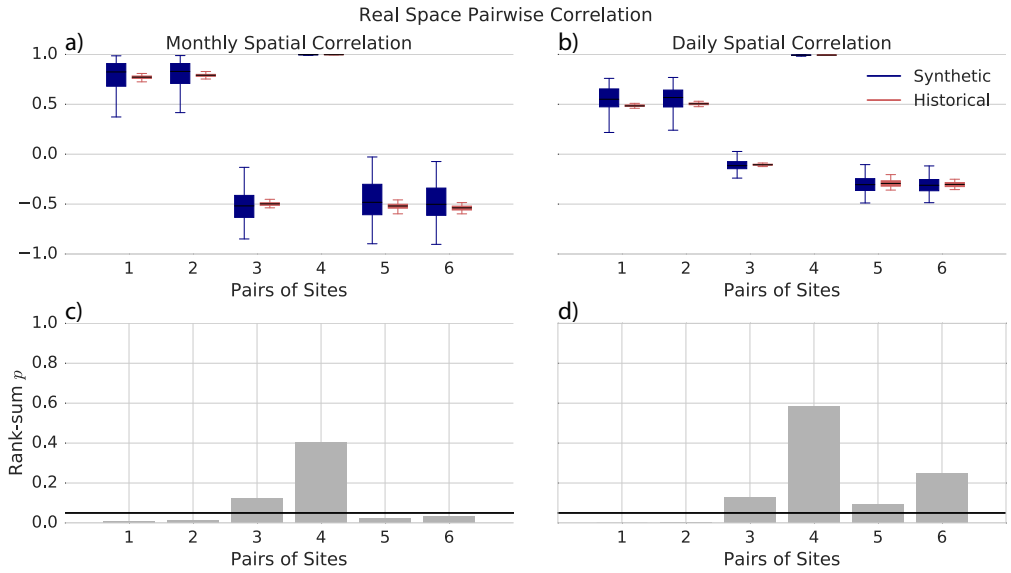

Figure 4: Boxplots of pairwise cross-correlations in monthly (panel a) and daily (panel b) historical (pink) and synthetic (blue) streamflows between sites as well as p-values for differences in median (panel c) and variance (panel d). Site pairs are listed in Table 1. In the boxplots, a black line is drawn at the median, while the box edges extend to the quartiles and the whiskers to 1.5 times the interquartile range beyond the quartiles.

Table 1

Pair Number	Sites
1	Marietta and Muddy Run
2	Marietta and Lateral Inflows
3	Marietta and Evaporation
4	Muddy Run and Lateral Inflows
5	Muddy Run and Evaporation
6	Lateral Inflows and Evaporation