Stationary and Non-Stationary Flood Frequency Analysis on the North Fork Virgin River Using GEV Distributions and Hidden Markov Models with GEV States.

Peter Van Dusen
CVEN 6833-01
2017 Project

This project investigated four flood frequency analysis models for annual peak mean daily flows on the North Fork Virgin River in Zion National Park, UT. Two of the methods, stationary and non-stationary GEV distributions, are traditional methods that have been explored through other research (Coles S. 2001, Katz 2010, Condon et al. 2014). *This project found that a non-stationary GEV distribution with the location and scale changing as a function of mean daily precipitation offered improvements over the stationary GEV model. It also was apparent that the median annual exceedance probabilities for high flows for the non-stationary model was lower than that for the stationary GEV distribution. The two other methods, stationary and non-stationary hidden Markov models with states as GEV distributions, have been less explored in literature. This project found that both a stationary and non-stationary two state HMM-GEV with only the location changing between states represented the historical data fairly well. It appeared the models might have explained possible historical shifts in higher median peak annual flows. These HMM-GEV models, however, need to be investigated further over a wider variety of sites and with more than two states.*

Introduction

Statistical methods used to determine high flow magnitudes and frequencies are crucial in hydrologic design. Flood analysis can give valuable insight into structure design requirements, safety risks posed to the public, and other potential impacts of probable floods (Chow and Maidment 1998). In the famous water resource paper "Stationarity Is Dead: Whither Water Management", it is argued that stationary methods and distributions are no longer appropriate in flood frequency analysis – the magnitude and return periods of flows will be changing as the climate factors that drive the hydrologic cycle change (Milly et al. 2008). The purpose of this project is to investigate a traditional and a non-traditional flood frequency model using stationary and non-stationary methods to improve flood forecast planning for the North Fork Virgin R. in Zion National Park, Utah.

A variety of stationary and non-stationary flood analysis techniques have been widely used and researched. The United States Federal government recommends the use of a three parameter log-Person Type III (LP3) distribution on above threshold flows or peak annual flows with a regional skew coefficient and weighting (Griffis and Stedinger 2007). This stationary technique has been the guideline in the United States since 1982. This federal code suggests that other methods can be used in flood analysis, as long as the process is justified and compared with the federal recommendations (IACWD 1982). Numerous other distributions have been studied (Millington et al. 2011). Vogel et al. (2008) investigated the use of a variety of distributions for flood frequency analysis in the US; it was found that the three parameter generalized extreme value distribution (GEV) and the two and three parameter log normal curves (LN2 and LN3) were comparable to the federally recommended LP3. Karim and Chowdhury (1995) found that the three parameter GEV was a more adequate model for flood frequency analysis in Bangladesh compared to LN2 and LP3 distributions. Recent research has demonstrated the improvements offered through a non-stationary GEV distribution, or a GEV distribution where the parameters are a function of time or covariate climate indices (Coles S. 2001, Katz 2010, Condon et al. 2014).

While used frequently in hydrology, a statistical method that has largely not been investigated in the field of hydrologic extremes is the use of hidden Markov models (HMM). In these models, it is assumed that a time series is described by an underlying Markov chain with states as probability distribution functions. The probability of being in a state can be a result of a random probability transition matrix (as in a traditional Markov chain) or from a logistic model that is a function of covariates (non-stationary) (Bracken et al. 2014). A variety of distributions and applications have been demonstrated for these hidden Markov Models (Bracken et al. 2014, Zucchini and Guttorp 1991, Evin et al. 1991). While Sun et al. (2013) and Pender et al. (2015) have investigated the use of hidden Markov models with GEV states, little research has been conducted to investigate the potential applications of non-stationary HMM-GEV to flood frequency analysis.

The purpose of this project is to perform annual maximum flood frequency analysis on the N. Fork Virgin River (USGS 09405500) using four methods:

1. Stationary GEV
2. Non-stationary GEV
3. Stationary HMM-GEV
4. Non-stationary HMM-GEV

Data

This analysis will be performed on annual peak mean daily flows for the North Fork Virgin River in Zion National Park, Utah (USGS gauge site 09405500). Flow values were obtained from the USGS website. Mean daily flows, expressed in cubic feet per second, were available on a daily timescale from 1928 to 2016 (USGS 2016).

Historical 'observed' covariate data was downscaled Livneh data provided by the Bureau of Reclamation. Livneh data is spatially gridded climate data that has been derived from approximately 20,000 NOAA observatory sites across the US. It is gridded at a spatial resolution of 1/16 degree latitude/longitude (Livneh et al. 2013). For this study, mean daily precipitation (expressed in millimeter), max temperature, and minimum temperature (both expressed in Celsius) were used as observed covariates. The data was available on a daily time scale from 1950 to 2005.

Future covariate data was downscaled CMIP5 data provided by the Bureau of Reclamation. This CMIP5 data set contained an ensemble of data from 64 GCM models that use different factors and scenarios to model variables (Taylor et al. 2011). For this study, mean daily precipitation (expressed in millimeters), max temperature, and minimum temperature (both expresses in Celsius) were used as forecasted covariates. The data was available on a daily time scale from 1950 to 2099.

Method

Generalized Extreme Value Distributions

For stationary and non-stationary GEV models, annual peak flow data was fit to the GEV distribution (eq 1) below using Eric Gilleland's 2016 R package 'extRemes'. μ is the distribution location, σ is the distribution scale, and ε the distribution shape. Parameters were fit using the method of maximum likelihood.

$$G(z) = 1 - exp\left\{-\left[1 - \varepsilon\frac{(z-\mu)}{\sigma}\right]^{\frac{-1}{\varepsilon}}\right\} \text{ .... eq 1}$$

For the nonstationary GEV model, the location and scale were linear functions of the covariates (***x***) (eq 2 and eq 3). The shape was kept constant to maintain physically realistic distributions (eq 4). It is important to note that the logarithm of the scale was modeled in order to maintain a positive scale factor. Beta values were solved using the method maximum likelihood.

$$\mu(t) = \beta_{0,\mu} + \beta_{1,\mu}x_1 + \cdots + \beta_{n,\mu}x_n \text{ .... eq 2}$$

$$\log(\sigma(t)) = \beta_{0,\sigma} + \beta_{1,\sigma}x_1 + \cdots + \beta_{n,\sigma}x_n \text{ .... eq 3}$$

$$\varepsilon = \beta_{0,\varepsilon} \text{ .... eq 4}$$

The data was fit to over 36 non-stationary GEV models and 1 stationary model, each with a unique combination of covariates used to model either the location, scale, or both the location and scale parameters. Likelihood ratio tests, Akaiki information criteria (AIC), and quantile-quantile plots were used to determine the best model.

After the best model selection process, the model was applied to the CMIP ensemble between 2020 and 2099. With 64 models, 5120 GEV distribution were obtained – one for each year of each trace. A distribution of return flows and exceedance probabilities for specific flows could be obtained from these GEV distributions.

Hidden Markov Model with Generalized Extreme Value Distributions

Stationary and non-stationary HMM-GEV models were fit using the 'extRemes' package and David Harte's 2017 R package 'HiddenMarkov'. Hidden Markov parameters were determined using the Baum-Welch algorithm (Baum et al. 1970). For the non-stationary model, a logistic regression was used to determine the probability of being in a specific state as a function of the covariates. AIC was used as criteria for the best logistic model selection.

The stationary HMM-GEV model was simulated using the 'HiddenMarkov' package. The transition probability matrix and a random number (from a uniform distribution) between 0 and 1 determine the state of the model in the next time step. A sample is then pulled from that GEV distribution. This simulation was repeated several thousand times with a resulting distribution of extreme flows.

Similarly, the non-stationary HMM-GEV was simulated. First, the logistic regression model was used to predict the probability of being in a state at a specific time step using the observed covariates. Then, at each time-step in the simulation, a random number between 0 and 1 (from a uniform distribution) was generated to determine the GEV state at that time-step. A sample was then pulled from that GEV distribution. This was repeated several thousand times for the historical record, and a simulated distribution of annual maximum flows at each year was determined. This process is outlined in more detail in Bracken et al. 2014.

Results and Discussion

Generalized Extreme Value Distributions

The best model selected for annual peak mean daily flows is shown in figure 1. The best model was non-stationary where the location and scale parameters were linear functions of annual average daily precipitation. The shape was kept constant. A likelihood ratio test with the negative log likelihood value of the stationary GEV model indicated strong evidence to reject the stationary model. The AIC of 827.81 compared to the stationary GEV model AIC of 834.89 confirms this.

Figure 2: Best GEV model selection

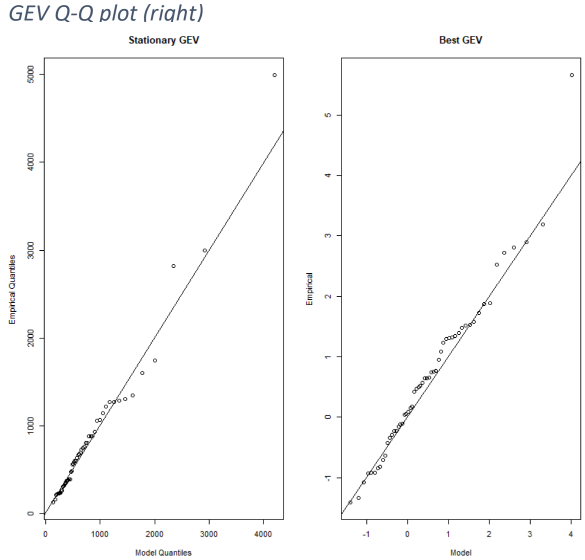
Best Model = fit.gev1.2p

Location		**Scale**				
μ_0	μ_1	ϕ_0	ϕ_1	**Shape**	**Nllh**	**AIC**
230.43	135.08	4.23	0.94	0.38	408.90	827.81

LR Test Against Sationary GEV Model: **p-value=0.003926**

Where: Location $\mu = \mu_0 + (\mu_1 * Pr)$
Scale $log(\varphi) = \varphi_0 + (\varphi_1 * Pr)$
Shape *(Constant)*

Figure 1: Stationary GEV Q-Q plot (left) and Non-Stationary GEV Q-Q plot (right)



A quantile-quantile plot comparing the stationary model and the best nonstationary model are shown in figure 2. This is a clear visual justification for the improvements offered by the non-stationary model. It is interesting to note that both models had a difficult time accounting for the largest extreme flow.

This best model was fit to the historical covariate data. The corresponding annual 1% exceedance flows (red), annual 5% exceedance flows (blue), and observed flows (black) were plotted. Both best model (dashed lines) and stationary model (solid lines) return levels were included for comparison. This is displayed in figure 3 below.

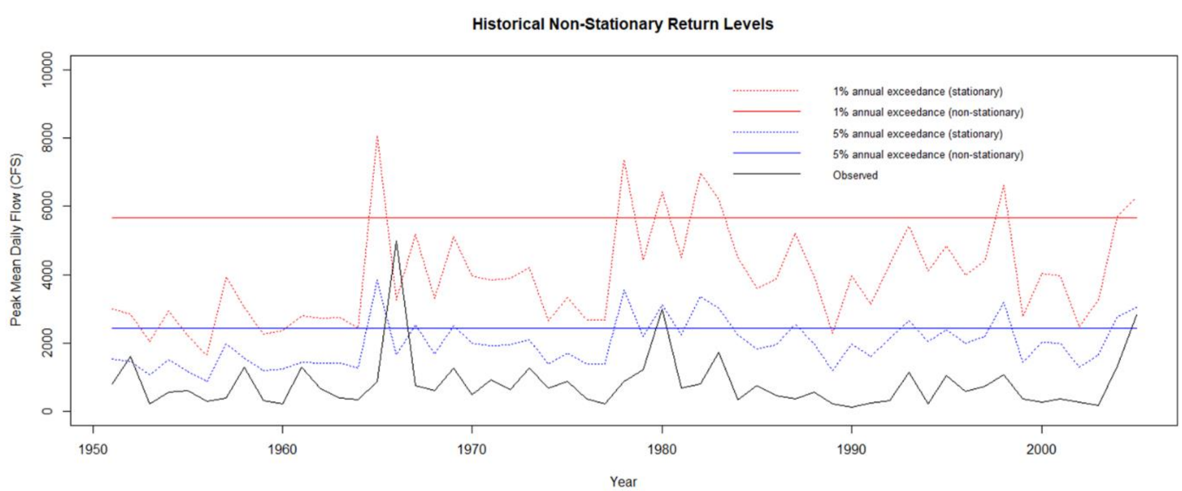


Figure 3 Observed flows (black) with stationary (solid) and nonstationary (dashed) 5% and 1% annual exceedance flows

It is apparent the nonstationary model does seem to follow some of the peaks (particularly when you look at the lower 2 year events) well. When comparing the non-stationary and stationary return levels. The stationary model generally has a higher return level on most of the years – however certain covariate conditions result in an occasional higher nonstationary return level. This best model was then applied to future CMIP data between 2020 and 2099. The distributions for exceedance probabilities of critical flows are shown below in figure 4. The red lines are the stationary exceedance probabilities. It is clear the stationary models have significantly higher exceedance probabilities than the non-stationary GEV model for high flows.

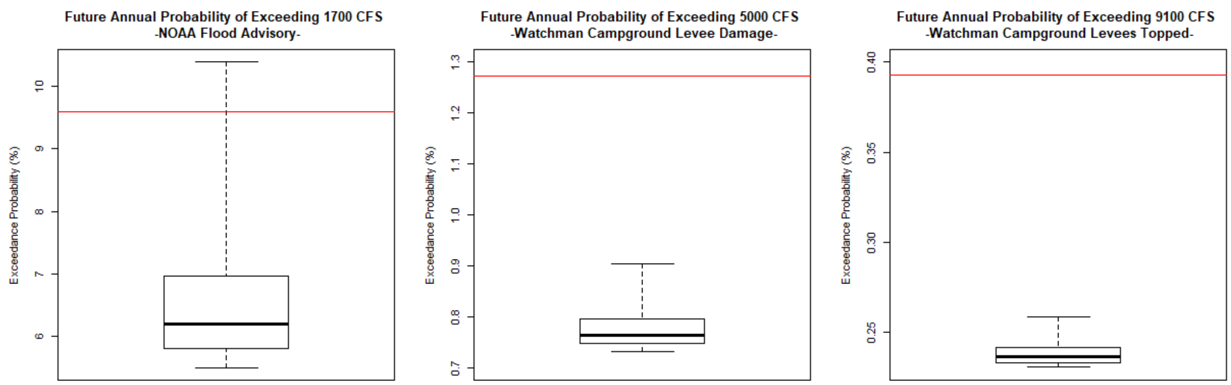


Figure 4: The distribution of exceedance probabilities for critical flows generated from the non-stationary GEV model with CMIP covariates. The stationary exceedance probability is also included (red).

Hidden Markov Model with Generalized Extreme Value Distributions

The stationary model parameters and corresponding transition probability matrix determined by the Baum-Welch algorithm are shown in figure 5. The model has only been successful in determining parameters for a two-state hidden Markov model. It is important to note the two states only differ by their location values. When the scale was also adjusted, the log likelihood decreased from -413.1 to -441.0, indicating a worse model. The two states and their scaled density functions are shown on the left figure 6 below. The graph on the right is depicting the most likely states for the observed data. It is clear that there are some groupings of points more likely to correspond to state 2, where there does appear to be a generally higher shift in the median return level.

Figure 5: HMM GEV state parameters and TPM

$GEV(1)$	$GEV(2)$
$\mu_1 = 392$	$\mu_2 = 767$
$\sigma_1 = 251$	$\sigma_2 = 251$
$\varepsilon_1 = .392$	$\varepsilon_2 = .392$

$$TPM = \begin{bmatrix} .87 & .13 \\ .42 & .48 \end{bmatrix}$$

The stationary HMM-GEV model was simulated with the results shown in figure 7. This model appears to perform well when compared with the observed. The model results in slightly lower 75th and 95th percentiles.

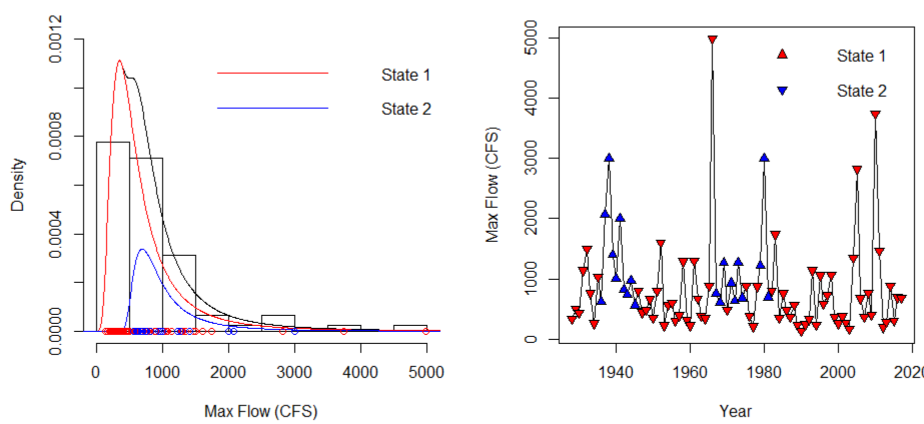


Figure 6: Two state HMM distribution (left). Decoded historical observations (right)

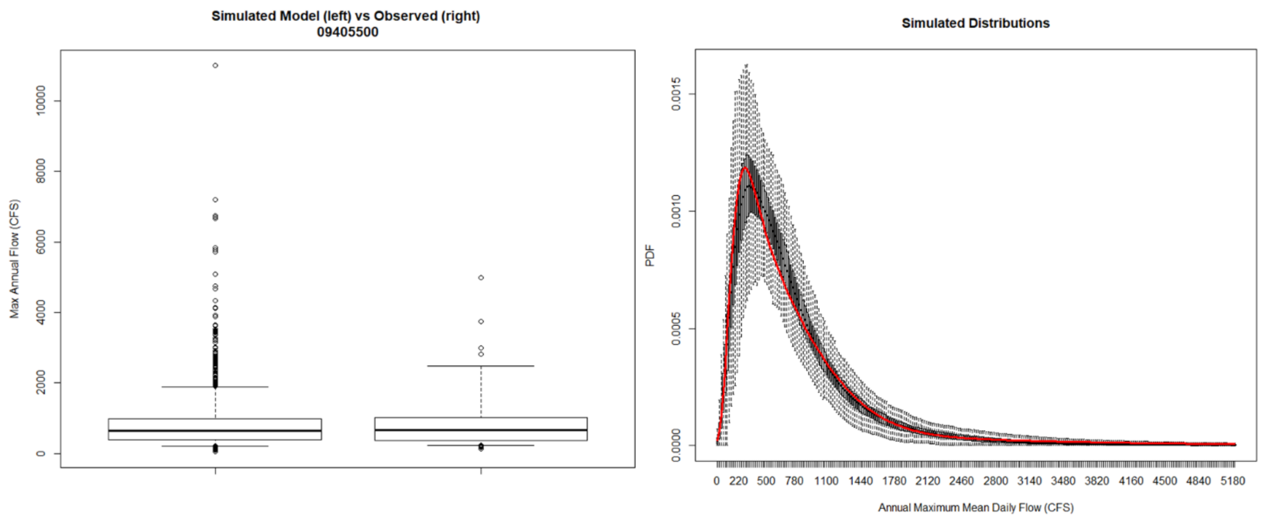


Figure 7: Simulate stationary HMM-GEV results. Left figure compares the simulated (left) with observed (right). The model on the right compares pdfs of simulate vs observed (red).

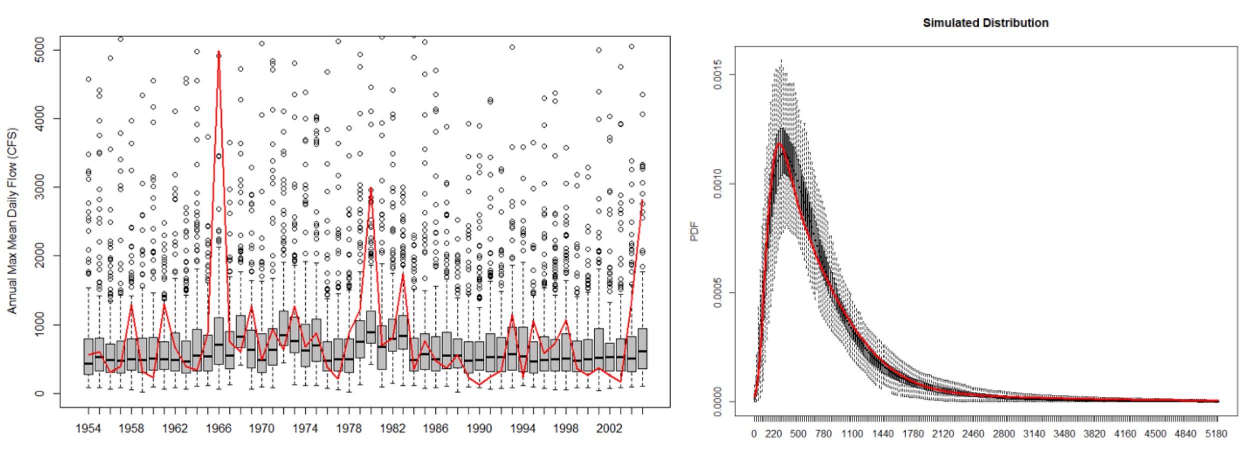
The same states in the stationary HMM-GEV model were used in the non-stationary HMM-GEV model. Using the decoded states of the observed data as the response variable, and the covariate data available for the observed data as predictors, a best logistic model, eq 5, was obtained using the covariates below.

$$logit(p_t) = 16.90 - 1.66(MaxT_t) + 2.55(Pr_{t-1}) + 20.62(m_{t-1}) - 18.91(m_{t-2})$$ ……………………….eq 5

MaxT = annual average max daily temperature (C)
Pr = annual average daily rainfall (mm)
m = Markov GEV state

The simulation results from the non-stationary GEV model is shown below in figure 8. The annual boxplots seem to capture some potential shifts in the median annual peak flow. There are two low periods in the early 1990s and late 1990s that do not seem to be captured as well. The simulated distributions compared to the observed appears to be a reasonable fit.

Figure 8: Annual simulated results compared to observed (red)



Summary

This project investigated four flood frequency analysis models for annual peak mean daily flows on the North Fork Virgin River in Zion National Park, UT. Two of the methods, stationary and non-stationary GEV distributions are traditional methods that have been explored in other research. This project found that a non-stationary GEV distribution with the location and scale changing as a function of mean daily precipitation offered improvements over the stationary GEV model. It also was apparent that the median annual exceedance probabilities for future critical flows is lower than what is predicted with the stationary GEV distributions.

The two other methods, stationary and non-stationary hidden Markov models with states as GEV distributions, have been less explored in literature. This project found that both a stationary and non-stationary two state HMM-GEV with only the location changing between states represented the historical data fairly well. It appeared the models might have explained possible historical shifts in higher median peak annual flows.

Further investigation and work needs to be done for HMM-GEV models. Analysis at other sites should be completed – the North Fork often responds to large short storms that annual climate shifts might not explain well. This project also did not use the HMM-GEV model for the CMIP projections, which would be interesting to investigate. A HMM model with more than two states should also be investigated. Seasonal analysis might also be interesting as shifts in seasonal timing might be capture by a shift in states in the HMM-GEV model.

References

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *Ann. Math. Statist.* 41, no. 1 (February 1970): 164–71. https://doi.org/10.1214/aoms/1177697196.

Bracken, Cameron, B Rajagopalan, and Edith Zagona. *A Hidden Markov Model Combined with Climate Indices for Multidecadal Streamflow Simulation*. Vol. 50, 2014. https://doi.org/10.1002/2014WR015567.

Chow, V.T, and D. R. Maidment. *Applied Hydrology*. Mays McGraw Hill, 1988.

Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

Condon, Laura, S Gangopadhyay, and T Pruitt. *Climate Change and Non-Stationary Flood Risk for the Upper Truckee River Basin*. Vol. 11, 2014. https://doi.org/10.5194/hessd-11-5077-2014.

Evin, Guillaume, J Merleau, and L Perreault. *Two-Component Mixtures of Normal, Gamma, and Gumbel Distributions for Hydrological Applications*. Vol. 47, 2011. https://doi.org/10.1029/2010WR010266.

Griffis Veronica W., and Stedinger Jery R. "Evolution of Flood Frequency Analysis with Bulletin 17." *Journal of Hydrologic Engineering* 12, no. 3 (May 1, 2007): 283–97. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:3(283).

Interagency Committee on Water Data (IACWD). "Guidelines for Determining Flood Flow Frequency." Bulletin No. 17B revised and corrected , Hydrology Subcommittee, Washington, D.C., 1982.

Karim, MD., and J. Chowdhury. "A Comparison of Four Distributions Used in Flood Frequency Analysis in Bangladesh." *HydrologicalSciences -Journal- Des Sciences Hydrologiques,40,l,* February 1995.

Katz, Richard W. "Statistics of Extremes in Climate Change." *Climatic Change* 100, no. 1 (May 1, 2010): 71–76. https://doi.org/10.1007/s10584-010-9834-5.

Livneh, Ben, Eric A. Rosenberg, Chiyu Lin, Bart Nijssen, Vimal Mishra, Kostas M. Andreadis, Edwin P. Maurer, and Dennis P. Lettenmaier. "A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions." *Journal of Climate* 26, no. 23 (July 12, 2013): 9384–92. https://doi.org/10.1175/JCLI-D-12-00508.1.

Millington, Nick, Samiron Das, and S.P. Simonovic. "The Comparison of GEV, Log-Pearson Type 3 and Gumbel Distributions in the Upper Thames River Watershed under Global Climate Model." THE UNIVERSITY OF WESTERN ONTARIO DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING, September 2011.

Milly, P. C. D., Julio Betancourt, Malin Falkenmark, Robert M. Hirsch, Zbigniew W. Kundzewicz, Dennis P. Lettenmaier, and Ronald J. Stouffer. "Stationarity Is Dead: Whither Water Management?" *Science* 319, no. 5863 (February 1, 2008): 573. https://doi.org/10.1126/science.1151915.

Mondal, Arpita, and P. P. Mujumdar. "Hydrologic Extremes Under Climate Change: Non-Stationarity and Uncertainty." In *Sustainable Water Resources Planning and Management Under Climate Change*, edited by Elpida Kolokytha, Satoru Oishi, and Ramesh S.V. Teegavarapu, 39–60. Singapore: Springer Singapore, 2017. https://doi.org/10.1007/978-981-10-2051-3_2.

Pender, Douglas, Sandhya Patidar, Gareth Pender, and Heather Haynes. "Stochastic Simulation of Daily Streamflow Sequences Using a Hidden Markov Model." *Hydrology Research*, May 28, 2015. https://doi.org/10.2166/nh.2015.114.

Sun, Wei, Hao Zhang, Ahmet Palazoglu, Angadh Singh, Weidong Zhang, and Shiwei Liu. "Prediction of 24-Hour-Average PM2.5 Concentrations Using a Hidden Markov Model with Different Emission Distributions in Northern California." *Science of The Total Environment* 443, no. Supplement C (January 15, 2013): 93–103. https://doi.org/10.1016/j.scitotenv.2012.10.070.

Taylor, Karl E., Ronald J. Stouffer, and Gerald A. Meehl. "An Overview of CMIP5 and the Experiment Design." *Bulletin of the American Meteorological Society* 93, no. 4 (October 7, 2011): 485–98. https://doi.org/10.1175/BAMS-D-11-00094.1.

U.S. Geological Survey, 2016, National Water Information System data available on the World Wide Web (USGS Water Data for the Nation), accessed [June 10, 2012], at URL [http://waterdata.usgs.gov/nwis/].

Vogel, Richard, WO THOMAS, and TA MCMAHON. *Flood-Flow Frequency Model Selection in Southwestern United States*. Vol. 119, 1993. https://doi.org/10.1061/(ASCE)0733-9496(1993)119:3(353).

Zucchini, Walter, and Peter Guttorp. *A Hidden Markov Model for Space-Time Precipitation*. Vol. 27, 1991. https://doi.org/10.1029/91WR01403.