ABSTRACT

PATSKOSKI, JASON. Utility of Tree Rings and Future Climate Change Projections in Reservoir Sizing. (Under the direction of Sankarasubramanian Arumugam).

Reservoir sizing is one of the most important aspects of water resources engineering as the storage in a reservoir must be sufficient to supply water during extended droughts. Typically, observed streamflow is used to obtain required storage estimates based on the Sequent Peak Algorithm (SQP). However, observed streamflow data is often limited to less than 80 years of data, thereby increasing the uncertainty in storage estimates under high reliability. To overcome this, streamflow is stochastically generated to create multiple streamflow records with longer record length. Based on the generated flows, the SQP is used on each realization creating a distribution of required storages. The main limitation in this approach is that the parameters of the stochastic model are purely derived from the observed record and does not have information related to prehistoric droughts. Paleo streamflow records, usually estimated based on tree ring chronologies, provide better estimates of prehistoric droughts, but typically reconstructed flows underestimate high flow values since tree ring growth reaches its potential limit during wet years. To overcome this, we propose an alternate approach by utilizing information from Sea Surface Temperature (SST) and tree ring chronologies. Given the role of El Nino Southern Oscillation (ENSO) in influencing hydroclimatology over the Southeastern US, we estimate the periodic component of streamflow using Nino 3.4 – an index representing ENSO – and the non-periodic component of streamflow using the non-periodic interannual variability in the tree rings. The proposed tree ring and SST hybrid approach is tested with the traditional principal component regression (PCR) approach based on Leave-Five-Percent-Out-Cross-Validation. Results from
the study show the inclusion of SST provides better annual streamflow reconstruction estimates during high flow years. Combination of the annual streamflow estimates from the two models – PCR and the hybrid approach– results in improved estimates of reconstructed annual streamflow records.

To reduce the uncertainty in parameters of stochastic streamflow generations, we propose obtaining synthetic streamflow by combining observed and the improved reconstructed streamflow records within a Bayesian framework. The proposed methodology of combining paleo and observed records is compared with the traditional approach of using observed streamflow records alone in estimating reservoir storages based on split-sample validation. Results show that combining observed and reconstructed streamflow yield stochastic streamflow generation parameters more representative of the longer streamflow record, resulting in improved reservoir storage estimates. We also observe that the uncertainty in storage estimates decreases by incorporating reconstructed streamflows having longer lengths and higher skill in predicting observed flow.

Given that the reservoir sizing is typically estimated to meet future increase in water demand, there is no guarantee that future streamflow over the planning period will be representative of past streamflow records. In this context, general circulation models (GCMs) produce 10 – 30 year near-term climate change projections which could be utilized to obtain future climate conditions. Using the projected precipitation and temperature from the GCMs over a lumped watershed model, projected streamflow are obtained for 10 – 30 years. We propose combining observed, reconstructed and projected streamflow to generate synthetic streamflow records using a Bayesian framework. Based on split-sample validation, the proposed approach of utilizing paleo, observed and future streamflow records is compared to
using observed streamflow alone in obtaining reservoir storage estimates. Findings from the study show that the observed and reconstructed combination yielded generated streamflow traces more representative of future streamflow conditions over the validation period. The addition of near-term climate change projections provided limited/no improvements in reducing uncertainty on storage estimates.
Utility of Tree Rings and Future Climate Change Projections in Reservoir Sizing

by
Jason Mark Patskoski

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Civil Engineering

Raleigh, North Carolina

2014

APPROVED BY:

Sankarasubramanian Arumugam
Committee Chair

E. Downey Brill

Emily Berglund

Brian Reich
DEDICATION

I give God all the glory.
BIOGRAPHY

Jason Patskoski was born on November 18, 1987 in Cocoa Beach, Florida. He and his family moved to the once prosperous mill town of Greer, South Carolina when he was four years old. In Greer, he lived in house that backed up to the woods and spent most of his childhood climbing trees, making forts and swinging on vines. He lived in Greer until he graduated from high school. Greer is the place where he learned many of important life skills such as how to read, ride a bike, start a fire with a magnifying glass, balance a checkbook, shoot a gun, throw a curve ball, patch drywall, drive a stick shift, and post up a defender. In June of 2006 he graduated from Greer High School. During his time there he had the ability to dunk a basketball and run a quarter mile in fifty-one seconds.

After high school graduation, Jason attended Clemson University. While at Clemson, he attended many football and basketball home games, often camping out days for tickets. He still refers to his blood type as “orange.” Jason had an amazing experience at Clemson in every aspect. While there, he learned the art of computer programming, a skill to which he attributes much of his success. Other things he learned at Clemson include fluid mechanics, hydrology, statistics, how to play rugby and coincidentally how to use crutches. In May of 2010, he received a Bachelor of Science in Civil Engineering with a concentration in Applied Fluid Mechanics.
He then moved to Raleigh, North Carolina to pursue his Masters of Science in Civil Engineering at North Carolina State University which received in 2012 and has been pursuing his PhD.
ACKNOWLEDGMENTS

Most importantly, I would like to thank God who makes everything possible. Without his strength, I would not have been able to complete this work. I am truly blessed. As it says in Proverbs 16:9, “In his heart, a man plans his course, but the Lord determines his steps.”

I would also like to thank my advisor, Dr. Sankar Arumugam, for the opportunity to work with him. Not many people that grew up where I did get the chance to go to graduate school, and I would like to express my appreciation to Dr. Arumugam for believing in me. I would also like to thank him for his guidance and patience with me over the past two years. I truly do not know how I could have accomplished this task without his consistent teaching and direction throughout the process. I cannot thank him enough. I want to express my gratitude to the other committee members, Dr. Zechman and Dr. Brill, for their support during my time here at N.C. State. I would like to thank Dr. Kaye at Clemson University for his guidance and support of my future and opportunity for undergraduate research during my undergraduate career.

I wish to give my most sincere thanks to my parents, Mark and Mary Patskoski, for everything they have done for me throughout the years. They have sacrificed a lot in order to do what they felt was best for my wellbeing and future. They were always there to give me support, encouragement, and unconditional love. I would like to especially thank my dad for always being willing to do things like play catch with me or take me fishing when I was a kid. I realize how lucky I am to have a dad like that. I also appreciate him and my mom
teaching me the importance and value of work ethic. I definitely could not have achieved what I have without that value. Thanks mom and dad for raising me the right way.

I would also like to thank all my friends I have met while in graduate school, especially those in “The Domination Station” and on the “Honey Badgers” intramural team and its affiliates. You guys were always there to celebrate with me in the good times, willing to grab a beer and help me out in the not so good times and consistently went to eat General Tso’s just about every week. I hope I can repay you in the same fashion.
# TABLE OF CONTENTS

LIST OF TABLES .............................................................................................................................................. x

LIST OF FIGURES ............................................................................................................................................. xi

Chapter 1: Background ..................................................................................................................................... 1

Chapter 2: Predicting Streamflow in the Southeastern United States using SST and Tree Ring Chronologies ............................................................................................................................................. 5

  2.1: Introduction ............................................................................................................................................. 5
  2.2: Hydroclimatic Data Sources .................................................................................................................. 10
      2.2.1: Tree Ring Data .......................................................................................................................... 10
      2.2.2: Annual Streamflow Data .......................................................................................................... 12
      2.2.3: SST Data .................................................................................................................................. 13
  2.3: SSA based Annual Streamflow Reconstruction Methodology .................................................................. 14
      2.3.1: PCA model using Tree Ring Chronologies (TR-PCA) ................................................................. 16
      2.3.2: SST and Tree Rings based Reconstruction (SST-TR) using SSA ........................................... 18
          2.3.2.1: Basis for Selecting Nino3.4 .............................................................................................. 18
          2.3.2.2: Motivation for the SST-TR using SSA ............................................................................ 18
          2.3.2.3: SSA Methodology - Brief Description ............................................................................. 19
          2.3.2.4: SST-TR Methodology ...................................................................................................... 21
  2.4: Results and Discussion ............................................................................................................................. 25
      2.4.1: Multi-model Combination ......................................................................................................... 28
      2.4.2: Performance under cross-validation ......................................................................................... 34
      2.4.3: Discussion .................................................................................................................................. 41
  2.5: Summary and Conclusions ....................................................................................................................... 46

Chapter 3: Improved Reservoir Sizing Utilizing Observed and Reconstructed Streamflow within a Bayesian Framework ....................................................................................................................... 47

  3.1: Introduction ............................................................................................................................................. 47
  3.2: Data ....................................................................................................................................................... 50
  3.3: Methodology: Bayesian Framework and Experimental Design ............................................................. 51
      3.3.1: Sequent Peak Algorithm (SQP) ............................................................................................... 51
4.4.3: Reservoir Sizing under Validation.......................................................... 107
4.5: Conclusions................................................................................................. 108
Chapter 5: Conclusions and Future Work.......................................................... 110
   5.1: Conclusions............................................................................................... 110
   5.2: Future Work.............................................................................................. 113
REFERENCES ..................................................................................................... 116
APPENDIX ........................................................................................................... 124
   Appendix A: SSA Methodology Description................................................. 125
LIST OF TABLES

Table 2.1: List of USGS streamgage stations and the corresponding tree ring chronologies used for annual streamflow reconstruction ................................................................. 9

Table 2.2: Details of tree ring chronologies used in the study .................................................. 11

Table 2.3: Number of selected predictors (in parenthesis) and the total variance explained by the predictors resulting from stepwise regression ........................................ 24

Table 2.4: Results from the right-tailed hypothesis test based on Hotelling-Williams test evaluating the differences in correlations between the combined model and the TR-PCA model are greater than zero ............................................................................. 42

Table 3.1: List of USGS Streamgage Stations and the corresponding tree ring chronologies used for annual streamflow reconstruction ................................................................. 51

Table 3.2: Time periods and the reconstruction skill for each site used for the split-sample validation and the application on reservoir storage ......................................................... 62

Table 4.1: Details of USGS gauging stations considered for combining paleo and observed data and future climate change projections ................................................................. 84

Table 4.2: Performance of tree-ring based reconstruction model and the ‘abcd’ model in estimating the observed annual streamflow. Table also provides the paleo-information used for combining the three suites of hydroclimate data .................. 86

Table 4.3: Quantiles of the mean, standard deviation and lag-1 correlation of the projected streamflow from the ‘abcd’ model forced with monthly CMIP5 change projections for period 1981 – 2010 ................................................................. 91
LIST OF FIGURES

Figure 2.1: Locations of eight streamflow and tree ring chronology sites considered for the study over the Southeast US. ................................................................. 12

Figure 2.2: Schematic diagram of the TR-PCA (white) and SST-TR (all) (grey) methodologies employed for annual streamflow reconstruction. The steps from the SST-TR methodology from Section 3.2 are indicated in black circles......... 15

Figure 2.3: Scatter plot between observed annual streamflow and estimated streamflow for (a) Site 2 (b) Site 5 under three flow categories based on TR-PCA and SST-TR(sel) methods.................................................................................................. 17

Figure 2.4: Performance of four different reconstruction models in terms of (a) correlation and (b) NRMSE in predicting the observed annual streamflow for the eight selected sites............................................................................................................. 31

Figure 2.5: Performance (NRMSE) of four different streamflow reconstruction models in predicting the observed streamflow during (a) above-normal flow conditions and (b) below-normal flow conditions over the eight selected sites. ....................... 32

Figure 2.6: Performance of four different reconstruction models in terms of (a) correlation and (b) NRMSE in predicting the observed annual streamflow under leave-five-percent-out cross-validation for the eight selected sites. ....................................... 36

Figure 2.7: Performance of four different reconstruction models in terms of NRMSE in predicting the observed annual streamflow during (a) above-normal flow conditions and (b) below-normal flow conditions based on under leave-five-percent-out cross-validation for the eight selected sites. ....................... 37

Figure 2.8: Adjusted $R^2$ for the TR-PCA and SST-TR models in predicting the observed annual streamflow under leave-five-percent-out cross-validation for the eight selected sites................................................................................................................. 40

Figure 3.1: Location of USGS streamflow sites and the tree ring chronologies used for reconstruction in this study. ................................................................. 52

Figure 3.2: Experimental Design for validation the utility of reconstructed streamflow in improving the reservoir design. ................................................................................. 60

Figure 3.4: Difference between true storage and estimated storages using different reconstructed streamflow lengths under split-sample validation for the selected
sites. Demand is assumed to be 80% of the mean annual streamflow for respective sites. .......................................................... 67

Figure 3.5: Storage Distribution for the selected sites obtained by the Bayesian combination of observed and reconstructed flow. .......................................................... 68

Figure 3.6: Difference between the 95th storage quantile obtained using the combined flows with current skill and the corrupted combined flows having different skill and sample length. Contours denote the difference in million acre-feet. ............... 71

Figure 3.7: Difference between the 95th storage quantile obtained using the combined flows with perfect skill and the corrupted combined flows having different skill and sample length. Contours denote the difference in million acre-feet ............... 72

Figure 4.1: Map of the streamflow and tree ring chronology sites used for this study ........ 82

Figure 4.2: Modeling framework for obtaining the reconstructed, observed and projected streamflow for the selected two stations .......................................................... 89

Figure 4.3: Comparison between the reconstructed and the observed streamflow for the two sites ........................................................................................................ 90

Figure 4.4: Comparison between observed and ‘abcd’ Model simulated streamflow for the assumed observational period over the selected sites ........................................ 92

Figure 4.5: Methodology for obtaining the likelihood ratio distribution for each streamflow combination scenarios .......................................................................................... 95

Figure 4.6: The CDF of the ratio of likelihood distribution to the likelihood under climatological streamflow .......................................................... 103

Figure 4.7: Difference between the flow values obtained under observed/observed-reconstructed streamflow traces and flow values under validation period for different exceedence probabilities. The secondary y-axis is shown for the actual FDC for the validation period .......................................................... 105

Figure 4.8: Difference between validation storage and estimated storages using different streamflow scenarios for the selected sites. Demand is assumed to be 80% of the mean annual streamflow for respective sites .................................................................. 106
Chapter 1: Background

One of the most important goals of water supply systems is having the ability to supply the water demand for designed uses (e.g., irrigation, domestic consumption) with as few failures as possible (i.e. failure probability). To ensure this, reservoirs are designed to have adequate storage so that demands during periods of below-normal inflows could be met. The required storage for a reservoir to meet the specified demand with inflows is usually determined using the Sequent Peak Algorithm (SQP) [Thomas 1963]. The SQP is a sequential process which estimates the required storage needed to meet the specified demand for the observed inflow. However, for the given time series this algorithm only provides one storage estimate which could vary depending on the observed streamflow record length. To overcome this, synthetic streamflow time series are often generated using the moments of observed streamflow to develop a distribution of storage estimates [Stedinger 1982, Vogel et al. 1987, Vogel 1988]. Based on the desired reliability (eg: 90%) of supply, the corresponding percentile (90%) of the storage distribution could be considered for design purposes.

The above approach to reservoir design is limited since it is dependent only on the information from the observed streamflow record, which is usually limited to around 80 years. Since the desired lifespan often exceeds the length of the observed streamflow record, estimating storage estimates for high reliability often has large sampling variability. Furthermore, analysis of paleo information can potentially show that the observed streamflow misses unobserved severe drought episodes. To gain insight into drought conditions prior to the observed period, streamflow is often reconstructed by developing
regressions between observed streamflow and paleo-data and then using the model to estimate annual streamflow for the available paleo-data [Fritts 1976, Cook et al. 1983, Woodhouse et al. 2006, 2010].

The most common type of paleo-data used in streamflow reconstruction is tree ring chronologies due to their long length as well as owing to their ability to indicate the available moisture. However, tree ring chronologies are limited in their ability to predict high flow years [Margolis et al. 2011]. Trees have a metabolic limit on growth, so streamflow during above-normal years are often underestimated. Furthermore, high streamflow years resulting from wet years are often accompanied by high cloud cover and lower temperatures, both of which are limiting factors in tree growth [Fritts 1976].

Seasonal streamflow forecasts in the Southeastern US have often been developed using Sea Surface Temperature (SST) conditions since streamflow in this area is significantly influenced by oceanic conditions [Oh et al. 2012, Caraway et al. 2012]. Unlike tree ring chronologies, a strong monotonic relationship exists between SST anomalies and streamflow, thereby having the ability to predict high flow years. One of the goals of this study is to improve streamflow reconstruction in high flow years by using both tree ring chronologies and SSTs. The proposed SST and tree ring model will be compared with the traditional approach of using just tree ring chronologies alone in estimating annual streamflow over the Southeastern US.

Apart from understanding the long-term hydroclimatology using tree rings, the reconstructed streamflow could also be utilized in reservoir design. Previous studies have used paleo-data for improving design estimates related to water management. Paleo-flood
records have been combined with observed annual maximum events to improve design flood estimates [Hosking 1986; Stedinger 1986]. Studies have also combined observed and reconstructed data in a Bayesian framework for hydrologic time series generation [Vicens et al 1975, Valdes et al. 1977, Prairie et al 2008, Henley et al. 2011]. Given the uncertainty in reconstructed streamflow records, most Bayesian frameworks have considered only the state of reconstructed flow (eg: wet/dry) and have resampled the observed conditioned on the state of reconstructed flow [Prairie et al 2008, Henley et al. 2011]. Given that we obtain improved reconstructed flow from this study, instead we combine reconstructed streamflow values with the observed streamflow values to gain insight into the long term drought record. Thus an allied goal of this study is to quantify the value of paleo information in reducing uncertainty in reservoir sizing. For this purpose we combine observed and reconstructed streamflow using a Bayesian framework. The combined streamflow will then be used in the stochastic streamflow generation model to obtain the distribution of reservoir storages.

Although, using reconstructed streamflow could improve reservoir storage estimates, this approach has limitations since it considers only information from the past. Water resources planning typically focuses on meeting future demand, which requires estimates of future inflow conditions. Furthermore, under climate change, future streamflow information may not be representative of the past streamflow records [Gleick 1987, Lettenmaier et al. 1992, Gleick and Chalecki 1999, McCabe and Wolock 1999, Sankarasubramanian et al. 2001, Sinha and Cherkauer, 2010]. One approach to estimate future streamflow is to use provided near-term climate change projections from general circulation models (GCMs) which are available for 10 – 30 years into the future. Using the projected precipitation and
temperature from the GCMs and using the statistical relationships between observed climate and streamflow, projected streamflow can be obtained 10 – 30 years into the future [Maurer et al. 2007, Singh et al. 2014]. As part of this study, we combine streamflow obtained from near-term climate change projections with the time series of observed and reconstructed streamflow for developing estimates of future reservoir storage. We estimate the parameters of the stochastic streamflow generation model by using the observed, reconstructed and projected streamflow in a Bayesian framework.

The outline of this dissertation is as follows: Chapter 2 provides the methodology for improving streamflow reconstructions by utilizing both SSTs and tree ring chronologies. In Chapter 3, these improved reconstructed flows are combined with observed streamflow to demonstrate the utility of reconstructed streamflow in reducing the uncertainty on reservoir storage estimates. Chapter 4 considers reservoir storage estimation under future planning period by considering the time series of observed, reconstructed and projected streamflows. Finally, the findings are summarized in Chapter 5 along with a discussion of future work.
Chapter 2: Predicting Streamflow in the Southeastern United States using SST and Tree Ring Chronologies

2.1: Introduction

Long records of annual streamflow provide crucial information on the available water for various uses including public water supply and irrigation. Time series of annual streamflow are also useful for reservoir sizing and for estimating annual release (also known as yield) for the specified reliability (usually 95%). The higher the reliability of release, the more we depend on accurate estimates of annual streamflow corresponding to periods of droughts. But, observed streamflow information for most sites over the Southeast United States is available only from 1930’s indicating considerable uncertainty could arise in the estimates of 95% reliable yield which is obtained from 80-years of annual streamflow record. Hence, studies have focused on extending the streamflow record beyond the observed realm using paleo-proxies such as tree rings [Cook et al. 1983; Woodhouse et al. 2001, 2006]. Developing reconstructed streamflow records can give insight into past variability and droughts, thereby reducing the uncertainty on reliability estimates of reservoir yields [Gangopadhyay et al. 2009; Devineni et al. 2013]. One limitation of using tree-ring chronologies for streamflow reconstruction is the underestimation of high flow values [Meko et al. 1995; Gangopadhyay et al. 2009]. This study proposes a new methodology to overcome this limitation by reconstructing annual streamflow using tree-ring chronologies and sea surface temperature (SST) over selected stations over the Southeast United States.
It is well established that reconstructed streamflow records using tree-ring chronologies explain a significant amount of observed variability in annual streamflow [Fritts et al. 1976, 1991; Stockton and Jacoby, 1976; Hidalgo et al. 2000]. Each year, trees add a new ring whose width is dependent on many factors that include amount of sunlight, air temperature, root depth, and most importantly, moisture availability [Fritts, 1976]. The basic idea for reconstructing streamflow is to develop a statistical relationship between observed annual streamflow records (predictand) and tree-ring chronologies (predictor) for the corresponding period such that the statistical relationship could be employed for extending the annual streamflow beyond the observed period using the past tree-ring chronologies. The annual streamflows within the Upper Colorado River Basin were reconstructed using the Principal Components (PCs) of tree-ring chronologies based on forward stepwise regression [Stockton and Jacoby, 1976; Fritts 1991; Hidalgo et al. 2000; Woodhouse et al. 2006]. Past studies indicate that this approach has become the most popular technique in annual streamflow reconstruction [Fritts et al. 1990, Brockway et al. 1995]. Given tree rings provide an accurate estimate of the frequency and intensity of droughts, studies have also focused on reconstructing Palmer Drought Severity Index (PDSI) using tree rings [Maxwell et al. 2011; Kaufman and Vonck; 2011]. Multivariate techniques such as canonical correlation techniques have also been considered for reconstructing PDSI based on tree rings [Cook and Jacoby, 1983]. More complicated models based on hierarchical Bayesian approach have also been employed for reconstructing streamflow [Devineni et al. 2013] as well as other climate fields [Li et al., 2010; Tingley and Huybers 2010a,b] using tree-ring chronologies.
Studies have also considered using both tree-ring chronologies and SST for improving the reconstructions. Kitzberger et al., [2007] associated local wildfire chronologies reconstructed from fire scars on tree rings across western North America with SSTs across the Pacific and Atlantic and found that SSTs over the Pacific Ocean influenced variations in fire at interannual to decadal scales, whereas the SST conditions over the Atlantic Ocean modulated the strength and spatial influence of wildfire occurrence at multidecadal scales. Recently, Anderson et al., [2012a, 2012b] improved reconstructions of snow water equivalent obtained using tree-chronologies for the Upper Green River basin, a major tributary of the Upper Colorado River basin, by considering Southern Oscillation Index and Pacific Decadal Oscillation Index as additional predictors using stepwise linear regression.

Irrespective of these advances on the methodologies and application of tree rings for reconstructing various hydroclimatic attributes, the primary limitation in using tree-ring chronologies for streamflow reconstruction is the underestimation of high flow values since tree ring growth is often limited by soil properties as well as by the metabolic growth rate of the tree [Meko et al. 1995; Gangopadhyay et al. 2009]. In addition, high flow values which are commonly due to increased precipitation are usually accompanied by significant cloud cover resulting in reduced sunlight and lower temperatures, both of which limit tree growth [Fritts 1976]. Recent studies also document this limitation indicating reconstructed annual streamflows yield conservative estimates of high flow events [Margolis et al. 2011].

The overall goal of this study is to propose a new methodology for reconstructing streamflow that better estimates the high flow events, thereby improving the overall skill in
reconstructing annual streamflow. The idea is to use Sea Surface Temperature (SST) conditions as an additional predictor along with tree-ring chronologies for improving the streamflow reconstruction over the Southeast United States. Studies have shown that El-Nino Southern Oscillation (ENSO) conditions significantly influence the hydroclimatology of the Southeast United States [Boyles and Raman, 2003; Hayhoe et al. 2007; Seager et al., 2009; Peterson et al. 2012]. Thus, we expect the proposed methodology to be applicable for basins whose annual streamflow is significantly influenced by anomalous conditions in SST. The methodology employs Singular Spectrum Analysis (SSA) developed by Ghil et al. [2002] for separating the quasi-periodic ENSO components from streamflow, so that a separate statistical relationship could be developed between the quasi-periodic components of annual streamflow and SST. We apply the hybrid methodology for eight selected basins over the Southeast United States having long time series of tree-ring chronologies. These eight basins belong to the Hydroclimatic Data Network (HCDN) [Slack et al., 1993; Sankarasubramanian and Vogel, 2002] whose streamflow records are minimally influenced by reservoir storage and groundwater pumping.
Table 2.1: List of USGS streamgage stations and the corresponding tree ring chronologies used for annual streamflow reconstruction

<table>
<thead>
<tr>
<th>Site Index</th>
<th>USGS Station Number</th>
<th>Station Name</th>
<th>Mean Annual Streamflow [cfs]</th>
<th>Drainage Area [mi²]</th>
<th>Tree Chronologies Used [ITRDB Number]</th>
<th>Observed Streamflow used for Model Development</th>
<th>Number of Reconstructed years (from 1857)</th>
<th>Correlation (Q_{an}, Nino3.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>02132000</td>
<td>Lynches River at Effingham, SC</td>
<td>1009</td>
<td>1030</td>
<td>SC004, SC006, NC008</td>
<td>1930 - 1985</td>
<td>74</td>
<td>0.236</td>
</tr>
<tr>
<td>2</td>
<td>02136000</td>
<td>Black River at Kingstree, SC</td>
<td>937</td>
<td>1252</td>
<td>SC004, SC006, NC008</td>
<td>1930 - 1985</td>
<td>74</td>
<td>0.282</td>
</tr>
<tr>
<td>3</td>
<td>02198000</td>
<td>Brier Creek at Millhaven, GA</td>
<td>603</td>
<td>646</td>
<td>SC004, SC006, GA003</td>
<td>1938 - 1985</td>
<td>82</td>
<td>0.234</td>
</tr>
<tr>
<td>4</td>
<td>02232500</td>
<td>St. Johns River near Christmas, FL</td>
<td>1298</td>
<td>1539</td>
<td>FL007, FL008</td>
<td>1934 - 2003</td>
<td>78</td>
<td>0.299</td>
</tr>
<tr>
<td>5</td>
<td>02236000</td>
<td>St. Johns River near De Land, FL</td>
<td>3055</td>
<td>3066</td>
<td>FL007, FL008</td>
<td>1934 - 2003</td>
<td>78</td>
<td>0.282</td>
</tr>
<tr>
<td>6</td>
<td>02246000</td>
<td>North Fork Black Creek near Middleburgh, FL</td>
<td>188</td>
<td>177</td>
<td>FL005, FL007, FL008</td>
<td>1932 - 1993</td>
<td>76</td>
<td>0.278</td>
</tr>
<tr>
<td>7</td>
<td>02321500</td>
<td>Santa Fe River at Worthington Springs, FL</td>
<td>410</td>
<td>575</td>
<td>FL005, FL007, FL008</td>
<td>1932 - 1993</td>
<td>76</td>
<td>0.330</td>
</tr>
<tr>
<td>8</td>
<td>02322500</td>
<td>Santa Fe River near Fort White, FL</td>
<td>1500</td>
<td>1017</td>
<td>FL005, FL007, FL008</td>
<td>1933 - 1993</td>
<td>75</td>
<td>0.268</td>
</tr>
</tbody>
</table>
The chapter is organized as follows: Section 2.2 provides information on the data sources related to streamflow, tree-ring chronologies, and SST. Following that, we describe the hybrid methodology using SSA that employs SST and tree-ring chronologies for reconstructing annual streamflow in eight selected HCDN stations over the Southeast United States. Next, we evaluate the performance of the SSA-based methodology with the traditional stepwise regression through rigorous cross-validation. Finally, we summarize the salient findings and conclusions along with potential applications arising from the improved reconstructed streamflow.

**2.2: Hydroclimatic Data Sources**

The primary challenge in reconstructing annual streamflow over the Southeast United States is the limited availability of tree-ring chronologies [Stahle, 2012]. After matching the HCDN stations with the available tree-ring chronologies, we found (Figure 2.1) eight HCDN stations (Table 2.1) for reconstructing annual streamflows over the Southeast United States.

**2.2.1: Tree Ring Data**

The tree-ring chronologies for the study were obtained from the National Atmospheric and Oceanic Administration (NOAA) International Tree Ring Data Bank (ITRDB) [available online at http://www.ncdc.noaa.gov/paleo/treering.html]. Tree Ring chronologies are standardized by removing the growth trend in the tree ring measurements due to the natural aging process of a tree. The selection of tree-ring records for streamflow reconstruction is often limited to chronologies located within the watershed of the selected station [Hidalgo et al. 2000; Woodhouse et al. 2006]. However, in the Southeast, number of
tree-ring chronologies are limited, which restricts the potential for hydroclimatic reconstructions [Cook et al. 1999; Seager et al. 2009; Stahle, 2012].

Table 2.2: Details of tree ring chronologies used in the study

<table>
<thead>
<tr>
<th>Tree Ring Chronology</th>
<th>Site Name</th>
<th>Start Year</th>
<th>End Year</th>
<th>Species</th>
<th>Embedding Dimension (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC008</td>
<td>Black River NC</td>
<td>365</td>
<td>1985</td>
<td>Baldcypress</td>
<td>45</td>
</tr>
<tr>
<td>SC004</td>
<td>Four Holes Swamp</td>
<td>1001</td>
<td>1985</td>
<td>Baldcypress</td>
<td>60</td>
</tr>
<tr>
<td>SC006</td>
<td>Black River SC</td>
<td>549</td>
<td>1993</td>
<td>Baldcypress</td>
<td>70</td>
</tr>
<tr>
<td>GA003</td>
<td>Ebenezer Creek</td>
<td>990</td>
<td>1985</td>
<td>Baldcypress</td>
<td>40</td>
</tr>
<tr>
<td>FL005</td>
<td>Suwannee River</td>
<td>1725</td>
<td>1993</td>
<td>Overcup Oak</td>
<td>40</td>
</tr>
<tr>
<td>FL007</td>
<td>Lower Withlacoochee River</td>
<td>1618</td>
<td>2005</td>
<td>Baldcypress</td>
<td>60</td>
</tr>
<tr>
<td>FL008</td>
<td>Middle Withlacoochee River</td>
<td>1646</td>
<td>2003</td>
<td>Baldcypress</td>
<td>80</td>
</tr>
</tbody>
</table>

In this study, we selected tree-ring chronologies that lie within 200 km of the USGS stream gaging stations (Figure 2.1). It is important to note that multiple chronologies were used for reconstructing streamflow at a given site (Table 2.1) due to limited number of tree-ring chronologies available over the region. Recent studies indicate that large number of species used for tree-ring-based reconstructions could improve reconstructed time series [Cook and Pederson, 2010; Maxwell et al. 2011; Pederson et al. 2013]. Most of the chronologies listed in Table 2.2 were developed from more than 10 trees in the region. Details of species for each tree-ring chronologies are provided in Table 2.2. With the exception of FL005 (Overcup Oak), all the sites had chronologies from Baldcypress trees. Baldcypress is a deciduous coniferous tree native to the Southeast United States, Mexico, and western Guatemala [Stahle et al. 2012]. Recently, Stahle et al. [2012] considered the tree-ring chronologies listed in Table 2.2 as well as other chronologies in the region and found that the
contribution of moisture from ENSO was indicated in the sub-seasonal chronologies over the Southeast United States. Additional details on the chronologies including the contributors could be found from ITRDB (http://www.ncdc.noaa.gov/paleo/treering.html).

Figure 2.1: Locations of eight streamflow and tree ring chronology sites considered for the study over the Southeast US.

2.2.2: Annual Streamflow Data

Given the purpose of the study is to reconstruct annual streamflow records using SST and tree-ring chronologies, we considered only undeveloped/virgin basins from the Hydro-Climatic Data Network (HCDN) [Slack et al. 1993]. Streamflow records in the HCDN
database are not impacted by upstream storage or groundwater pumping. USGS has identified these basins as virgin watersheds primarily for the purpose of understanding the role of climate in influencing the land-surface response. Several studies have considered HCDN sites for quantifying the role of climate in influencing the land-surface response at both regional scale [Oh and Sankarasubramanian, 2012] and at continental scale [Petersen et al., 2012]. The locations of the selected streamflow sites can be seen in Figure 2.1 along with the locations of tree-ring chronologies used for reconstruction. Information summarized in Table 2.1 show that drainage area of the basins ranges from 177 square miles (mi²) to 3066 mi², which typically capture the role of climate variability on basin response.

2.2.3: SST Data

To improve the prediction of high flow values in streamflow reconstruction, we considered SST as an additional predictor. As discussed in Section 2.1, studies have shown that El-Nino Southern Oscillation (ENSO) conditions significantly influence the hydroclimatology of the Southeast United States [Boyles and Raman, 2003; Hayhoe et al. 2007; Seager et al., 2009; Peterson et al. 2012]. Previous studies have used Nino3.4 index to denote the ENSO state [Trenberth and Stepaniak 2001; Devineni and Sankarasubramanian, 2010a, 2010b]. Nino3.4 index denotes the average anomalous SST conditions over the area 5°S–5°N and 170°–120°W in the tropical Pacific. We considered the annual average Nino3.4 (October-September) for the period 1857 through 2003 for annual streamflow reconstruction. Since it is typical to consider annual streamflow values in water year, which is from October-September, we considered annual average Nino3.4 over the same period. In the study, annual
average annual Nino3.4 was calculated using Kaplan’s analyses of
(http://iridl.ldeo.columbia.edu/SOURCES/.KAPLAN/.Indices/.NINO34/) global sea surface
temperatures [Kaplan et al. 1998]. Table 2.1 shows the correlation between annual
streamflow and Nino3.4 for the observed period over the water year. All the correlations
reported in Table 2.1 are statistically significant at 95% confidence interval. Typically,
positive (negative) anomalous conditions in Nino3.4, El Nino (La Nina), over the tropical
Pacific usually results in wet (dry) and cool (warm) winters [Ropelewski and Halpert, 1987]
over the Southeast United States. ENSO influence over the Southeast United States is
usually weak during the summer exhibiting dry (wet) and warm (cold) climatic conditions
during El Nino (La Nina) in the tropical Pacific [Ropelewski and Halpert, 1987, Raman and
Boyles, 2003; Seager et al., 2009]. Our interest here is to utilize the dependency of annual
hydroclimatology of the Southeast United States with ENSO so that reconstructed
streamflow provide better estimates during high flow conditions. Given that SST data is
available from 1856, we intend to improve the reconstructed streamflow starting from the
water year 1857 till the beginning of the instrumental records available for a given site (Table
2.1).

2.3: SSA based Annual Streamflow Reconstruction Methodology

This section provides a detailed description of the employed methodology based on
SSA utilizing SST and tree-ring chronologies. Our baseline comparison of the SSA
methodology is the traditional approach that uses principal components analysis (PCA) of
tree-ring chronologies (TR-PCA) as predictors for developing reconstructed streamflow
using stepwise regression [Hidalgo et al. 2000; Woodhouse et al. 2006]. To improve upon the limitations of the TR-PCA model (white-boxes in Figure 2.2), a model combining tree rings and SSTs (SST-TR) was developed as shown (gray-shaded boxes in Figure 2.2) using SSA [Ghil et al. 2002]. Figure 2.2 provides the detailed flow chart for developing reconstructed streamflow for both TR-PCA and SST-TR.

Figure 2.2: Schematic diagram of the TR-PCA (white) and SST-TR (all) (grey) methodologies employed for annual streamflow reconstruction. The steps from the SST-TR methodology from Section 3.2 are indicated in black circles.
2.3.1: PCA model using Tree Ring Chronologies (TR-PCA)

The proximity of tree-ring chronologies cause them to be inter-correlated and the most common approach to address the linear codependency in streamflow reconstruction is Principal Component Analysis (PCA) [Hidalgo et al. 2000]. When a multiple linear regression model contains predictors that are highly inter-correlated, regression coefficients estimates can be inaccurate and unstable [Weisberg, 1985]. PCA creates a new set of variables called PCs, which are orthogonal to one another, by transforming the original data set into linear combinations of the original variables [Almanaseer and Sankarasubramanian, 2012]. Eigen-value analysis from a covariance matrix of the original data set extracts the principal components (PCs) from the data set. PCA will yield the same number of PCs as the number of variables in original data set with each explaining a percentage of the total variance. In the study, PCA is performed on the tree-ring chronologies, retaining 90% of the total variance, and stepwise regression is used to develop a relationship between observed annual streamflow values and the PCs (TR-PCA model). Previous studies have employed stepwise regression for streamflow reconstructions in the Upper Colorado River Basin [Hidalgo et al. 2000; Woodhouse et al. 2006]. Using stepwise regression greatly reduces the possibility of model “over fitting” by selecting one predictor at a time based on statistical significance (Figure 2.2) [Anderson et al., 2012a, 2012b].
Figure 2.3: Scatter plot between observed annual streamflow and estimated streamflow for (a) Site 2 (b) Site 5 under three flow categories based on TR-PCA and SST-TR(sel) methods.
2.3.2: SST and Tree Rings based Reconstruction (SST-TR) using SSA

2.3.2.1: Basis for Selecting Nino3.4

Seager et al. [2009] provided detailed analyses on the source of variability in precipitation over the Southeast United States. Based on both observed SST data and predicted SST data, Seager et al. [2009] clearly showed that the winter precipitation (November-March) is modulated by anomalous conditions in the tropical Pacific even though the tropical SST signal was weak during the period 1922-1950 (see Figure 2.3 in Seager et al. 2009). The study also clearly showed that summer precipitation (April-October) over the Southeast United States is purely influenced by the internal atmospheric variability (see Figure 2.5 in Seager et al. 2009). Perhaps the most interesting finding from the study of Seager et al. [2009] is that both observational analysis and model results have indicated no influence of SST anomalies over the Atlantic Ocean on the Southeast precipitation. Given this background on climate variability over the Southeast United States, we only considered one additional SST predictor over the tropical Pacific which is denoted by Nino3.4. Thus, the primary predictors for reconstructing the streamflow under TR-SST are the observed tree-ring chronologies and annual average Nino3.4 over the water year.

2.3.2.2: Motivation for the SST-TR using SSA

The simplest approach one could consider for reconstruction using SST is to develop a principal components regression similar to TR-PCA using both tree rings and Nino3.4 as predictors. Instead, we separate the ENSO signal on observed streamflow as periodic and non-periodic components using SSA [Ghil et al., 2002] and develop separate regression
relationships. Since ENSO is the primary source of moisture transport from outside the region [Seager et al., 2009; Devineni and Sankarasubramanian, 2010a], we intend to relate the statistically significant periodic component of streamflow with the periodic component of Nino3.4 (Figure 2.2 – shaded gray boxes under Nino3.4). The remaining component of streamflow, the non-periodic component from SSA, which primarily constitutes the moisture transport within the region will be captured by the tree-ring chronologies (shaded gray boxes under tree-ring chronologies in Figure 2.2). Thus, a separate regression relationship is developed between the non-periodic components of the streamflow and the tree-ring chronologies using stepwise regression. Both the estimated components – periodic components from SST and the non-periodic components from tree-ring chronologies – are added together to develop the reconstructed streamflow.

2.3.2.3: SSA Methodology - Brief Description

To separate the periodic and non-periodic components of streamflow and tree-ring chronologies, we employ SSA developed by Ghil et al. [2002]. SSA is a time series approach that quantifies the lag dependence in streamflow through estimates of lagged cross-covariance [Shun and Duffy, 1999; Ghil et al. 2002]. Like PCA, SSA produces orthogonal PCs, the variance of the PCs (eigen values) and the eigenvectors corresponding to each PC. Similarly, SSA also brings each PC obtained from the lagged cross-covariance matrix back to the original space creating reconstructed components (RCs). If one uses all the RCs for reconstruction, it will reproduce the original time series. Given that SSA employs PCA on
lagged time series of tree rings, year-to-year storage characteristics or persistence of annual streamflow would also be incorporated in the reconstructed streamflow.

Given the time series of tree-ring chronology at a given site or annual average Nino3.4 over the tropical Pacific with length $N$, SSA creates a lagged matrix of the time series by considering embedding dimension, $M$, which denotes the number of lagged time series to be created. A concise mathematical description and related equations for SSA is provided in Appendix A. Table 2.2 provides the embedding dimension for each tree-ring chronology. The embedding dimension is usually chosen within $1/4$ to $1/3$ of $N$. A covariance matrix of dimension $M \times M$ is estimated using the lagged matrix and then decomposed into the respective eigenvectors and eigenvalues using orthogonal decomposition. Thus, eigenvectors of the decomposed covariance matrix could be multiplied the original values of the time series to get the principal components from the SSA. Thus, there will be a PC for each column of the lagged matrix totaling $M$. Similar to PCA, the orthogonal PCs of the SSA also could be brought back to the original time space by convolving and normalizing the corresponding PCs and eigenvectors to create reconstructed components (RCs). Given that the eigenvector from the lagged covariance matrix has temporal signature, they carry the periodic signals exhibited by the original time series. Thus, each RC will also share the same periodicity as its associated eigenvector. RCs can be added together to recreate the original time series. Thus, by adding RCs that exhibit similar periodicities, we can extract the desired periodic components of the original time series. Similarly, all the non-periodic RCs could be added together to obtain the non-periodic component of the time series. Thus,
periodic components of a time series could be extracted based on signal to noise separation of oscillatory pairs of eigen elements [Ghil et al. 2002].

Monte Carlo Singular Spectrum Analysis (MC-SSA) are typically employed to ensure that the identified oscillatory pairs of eigen elements are statistically significant [Allen, 1992]. Climate time series, such as Nino3.4, often have greater power at lower frequencies, so a red noise “null hypothesis” is often used for periodic component detection [Allen, 1992; Ghil et al. 2002]. MC-SSA generates simulated red noise data (of length N) from the original time series and SSA is performed on the simulated data to obtain the eigenvalues. The generation of red noise data and subsequent calculation of eigenvalues are performed a large number of times giving an ensemble of eigenvalues. If a corresponding eigenvalue from SSA of the original time series lies outside the 95% confidence range of the ensemble of eigenvalues, then the null hypothesis that the eigenvalue exhibits periodicity is accepted. Thus, the RCs associated with an eigenvalue outside the confidence range can be said to be “periodic.” Since ENSO exhibits periodicity of three to seven years, the RCs associated with ENSO will have a similar periodicity.

2.3.2.4: SST-TR Methodology

The SST-TR model utilizes MC-SSA to separate periodic and non-periodic components of streamflow, tree-ring chronologies, and SSTs so that separate regressions for periodic and non-periodic components of streamflow can be created as seen in Figure 2.2. Steps for reconstructing streamflow for a given site using the proposed methodology are given below (Figure 2.2).
1. Given ‘p’ number of annual tree-ring chronologies available for a site and the annual (Oct-Sep) average Nino3.4, separate both predictors into periodic components and non-periodic components based on the eigen-value spectrum using MC-SSA. Periodic components only in the ENSO frequency of 3-7 years are only separated.

2. Reconstruct all the non-periodic components (from step 1) into the original space for a given tree-ring chronology (Equation A-4) and add all the resulting RCs together to create an original time series of tree-ring chronologies that has information only on the non-periodic component. Repeat this step for all the ‘p’ tree-ring chronologies to develop ‘p’ tree-ring chronologies with no periodic component.

3. Similarly, reconstruct the periodic components of Nino3.4 to the original space and add them together to develop the periodic components of Nino3.4.

4. Separate the predictand, the time series of average annual streamflow, into periodic (3-7 years) and non-periodic components (see bottom of Figure 2.2) using SSA. Reconstruct all the non-periodic and periodic components of the streamflow back in the original space using the respective eigen vectors (Equation A-4) obtained from SSA on streamflow. Add all the reconstructed non-periodic components to create an original streamflow time series that has only information on the non-periodic component. Similarly, add the reconstructed periodic components to create a time series of streamflow that has only periodic component.

5. Employ stepwise regression [Hocking, 1976] between the periodic reconstructed component of streamflow (step 4) and the periodic reconstructed components of
Nino3.4 (step 3). This regression explains the variability in the moisture transport due to the tropical Pacific.

6. Since the ‘p’ reconstructed non-periodic component of tree-ring chronologies (from step 2) are correlated to each other, perform PCA on them to obtain principal components of non-periodic component of tree-ring chronologies.


8. Use regression relationship from step 5 to estimate the periodic component of streamflow from the periodic component of Nino3.4 (step 3).


10. Add periodic component estimate of streamflow from step 8 and non-periodic component of streamflow from step 9 to obtain the reconstructed streamflow for both the observed period and the reconstruction period beginning 1857.

We considered two different models of SST-TR with the outlined approach above using stepwise regression to select the principal components of non-periodic components of streamflow. We call this as SST-TR (all). We also considered another variation, SST-TR (selected), in which the reconstructed non-periodic components for a given tree-ring were selected based on their correlation with the reconstructed non-periodic component of streamflow. Thus, the components with significant correlation were alone added together to
recreate the non-periodic reconstructed component for a given tree ring in step 2. Rest of the procedure (steps 3-10) remains the same for SST-TR (sel).

Table 2.3: Number of selected predictors (in parenthesis) and the total variance explained by the predictors resulting from stepwise regression

<table>
<thead>
<tr>
<th>Site</th>
<th>Percent Variance Explained by Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tree Rings [TR-PCA]</td>
</tr>
<tr>
<td>1</td>
<td>0.730 (1)</td>
</tr>
<tr>
<td>2</td>
<td>0.730 (1)</td>
</tr>
<tr>
<td>3</td>
<td>0.711 (1)</td>
</tr>
<tr>
<td>4</td>
<td>0.630 (1)</td>
</tr>
<tr>
<td>5</td>
<td>0.630 (1)</td>
</tr>
<tr>
<td>6</td>
<td>0.492 (1)</td>
</tr>
<tr>
<td>7</td>
<td>1.000 (3)</td>
</tr>
<tr>
<td>8</td>
<td>0.492 (1)</td>
</tr>
</tbody>
</table>

Table 2.3 provides the percentage variance explained by the predictors under each of the models that include TR-PCA, SST-TR (sel) and SST-TR (all). For instance, in Site 1, the stepwise regression in the TR-PCA model selects one of the available predictors (PCs) which explain 73% of the total variance of all the tree-ring chronologies associated with this site. Since the “Periodic SST” RCs are used under both SST-TR models, it is not specified under a particular SST-TR methodology (Table 2.3). From Table 2.3, it is very clear that SST-TR(all) explain more variability of the predictors as opposed to the variance explained by the SST-TR (sel). This is to be expected since the SST-TR(sel) considers only the reconstructed components that are significantly correlated with the periodic component of streamflow. A large embedding dimension had to be chosen for the length of each tree ring (see Table 2.2)
making the variance explained by a single RC small. The stepwise regression only selected a relatively small number of RCs under SST-TR (sel), but they were significantly correlated with the non-periodic component of streamflow. In the next section, we evaluate the performance of these three streamflow reconstruction models (TR-PCA, SST-TR (all) and SST-TR (sel)) in predicting the observed streamflow through rigorous cross-validation. The selected eight stations span across three states, FL, GA and SC, over the Southeast United States.

2.4: Results and Discussion

The primary idea behind the proposed reconstruction methodology is to improve the estimation of high flows in the reconstructed flows for the eight selected stations over the Southeast United States. Studies have used evaluation criteria, e.g., correlation, reduction of error and coefficient of efficiency, for comparing the performance between observed streamflow and reconstructed streamflow in the tree-ring literature [Fritts, 1976, Cook et al. 1999, Devineni et al. 2013]. Reduction of error and coefficient of efficiency are similar to the coefficient of determination ($R^2$) with the former using the climatology of the observed streamflow and the later using the mean of the observed streamflow used for validation for calculating the total variance in the observed flow. For additional details, see Cook et al. 1999. In this study, the performance of three reconstruction models was compared using Pearson correlation coefficient and Normalized Root Mean Square Error (NRMSE) in predicting the observed streamflow. Correlation measures the linear dependency between the observed ($Q_t$) annual streamflow and reconstructed streamflow ($\hat{Q}_t$) with values ranging
between -1.0 to 1.0. NRMSE is simply the ratio of the root mean square error (RMSE) between the observed and reconstructed flow to the mean annual flow ($\bar{Q}$) (Equation 2.1):

$$NRMSE = \frac{RMSE}{\bar{Q}} \quad (2.1)$$

where $RMSE = \sqrt{\frac{\sum (\hat{Q}_i - Q_i)^2}{n}}$ with $n$ denoting the total years of observed streamflow (Table 2.1) at each site. The advantage in using NRMSE helps in comparing the performance of three different models across the eight sites.

We first compare the performance of TR-PCA with SST-TR (sel) in predicting the observed streamflow (expressed in ft$^3$/sec, cfs) for two different sites (Site 2 and Site 5) under below-normal, normal and above-normal flow categories (Figure 2.3). We obtain the observed flows categories by grouping them into below-normal, normal and above-normal categories if the observed flow falls within <33$^{rd}$ percentile, 33-67$^{th}$ percentiles and >67$^{th}$ percentiles of the observed flows respectively. We are not plotting the reconstructed streamflow from SST-TR (all) model since the estimated streamflow overlapped with the SST-TR (sel) and TR-PCA models for many years. The correlations between the observed streamflow and predicted streamflow for site 2 and site 5 for the TR-PCA (SST-TR (sel)) model are 0.59 (0.66) and 0.48(0.58) respectively indicating the improved performance of SST-TR model. Similarly, the NRMSE between the observed streamflow and predicted streamflow for site 2 and site 5 for the TR-PCA (SST-TR (sel)) model are 0.44 (0.41) and 0.31(0.28) respectively. It is important to note the reduction in NRMSE between the two models in predicting the observed flows are expressed as a ratio to the mean annual
streamflow observed at each respective site (see Table 2.1 for values). Thus, the proposed methodology reduces the error in reconstructing the flow by around 3% of their respective mean annual flows.

Apart from the improved performance of SST-TR model, Figure 2.3 also reveals that the estimated streamflow from SST-TR (sel) reduces the error in predicting the observed streamflow during above-normal flow conditions. This could be better understood since many points from SST-TR (sel) lie closer to the 1:1 line in comparison to the TR-PCA model. The NRMSE for the TR-PCA (SST-TR (sel)) model during above-normal flow conditions for site 2 and site 5 are 0.57 (0.53) and 0.39 (0.30) respectively. This clearly indicates the improved performance of SST-TR model in predicting above-normal flow conditions results in 4-8% error reduction in relation to the respective site’s mean annual streamflow. Comparing the performance during below-normal flow conditions shows that NRMSE for the TR-PCA (SST-TR (sel)) model for site 2 and site 5 are 0.41 (0.37) and 0.32 (0.35) respectively. Even though the proposed SST-TR (sel) resulted in improved performance during high-flow conditions, it resulted in reduced performance during below-normal conditions for site 5. In fact, we observed increased NRMSE in reconstructing below-normal flows for sites 5-8. To address this limitation, we considered multi-model combination methodology that combines the reconstructed streamflow from these three different models (TR-PCA, SST-TR (all) and SST-TR (sel)) based on their ability to predict the flow under a particular flow conditions.
2.4.1: Multi-model Combination

Recent studies on multimodel combinations clearly show that combining multiple streamflow predictions conditioned on the predictor/flow conditions result in overall improved performance (Devineni et al., 2008, Li and Sankarasubramanian, 2012). The idea of the combination methodology is to assess the performance of candidate models under a given flow conditions and assign higher weights for the model that performs the best under similar flow conditions. The presented methodology is similar to the approach pursued by Li and Sankarasubramanian [2012] for combining multiple watershed models. We apply this methodology for combining the streamflow from three models (TR-PCA, SST-TR(all), SST-TR(sel) to develop an improved reconstructed streamflow.

Given the observed streamflow \( Q_t \) and predicted streamflows, \( \hat{Q}_t^m \), where \( m=1,2,3,...,M \) denoting a given model over the given time period, \( t=1,2,3,...,n \), we can estimate the error in reconstruction (Equation 2.2), \( \epsilon_t^m \), for each time step \( t \) over the observed period.

\[
\epsilon_t^m = (\hat{Q}_t^m - Q_t)^2 \quad (2.2)
\]

Given the reconstructed streamflow, \( \hat{Q}_t^m \), with \( i \) denoting the year for which we are interested in obtaining the multimodel estimate of reconstructed estimate, we obtain similar reconstructed flow conditions for each model by identifying ‘\( K \)’ neighbors based on simple Euclidean distance. It is important to note that \( K \) nearest neighbors are identified for each model over the period for which we have instrumental record is available. Then, using Equation 2.3, we evaluate the performance of model by comparing with the observed over
the on the identified ‘K’ similar years by computing mean square error, $\lambda_{i,K}^m$ in $\hat{\bar{Q}}_i^m$ over ‘K’ nearest neighbors.

$$\lambda_{i,K}^m = \frac{1}{K} \sum_{j=1}^{K} e_{i(j)}^m$$  \hspace{1cm} (2.3)$$

The subscript $j$ denotes the identified neighbors based on the reconstructed flow, $\hat{Q}_i^m$, for model $m$ during the observed period. For obtaining $K$ neighbors if $\hat{Q}_i^m$ is in the observed period, then the model leaves out the error, $e_{i}^m$, for the conditioning value ($\hat{Q}_i^m$). We considered $K=10$ neighbors for all the sites. The selection of nearest neighbors $K$ has been addressed in detail in the semi-parametric and nonparametric statistics literature. It has been shown that the optimum number of neighbors is equal to the square root of the number of the points available for identifying the neighbors ($n^{0.5}$) as $K$ approaches infinity [Fukunaga, 1990]. Weight terms, $w_{i,K}^m$, for each model are calculated based on the inverse of $\lambda_{i,K}^m$ indicating that higher weights for models that provide lower mean square error over the conditioning neighborhood.

$$w_{i,K}^m = \frac{(\lambda_{i,K}^m)^{-1}}{\sum_{m=1}^{M}(\lambda_{i,K}^m)^{-1}}$$  \hspace{1cm} (2.4)$$

The weights of each model and the model predictions are then used to obtain the combined model prediction $\hat{Q}_i^c$.

$$\hat{Q}_i^c = \sum_{m=1}^{M} w_{i,K}^m * \hat{Q}_i^m$$  \hspace{1cm} (2.5)$$

Thus, the combination methodology provides higher weight for the model that performs better under similar flow conditions. This results in a total of four streamflow reconstruction
models (TR-PCA, SST-TR (sel), SST-TR(all), combined). The performance of these four models are evaluated by first predicting the observed flow and then by rigorous cross-validation for the eight selected sites.

Figure 2.3 also shows the performance of the combined model with the TR-PCA model and the SST-TR(sel) model. The NRMSEs for site 2 and site 5 under the combined (TR-PCA) model are 0.406 and 0.278 (0.442 and 0.306) respectively. Particularly for site 5, the NRMSEs in predicting the below-normal and above-normal flows for the combined (TR-PCA) model are 0.314 and 0.338 (0.322 and 0.397) respectively indicating the improved performance of the combined model. This could be inferred from Figure 2.3 that in most situations that the performance of the combined model is much closer to the 1:1 line. The main advantage of combination approach is that it evaluates the individual model performance under similar reconstructed flows and assign higher weights for the model performing better under similar conditions. We next discuss in detail the performance of the combination model over the eight selected sites.
Figure 2.4: Performance of four different reconstruction models in terms of (a) correlation and (b) NRMSE in predicting the observed annual streamflow for the eight selected sites.
Figure 2.5: Performance (NRMSE) of four different streamflow reconstruction models in predicting the observed streamflow during (a) above-normal flow conditions and (b) below-normal flow conditions over the eight selected sites.
Figure 2.4 shows the performance of four streamflow reconstruction models in predicting the observed streamflow in terms of correlation (4a) and NRMSE (4b) over the selected eight sites. The SST-TR (sel) model had a higher correlation than the traditional approach of TR-PCA in every site except sites 7 and 8 (Figure 2.4). SST-TR (sel) model also outperformed the SST-TR (all) which is expected since the non-periodic tree-ring RCs were selected based on ability to capture the variability on the non-periodic component of annual streamflow. This improved performance could be seen under both correlation and NRMSE. SST-TR (all) model performs better than the TR-PCA model only in sites 4 and 5. Since we don’t employ any stepwise regression in selecting the non-periodic tree ring RCs, it is expected SST-TR (all) to perform poorly. On the other hand, both TR-PCA and SST-TR (sel) employs stepwise regressions in selecting the PCs or RCs. The most interesting information from Figure 2.4 is that the proposed combination approach performs better than the three models in all the sites except site 4 in terms of correlation. In terms of NRMSE, the combination model has the lowest value in almost all the sites except site 4. Thus, combining different models based on their performance over the observed period improves the reconstructed streamflow estimates. As we mentioned earlier, for the observed period, the combined model did not consider the conditioning time step for calculating the MSE over $K$ neighbors.

Figure 2.5 compares the performance of the four models in estimating the streamflow during the above-normal and below-normal flow conditions based on NRMSE. Thus, the RMSE is primarily calculated under the respective categories of observed flows and then normalized by the mean annual streamflow recorded at the site. Based on Figure 2.5a, it is
clear that both SST-TR models perform better than TR-PCA model in predicting the above-normal flows across all the sites except site 7. SST-TR (sel) performs better than the three models in all the sites except site 7. The performance of combination model is better than the SST-TR(all) model in almost all the sites which indicates that the combined model captures the superior performance of SST-TR(sel) model in predicting above-normal flows. Analyses from Figure 2.5a clearly show that our approach of developing a hybrid model considering SST and tree ring certainly has resulted in improved estimates of above-normal flows. Figure 2.5b shows the performance of four models in predicting the below-normal flow conditions. Inclusion of SSTs explicitly improved the estimation of below normal flows only for site 2. Overall, the combination model performed better followed by TR-PCA model. Thus, incorporation of SSTs did not result in significant improvement in predicting below-normal flow conditions as it did with above-normal flows.

2.4.2 Performance under cross-validation

Validation of reconstructed streamflow is difficult since we don’t have observed streamflow for the long period over which we are extending the record. Given that all the comparison of reconstruction models (in Section 2.4.1) is primarily within the observed period, we perform leave-five-percent-out cross-validation in this section. This will help us to evaluate the performance of reconstruction models for periods that were not used for model fitting. For this purpose, we employ leave-five-percent-out cross-validation for evaluating the performance of the three reconstruction models and the combined model estimates. Detailed steps for performing of leave-five-percent-out cross-validation are as follows: For a given
year, $t_i$, the predictors (tree-ring chronologies and SST) and predictand (streamflow) in year $t_i$ and a random five percent of predictors and predictand over the remaining years in the observed period ($t=1, 2, \ldots, n$) were first removed from the data set. The entire methodology in Figure 2.2 (discussed in Sections 2.3.1 and 2.3.2 was performed over the remaining data points of predictors and predictand by fitting TR-PCA, SST-TR(all) and SST-TR(sel) models. Using the left-out observation for the year, $t_i$, we estimate the annual streamflow estimate, $\hat{Q}^{\mu}_{t_i}$ based on the fitted three reconstruction models. By using the estimates of $\hat{Q}^{\mu}_{t_i}$, we also obtain the annual streamflow for the left-out year $t_i$ based on the combination methodology. This procedure was repeated over all the $N$ years of observation to obtain the reconstructed flow estimates. The performance of the four models was compared based on correlation and NRMSE by comparing estimates of $\hat{Q}^{\mu}_{t_i}$ and the combined model with the actual observed flow over the entire observed period. This entire procedure over validation was repeated 50 times to develop 50 sets of correlation and NRMSE for evaluating the four model performance under validation.
Figure 2.6: Performance of four different reconstruction models in terms of (a) correlation and (b) NRMSE in predicting the observed annual streamflow under leave-five-percent-out cross-validation for the eight selected sites.
Figure 2.7: Performance of four different reconstruction models in terms of NRMSE in predicting the observed annual streamflow during (a) above-normal flow conditions and (b) below-normal flow conditions based on under leave-five-percent-out cross-validation for the eight selected sites.
Figure 2.6 shows the box plots of correlation and NRMSE from cross-validation where the numbers above each box plot indicate the number of times each method performed the best (highest correlation or lowest NRMSE) out of 50 trials for that site. The main advantage in analyzing the performance of these models under cross-validation is in its ability in predicting the annual streamflow outside the data used for model fitting. It is reasonable to expect that if a particular approach performs the best under cross-validation, the model could be expected to perform well under reconstruction period too. Based on Figure 2.6, the combination model performs the best in almost all sites except sites 1 and 4 under which SST-TR (sel) performed better. As we inferred in predicting the observed flow, SST-TR (sel) model outperforms the TR-PCA model in all the sites except sites 7 and 8 and also had higher skill than SST-TR (all) in every site except Site 7. This confirms the point that the reconstruction skill can be improved substantially through the selection of non-periodic tree ring RCs. Another important point from Figure 2.6 is that the reconstruction skill can be improved by combining models conditioned on the predicted flow. Basically, the proposed model combination approach gives the highest weight for the best-performing model under a given flow conditions resulting in improved annual streamflow reconstruction.

We also evaluate the performance of four models in predicting above-normal and below-normal flow conditions (Figure 2.7) in terms of NRMSE. From Figure 2.7a, SST-TR models have the highest skill in estimating the flows during above-normal years. SST-TR (sel) model performs the best in almost all the sites except site 7 under which SST-TR (all) performs better. Furthermore, the SST-TR (all) model showed higher above normal skill than the TR-PCA model in every site validating the argument made in Section 2.1 that the
inclusion of SSTs as predictors would result in better prediction of higher flow values. However, the ability to predict below-normal flow conditions differed substantially across all the models (Figure 2.7b). TR-PCA model performed the best under sites 7 and 8 with SST-TR (sel) model performing the best under sites 1 and 4. The combination model overall performed the best in the rest of the sites (sites 2,3,5 and 6). Even though the combination model did not perform well in the other sites (1,4, 7 and 8), based on the spread of the box-plot and the number of times it had the lowest NRMSE, we can clearly infer that it performs the best in predicting below-normal flows. Thus, based on Figures 2.6 and 2.7, we infer that combining multiple streamflow reconstruction models result in improved performance over all the years as well as over below-normal flow years. During above-normal flow conditions, the performance of SST-TR model is clearly better than the rest of the models.

Though we overall see SST-TR (sel) model performs better than the TR-PCA model, the improved performance should not result due to addition of more parameters in the SST-TR (sel) model. Since two regressions are included in the SST-TR models (periodic and non-periodic stepwise regressions), SST-TR models include more predictors for regression than TR-PCA. It is well known that the inclusion of predictors would always result in improved model performance even if the predictors are unrelated to the predictand. To determine if the improvement of the SST-TR models over the TR-PCA is not merely due to addition of more predictors, adjusted $R^2, \overline{R^2}$, for each model is calculated under leave-five-percent-out cross-validation (Figure 2.8). The calculation of $\overline{R^2}$ is similar to the calculation of the coefficient of determination ($R^2$), but the addition of predictors is penalized [Ramanathan 1998]. A
model’s $\bar{R}^2$ shows its skill normalized by the number of predictors. The calculation can be seen below for a model with $k$ predictors and length $n$.

$$\bar{R}^2 = 1 - \frac{\sum_t (\hat{Q}_t - Q_t)^2}{\sum_t (Q_t - \bar{Q})^2} \times \frac{n - 1}{n - k - 1} \quad (2.6)$$

![Figure 2.8](image)

Figure 2.8: Adjusted $R^2$ for the TR-PCA and SST-TR models in predicting the observed annual streamflow under leave-five-percent-out cross-validation for the eight selected sites.

Total number of predictors and the variance explained by each model could be inferred from Table 2.3. Typically, TR-PCA model employs one predictor except Site 7 (Table 2.3), whereas SST-TR(all) considers two predictors with one from non-periodic component and one from periodic component and SST-TR(sel) typically considers three
predictors with two representing the non-periodic component and one representing the periodic component. Figure 2.8 shows $\bar{R}^2$ for the TR-PCA and SST-TR models. The combination model is not shown, since it incorporates flow estimates from all the models. TR-PCA model have a higher $\bar{R}^2$ than the SST-TR models only in sites 3 and 8. It is clear from Figures 2.4-2.7 that TR-PCA performs well in site 8. Only in site 3, in which SST-TR (sel) performs poorly based on adjusted $R^2$ which reverses the findings from Figures 2.4-2.7. This poor performance is due to the increase in the number of predictors to four for site 3 under SST-TR (sel) model. Thus, selection of too many non-periodic components under stepwise regression could result in over-fitting and poor performance due to non-parsimonious model. Both SST-TR models perform better in the rest of the sites (1-2, 4-7) with SST-TR (all) performing better than SST-TR (sel) model only in site 7. Thus, the analyses based on adjusted $R^2$ show that SST-TR overall performs better, but selection of too many reconstructed components by the SST-TR (sel) could potentially result in the development of non-parsimonious model. Hence, caution should be employed in adding more reconstructed components representing the non-periodic component for model development.

2.4.3: Discussion

The primary motivation of this study is to develop a methodology that can improve the high flow estimates of reconstructed streamflows obtained using tree-ring chronologies. For this purpose, we considered annual streamflow and tree-ring chronologies available from eight HCDN basins over the Southeast United States. Given the significant influence of
ENSO on the hydroclimatology of the region [Seager et al. 2009, Stahle et al. 2012] during the first half (October to March) of the water year, we considered Nino3.4 as an additional predictor for improving the reconstructed streamflow estimates. The study employed singular spectrum analyses for associating the periodic and non-periodic components of streamflow with ENSO conditions and tree-ring chronologies respectively. Comparing the observed streamflow with the estimated streamflow from the developed SST-TR models showed a significant improvement in above-normal estimates, but overestimates streamflows during below-normal years. To overcome this limitation, we considered a model combination approach that provides higher weights for the model performing well under similar conditions [Li and Sankarasubramanian, 2012]. This resulted in improved streamflow estimates under all the streamflow conditions, since the combination methodology provided higher weights for the model that performs well under a given condition.

Table 2.4: Results from the right-tailed hypothesis test based on Hotelling-Williams test evaluating the differences in correlations between the combined model and the TR-PCA model are greater than zero.

<table>
<thead>
<tr>
<th>Site</th>
<th>TR-PCA correlation</th>
<th>Combined model correlation</th>
<th>Hotelling-Williams Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.55</td>
<td>0.62</td>
<td>1.65</td>
<td>0.0524</td>
</tr>
<tr>
<td>2</td>
<td>0.59</td>
<td>0.68</td>
<td>2.52</td>
<td>0.0073</td>
</tr>
<tr>
<td>3</td>
<td>0.49</td>
<td>0.66</td>
<td>2.69</td>
<td>0.0051</td>
</tr>
<tr>
<td>4</td>
<td>0.44</td>
<td>0.64</td>
<td>3.72</td>
<td>0.0002</td>
</tr>
<tr>
<td>5</td>
<td>0.48</td>
<td>0.62</td>
<td>2.84</td>
<td>0.0030</td>
</tr>
<tr>
<td>6</td>
<td>0.51</td>
<td>0.64</td>
<td>2.77</td>
<td>0.0037</td>
</tr>
<tr>
<td>7</td>
<td>0.64</td>
<td>0.68</td>
<td>0.72</td>
<td>0.2372</td>
</tr>
<tr>
<td>8</td>
<td>0.50</td>
<td>0.62</td>
<td>2.24</td>
<td>0.0144</td>
</tr>
</tbody>
</table>
It is important to note that the reconstructed streamflow are primarily useful in capturing the past extremes using tree-ring chronologies available before the instrumental period. Given that reconstructed streamflows consider only tree rings (and SST data), the developed model typically explains only around 35-65% variability in observed flows [Woodhouse et al. 2006; Kitzberger et al. 2007, Anderson et al. 2012a, 2012b]. One of the primary challenges in reconstructing streamflow over the Southeast is the limited availability of tree-ring chronologies [Stahle et al. 2012]. Despite this, the presented results (Figures 2.4-2.5) for the combination model showed substantial improvements in reconstructing annual streamflows for the selected eight sites. To ensure that the reported correlation (Table 2.4) resulting from the combined model is statistically significant improvement over the correlation of the TR-PCA model, we performed the Hotelling-Williams test [Bobko, 1995; Devineni and Sankarasubramanian, 2010a], which evaluates the statistical significance of the difference between two dependent correlation estimates. Hotelling-Williams test statistic, 

\[
(r_{12} - r_{13}) \sqrt{\frac{n-1(1+r_{23})}{2(n-1)/(n-3)R + (1-r_{23})^2}}
\]

follows a ‘t’ distribution with (n-3) degrees of freedom where \( r_{12} (r_{13}) \) denotes the correlation between the observed streamflow and the reconstructed streamflow from the combined (TR-PCA) model, \( r_{23} \) denotes the correlation between the reconstructed streamflow between the combined model and the TR-PCA model and ‘n’ denotes the total number of years of observation with

\[
\bar{r} = (r_{12} + r_{13})/2 \quad \text{and} \quad R = (1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}).
\]

Table 2.4 reports the Hotelling-Williams test statistic and the p-values for a right-tailed t-test evaluating the null hypothesis \( r_{12} \geq r_{13} \) being greater than zero. Based on Table 2.4, we clearly infer that the improvements resulting
from the combined model are statistically significant for all the sites except site 7. Given the site 7 requires more than one principal component of the tree rings (see Table 2.3) in developing the regression, additional tree-ring chronologies may be required for capturing the variability in annual flows at that site. We also infer from Figure 2.5 that the combined model provided reduced NRMSE in estimating the above-normal flows for all the sites except site 7. This indicates that combining SST-TR model with the TR-PCA model results in overall improvements in developing reconstructed streamflow.

Though the study demonstrated the advantages of using SST data along with tree-ring chronologies using rigorous cross-validation, application of the SST-TR methodology to additional sites over the Southeast United States could have provided more insights. This was not possible due to the limited availability of tree-ring chronologies in the region in comparison to other parts of the country [Stahle et al. 2012]. Since natural streamflow time series without anthropogenic influence is needed for developing the reconstruction model, we primarily selected stations from the HCDN database that constitute streamflows from virgin watersheds. For reconstructing streamflow for other sites with significant anthropogenic influence (e.g., reservoir storage and groundwater pumping), one could obtain naturalized streamflow from a watershed model for reconstructing annual streamflow if sufficient tree-ring chronologies are available within the watershed. However, using a watershed model could potentially introduce watershed model error in the reconstructed streamflow estimates. Studies have shown that streamflow estimates obtained from multiple watershed models result in improved estimates [Renard et al. 2010, Mahanama et al. 2012, Li and Sankarasubramanian, 2012]. This is another potential area of research in which annual
streamflow from multiple watershed models could be used for reconstructing streamflow for watersheds with significant anthropogenic influence.

Another potential challenge in the broader application of the proposed SST-TR methodology is in the identification of SSTs that influence the annual hydroclimatology of the selected watershed/region. It is well-known that ENSO and Pacific Decadal Oscillation influence the hydroclimatology of the Southern and Northwestern parts of the United States [Ropelewski and Halpert, 1987, Gershunov et al. 1998, Gershunov and Barnett, 1998; McCabe and Dettinger, 1999, Tootle and Piechota, 2006]. These two SST conditions could be considered as predictors for developing reconstructed streamflows over these regions. Perhaps the potential utility of the proposed methodology is in developing countries where limited/no records of streamflow are available (http://www.sage.wisc.edu/riverdata/). Studies in hydrology have extensively focused on Prediction in Ungauged Basins (PUB) considering both process and scale issues [Blöschl et al. 2013]. Given the extensive availability of tree-ring records in many parts of the globe [Grissino-Mayer and Fritts, 1997], utilizing both relevant SST conditions and tree-ring chronologies could provide an opportunity to obtain the streamflow data beyond the period for which meteorological records (primarily precipitation and streamflow) are available. Such efforts not only reduce the uncertainty in annual streamflow estimates, but also could improve the estimation of annual flow duration curves which are critical for water infrastructure design. We intend to consider this as part of our future study in which the utility of reconstructed annual streamflows could be explicitly assessed in improving the water infrastructure sizing.
2.5: Summary and Conclusions

Given the availability of streamflow records only from 1930, studies have considered tree-ring chronologies for extending the observed streamflow records over a longer period of time. The most common approach to reconstruct annual streamflow is to develop statistical regression relationships between principal components of tree rings and observed annual flow values and then extend the relationship to estimate annual streamflow values over the period for which tree ring data is available. The primary limitation of this approach is in estimating high flow values since tree ring growth reaches its potential limit during wet years. To overcome this limitation, we considered SST data over the tropical Pacific Ocean as an additional predictor for reconstructing the streamflow using SSA in eight virgin watersheds over the Southeast United States. Addition of SSTs to tree-ring chronologies in annual streamflow reconstruction improved the streamflow estimates during high flow years, but also resulted in overestimation of low flow values. Given this, the study combined streamflow estimates from multiple models – the traditional approach (TR-PCA) and the hybrid approach (SST-TR) – for reconstructing annual streamflow values. Since the model combination approach provides higher weights for the model that performs better during similar flow conditions, reconstructed streamflow using the combined model provided improved estimates under all the (below-normal, normal and above-normal) flow categories. This improved skill resulting from the hybrid approach of using tree-ring chronologies and SST data could potentially add value in utilizing both instrumental records and reconstructed streamflows for reducing uncertainties in annual streamflow estimates.
Chapter 3: Improved Reservoir Sizing Utilizing Observed and Reconstructed Streamflow within a Bayesian Framework

3.1: Introduction

One of the main goals of reservoirs is to supply water for specified uses even under extreme drought and to protect the downstream under severe floods. Typically, reservoir sizing requires estimation of storage for the design yield of specified reliability. The required storage is often estimated using the Sequent Peak Algorithm (SQP) using either observed streamflow or synthetic streamflow that preserves observed streamflow characteristics [Thomas et al. 1963, Vogel et al. 1988]. The SQP is a sequential process in which the maximum storage for a given trace of streamflow is obtained. This is most beneficial for over-year reservoirs where the reservoir is not designed to refill every year, but rather the storage of the reservoir is carried over to the next year, as opposed to within-year reservoirs which refill every year [Vogel et al. 1987, 1988]. For this reason, the SQP is most often used with annual inflow time series. One limitation of this approach is only one storage estimate can be obtained from an observed streamflow trace. To overcome this limitation, previous studies have employed synthetic streamflow traces so that the probability of meeting a given demand can be calculated [Stedinger et al. 1982, Vogel et al. 1987, Vogel et al 1988]. Most commonly, a lognormal model is used to generate synthetic streamflow traces since it ensures streamflow traces to be positive [Vogel et al. 1988]. Typically, multiple synthetic streamflow realizations having the observed length of annual streamflows are generated and a distribution of storage estimates are obtained using SQP.
Although the synthetic streamflow generation relates storage distribution with reliability of supply, the storage estimates for the desired reliability are derived only using observed streamflow characteristics. Multiple traces generated by stochastic streamflow generation models only address the sampling variability of storage estimates and they are not representative of extremes outside the envelope of instrumental only record. Further, in the continental US, observed streamflow is limited to around 80 years of data, which does not guarantee the long term drought record is represented.

For robust design and planning of water supply systems, information related to prehistoric events could be very useful in estimating the system reliability. Studies have used paleo-proxies such as tree rings to extend the streamflow record beyond the observational period and such reconstructed series insight into the pre-observational drought record could be used to reduce the uncertainty of reliability estimates of reservoir yields [Cook et al. 1983; Gangopadhyay et al. 2009; Woodhouse et al. 2001, 2006, 2010; Devineni et al. 2013]. Studies have used paleo-flood records with observed annual maximum events for improving design flood estimates [Hosking 1986; Stedinger 1986]. However, the values of reconstructed streamflow using tree rings under estimates high flow events, meaning there is great uncertainty in the values of reconstructed streamflow, which obviously limits their use in reservoir design [Prairie et al. 2008]. This limitation can be addressed by including Sea Surface Temperature (SST) data such as El Nino Southern Oscillation (ENSO) or Pacific Decadal Oscillation (PDO) as streamflow reconstruction predictors as in Chapter 2.

The primary motivation in reconstructing the streamflow time series is to incorporate past drought information for improving the design. Despite the high skill of reconstructed
streamflow, the values in the reconstructed time series typically underestimate the observed variance. Studies have used Bayesian frameworks have been used to combine observed and reconstructed data in the generation of hydrologic time series [Vicens et al 1975, Valdes et al. 1977]. One way to reduce the uncertainty in streamflow reconstruction is to resample observed flows based on the state of the reconstructions. Reconstructed streamflow tercile categories of flow (above or below normal) have been employed inside a Bayesian framework to generate streamflow of the Colorado River better representative of the overall drought record [Praire et al 2008]. Still, this approach used only reconstructed streamflow to generate above/below normal states. The generated streamflow were obtained using observed streamflow records, conditional on the generated tercile categories from reconstructed flows. Henley et al. [2011] used a hierarchical Bayesian framework to generate seasonal rainfall using observed and reconstructed inter decadal Pacific Oscillation and pacific decadal oscillation indices. Like in Praire et al [2008], the reconstructed data was used to generate the state of the precipitation (wet/dry), and observed precipitation was used to generate the values.

The goal of this study is to investigate the utility of reconstructed streamflow in reducing the uncertainty in stochastic streamflow generation and estimating reservoir storage estimates. We employ stochastic streamflow generation models, but unlike previous studies, reconstructed streamflow estimated values will be combined with the observed streamflow Bayesian framework to obtain distributions of streamflow parameters. The synthetic streamflow traces are used with the SQP to obtain storage distributions conditional on both observed and reconstructed streamflow. This approach is tested through a split-sample
validation where the true storage distribution is known and can be compared to the storage
distribution obtained using this approach. The storage distribution of the traditional approach
will be obtained from the SQP and generated streamflow traces using characteristics from the
observed. The performance of the proposed methodology will compared to the traditional
approach to test the argument that the inclusion of reconstructed streamflow will reduce the
uncertainty of reservoir sizing estimation.

The chapter is organized as follows: Section 3.2 provides information on the data
related to streamflow, tree-ring chronologies, and SST used for streamflow reconstruction.
Following that, Section 3.3 describes the streamflow generation procedure as well as the
reservoir sizing estimation. The Bayesian framework and experimental design is also detailed
in this section. Section 3.4 provides the results of the experiments testing whether the
addition of reconstructed streamflow reduces uncertainty in reservoir sizing. Finally, Sections
3.5 and 3.6 summarizes the salient findings and conclusions along with the potential of the
proposed methodology.

3.2: Data

The goal of this study is to improve reservoir sizing estimated by combing observed
and reconstructed streamflow. Since the streamflow records are being reconstructed, only
undeveloped basins from the Hydro-Climatic Data Network (HCDN) [Slack et al. 1993].
Streamflow sites in the HCDN are nearly void of upstream storage and groundwater
pumping. Since the reconstructed streamflow will be calculated using tree ring chronologies,
which are annually, annual streamflow based on water year is used. For this study,
streamflow was reconstructed as in Chapter 2, so the chosen streamflow sites are greatly influence by El Nino Southern Oscillation (ENSO). The locations of the selected streamflow sites, the rivers they are located on, and the tree ring chronologies used for reconstruction can be seen in Figure 3.1. The characteristics for each are summarized in Table 3.1.

Table 3.1: List of USGS Streamgage Stations and the corresponding tree ring chronologies used for annual streamflow reconstruction

<table>
<thead>
<tr>
<th>Site Number</th>
<th>USGS Site Number</th>
<th>Average Annual Streamflow [cfs]</th>
<th>Drainage Area [mi²]</th>
<th>Tree Chronologies Used [ITRDB Number]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>02132000</td>
<td>1009</td>
<td>1030</td>
<td>SC004, SC006, NC008</td>
</tr>
<tr>
<td>2</td>
<td>02236000</td>
<td>3055</td>
<td>3066</td>
<td>FL007, FL008</td>
</tr>
</tbody>
</table>

The tree ring chronologies used for reconstruction were obtained from the National Atmospheric and Oceanic Administration (NOAA) International Tree Ring Data Bank (ITRDB) [available online at http://www.ncdc.noaa.gov/paleo/treering.html]. The tree rings in Figure 3.1 were used along with the Nino-3.4 index (an index indicating the ENSO condition) to reconstruct streamflow using the Combined model from Chapter 2. All reconstructed streamflow in this study was obtained using this process.

3.3: Methodology: Bayesian Framework and Experimental Design

3.3.1: Sequent Peak Algorithm (SQP)

As discussed in Section 3.1, the maximum storage required from an inflow time series can be calculated using the SQP which can be seen in Equations 3.1 and 3.2 [Thomas et al. 1963].
\[ S = \text{Max}(S_t) \quad (3.1) \]

\[ S_t = \begin{cases} 
S_{t-1} + D + E - Q_t & \text{if positive} \\
0 & \text{otherwise} 
\end{cases} \quad t = 1 \ldots n \quad (3.2) \]

Figure 3.1: Location of USGS streamflow sites and the tree ring chronologies used for reconstruction in this study.

The storage required at time \( t \) is the sum of the storage required at the previous time step plus the water supply demand \((D)\) and evaporation \((E)\) minus the incoming streamflow at time \( t \). If the inflow is greater than the previous storage, demand and evaporation, the required storage at time \( t \) is zero. The storage required for the entire trace \((t = 1 \ldots n)\), where \( n \)
is the length of the streamflow record) is found by taking the maximum of the storage over the entire trace. The initial storage \(S_0\) is set to zero, which could result in an underestimation of storage if low flows occur at the beginning of time series due to the iterative process in the SQP. It is important to note that the streamflow time series can be any time scale, but the SQP is used for over-year reservoirs, so annual streamflow is most common and is the temporal scale used in this study.

3.3.2: Streamflow Generation

The SQP results in one required storage per inflow time series, but as outlined in Section 3.1, a distribution of reservoir storages is desired so the reliability yield can be calculated. This distribution can be calculated by generating many inflow time series using characteristics from the observed inflow [Stedinger et al. 1982, Vogel et al. 1987, Vogel et al. 1988]. To account for the positive constraint on streamflow, the lognormal streamflow generation (Equation 3.3) is most often used.

\[
Y_{t+1} = \mu + \rho(Y_t - \mu) + \varepsilon_t \sigma(1 - \rho^2)^{1/2} \quad t = 1 \ldots n \tag{3.3}
\]

Using this approach, a log-transformed streamflow time series, \(Y_t\), of length \(n\), can be generated using the observed mean, \(\mu\), standard deviation, \(\sigma\), and lag-one autocorrelation, \(\rho\), of the observed log-transformed streamflow.

3.3.3: Bayesian Framework

As discussed in Section 3.1, streamflow is generated using the lognormal model (Equation 3.3) parameters whose distributions will be obtained using a Bayesian framework, by combining observed and reconstructed annual streamflow. Bayesian frameworks have
been used to generate streamflow on an annual scale due to the ability to incorporate model uncertainty [Vicens et al 1975, Valdes et al. 1977]. Bayes theorem obtains posterior probability density functions (PDFs) of parameters, \( \theta \), for a given data set, \( Y \), which consists of both observed and reconstructed streamflow datasets. The Bayes formulation could be written as:

\[
P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \quad (3.4).
\]

\( Y \) denotes the log transformed streamflow data set, and \( \theta \) denotes the parameters of interest. \( P(\theta|Y) \) is the the posterior distribution of parameters \( \theta \) given the data set \( Y \). The likelihood function is \( P(Y|\theta) \), and the prior distribution is \( P(\theta) \). The prior distribution is a PDF representing previous knowledge on the parameters. It can represent a strong belief, or it can be uninformative where each parameters has equal probability of occurrence. The marginal likelihood, \( P(Y) \), is a constant, so Bayes theorem is often written as (Equation 3.5)

\[
P(\theta|Y) \propto P(Y|\theta)P(\theta) \quad (3.5).
\]

The goal of this Bayesian framework is to obtain parameters of the lognormal distribution considering both observed and reconstructed streamflow.

**Step A: Calculate the likelihood**

The Bayesian framework is designed to obtain posterior distributions of lognormal streamflow parameters, so the observed and reconstructed streamflow are transformed into the log plane using Equation 3.6.

\[
Y_i = \log(Q_i) \quad (3.6)
\]
Q and Y represent the natural and log transformed streamflow respectively and assumes that the log-transformed flows follow the normal distribution (Equation 3.7).

\[
P(Y_t|\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{Y_t - \mu}{\sigma} \right)^2 \right] \quad (3.7)
\]

If the data points Y are independent, then the likelihood function for the data set can be found by multiplying the likelihood of each individual y. Since streamflow often has a high year to year dependence, the joint density function is used for n correlated normal points. Normal variables could be estimated using Equation 3.8 where Σ denotes covariance matrix.

\[
P(Y|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(\Sigma)}} \exp \left[ -\frac{1}{2} [Y_1 - \mu_1, \ldots, Y_n - \mu_n] \Sigma^{-1} [Y_1 - \mu_1, \ldots, Y_n - \mu_n]^T \right] \quad (3.8)
\]

The parameter ‘μ_t’ is equal to the mean of the observed log transformed flows. Σ, the covariance matrix, will be obtained from the variance, σ^2, and lag one correlation, ρ, of the log transformed streamflow (Equation 3.9).

\[
\Sigma = \begin{bmatrix}
\sigma^2 & \rho \sigma^2 & \rho^2 \sigma^2 & \rho^3 \sigma^2 & \cdots & \rho^{n-1} \sigma^2 \\
\rho \sigma^2 & \sigma^2 & \rho \sigma^2 & \rho^2 \sigma^2 & \cdots & \rho^{n-2} \sigma^2 \\
\rho^2 \sigma^2 & \rho \sigma^2 & \sigma^2 & \rho \sigma^2 & \cdots & \rho^{n-3} \sigma^2 \\
\rho^3 \sigma^2 & \rho^2 \sigma^2 & \rho \sigma^2 & \sigma^2 & \cdots & \rho^{n-4} \sigma^2 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \rho^{n-2} \sigma^2 \\
\rho^{n-1} \sigma^2 & \rho^{n-2} \sigma^2 & \rho^{n-3} \sigma^2 & \rho^{n-4} \sigma^2 & \cdots & \sigma^2
\end{bmatrix} \quad (3.9)
\]

It is important to note that we are only considering the lag-one autocorrelation since the streamflow beyond two years are not exhibiting any dependence. Equations 3.8 and 3.9 are to calculate the likelihood of the streamflow parameters μ, σ, and ρ on the observed and
the reconstructed streamflow. Since the observed and reconstructed streamflows are combined, one matrix of \( \mu \) and one covariance matrix will be used to calculate the likelihood of the combined data set.

**Step B: Obtain the Prior Distributions**

The prior distributions of the parameters in this framework are uninformative, indicating each parameter follows a uniform distribution. For the mean, the streamflows have been log-transformed, ensuring they will be positive when brought back to the original space. For this reason, any number can be the mean of the log transformed streamflow, so the prior distribution is a uniform distribution between negative and positive infinity. This makes the probability of all values of the mean a constant which cancel out in the proportionality shown in Equation 3.5.

The standard deviation must be greater than zero, but all values over zero are equally likely. The prior distribution for the standard deviation can be seen below where \( c \) is the probability density for the uniform distribution.

\[
P(\sigma) = \begin{cases} 
  c & \text{if } \sigma > 0 \\
  0 & \text{otherwise} 
\end{cases} \quad (3.10)
\]

The lag-one autocorrelation, by definition, must have a value between negative one and one. All values in this range are equally likely, so its prior distribution is similar to the prior distribution for standard deviation.

\[
P(\rho) = \begin{cases} 
  0.5 & \text{if } |\rho| \leq 1 \\
  0 & \text{otherwise} 
\end{cases} \quad (3.11)
\]
If we assume that the mean, standard deviation and lag-one autocorrelation are independent, their joint prior can be found by multiplying each individual prior distribution (Equation 3.12).

\[ P(\mu, \sigma, \rho) = P(\mu)P(\sigma)P(\rho) = \begin{cases} 
0.5 * c_\sigma & \text{if } \sigma > 0 \text{ and } |\rho| \leq 1 \\
0 & \text{otherwise} 
\end{cases} \]  

(3.12)

*Step C: Combining the likelihood and prior*

Since constants do not affect the proportionality equation, Equation 3.5 can be updated as seen below where \( P(Y|\mu, \sigma, \rho) \) is Equation 3.8.

\[ P(\mu, \sigma, \rho|Y) \propto P(Y|\mu, \sigma, \rho) \begin{cases} 
0.5 & \text{if } \sigma > 0 \text{ and } |\rho| \leq 1 \\
0 & \text{otherwise} 
\end{cases} \]  

(3.13)

\[ P(\mu, \sigma, \rho|Y) \propto \begin{cases} 
P(Y|\mu, \sigma, \rho) & \text{if } \sigma > 0 \text{ and } |\rho| \leq 1 \\
0 & \text{otherwise} 
\end{cases} \]  

(3.14)

*3.3.4: Markov Chain Monte Carlo*

The distribution of streamflow generation parameters is obtained using the Bayesian framework in Equation 3.14 and Markov Chain Monte Carlo (MCMC). MCMC operates by changing the parameters in a Bayesian framework using a proposal distribution. With an initial set of streamflow parameters, the likelihood of the streamflow time series conditional on those streamflow parameters (Equation 3.14) is calculated. New parameters are then generated using a proposal distribution. In this Bayesian model, the proposal distributions considered is an independent normal (Equations 3.15 - 3.17) proposal distribution. The likelihood of the new parameter is calculated using Equation 3.14, and the likelihood ratio is determined by dividing the likelihood of the new parameter set by the likelihood of the previous parameter set. This likelihood ratio is the probability that a given parameter will be
updated during the iteration. MCMC uses the proposal distribution many times, recording the parameter set at the end of each iteration to obtain a distribution of the parameter set.

\[ \mu' \sim N(\mu', \sigma_{\mu}) \quad (3.15) \]
\[ \sigma' \sim N(\sigma, \sigma_{\sigma}) \quad (3.16) \]
\[ \rho' \sim N(\rho, \sigma_{\rho}) \quad (3.17) \]

The proposal distribution updates all three parameters one at a time, calculating the likelihood ratio and updating a given parameter before going to the proposal distribution of the next parameter. The order of the parameter updates is mean, standard deviation and lag-one autocorrelation, with all distributions being normal with a mean of the previous value of a parameter. The standard deviation of each distribution is found by updating until the acceptance rate of a given parameter is between 0.2 and 0.4.

The Bayesian model can be used on any streamflow time series, but the issue of how to combine the reconstructed and observed streamflow needs to be addressed. Since reconstructed streamflow is an estimate, its variance is going to be less than that of the observed streamflow. In this study, the variance of streamflow is being estimated and would be greatly underestimated by the reconstructed streamflow values. To address this issue, random noise with mean equal to zero and variance equal to the unexplained variance from the reconstruction regressions is added to the reconstructed streamflow at the beginning of each MCMC iteration. The reconstructed streamflow with noise is then combined with the observed chronologically, creating one time series. Once the noise is added and the time series are combined, the combined time series can be analyzed using the Bayesian model.
3.3.5: Experimental Design

One of the limitations of using reconstructed streamflow in reservoir sizing is that their values are unknown. A split-sample validation was designed to test the validity of using reconstructed streamflow for reservoir sizing (Figure 3.2). An example would be a site where there are 70 years where both streamflow and tree ring chronology records are available. The true distribution of the storages (white) can be found for the observed 70 years based on the obtaining streamflow parameter distributions from the Bayesian Model and then using them to obtain storages using the SQP with the synthetic streamflow traces. To obtain a synthetic streamflow trace, a set of mean, standard deviation and autocorrelation is selected and a streamflow trace of 70 years is generated. The SQP is then used on this time series to get a storage value. This process is repeated many times, resulting in a true distribution of storages.

To split the sample, the most recent 40 years of streamflow records will be considered as “observed” and the oldest 30 years of records will be “unobserved.” The “observed” distribution of storages (dark grey in Figure 3.2) is found through synthetic streamflow traces and with the SQP.
Figure 3.2: Experimental Design for validation the utility of reconstructed streamflow in improving the reservoir design.
To obtain reconstructed streamflow records in the split-sample validation, reconstruction regressions are found between for the 40 years of “observed” data. These regressions are used to estimate the streamflow for the 30 years of “unobserved” data. Similar to the procedure for the “observed” distribution of storages, the “observed” 40 years of streamflow and the “unobserved” 30 years of reconstructed streamflow will be combined using the Bayesian Model (Figure 3.2) to obtain posterior distributions of lognormal parameters, which will be subsequently used with the lognormal streamflow generation scheme and the SQP to obtain the combined distribution of storages (light grey Figure 3.2).

3.4: Results

The streamflow generation parameters and required storage estimates were obtained using the methodology outlined in Section 3.3. For each site, the demand used to calculate the required storage was 80% of the mean annual streamflow which can be found in Table 3.2. The quantiles of the required storages were calculated from the calculated storages from 10,000 generated streamflow traces. In each experiment, the length of each generated streamflow trace was the length of the “True” streamflow record.
Table 3.2: Time periods and the reconstruction skill for each site used for the split-sample validation and the application on reservoir storage

<table>
<thead>
<tr>
<th>Site</th>
<th>Mean Annual Streamflow [cfs]</th>
<th>Split-sample Validation</th>
<th>Reservoir Storage Application</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>“Observed Period”</td>
<td>“Unobserved Period”</td>
</tr>
<tr>
<td>1</td>
<td>1035</td>
<td>1946 – 1985</td>
<td>1930 – 1945</td>
</tr>
<tr>
<td>2</td>
<td>3057</td>
<td>1964 – 2003</td>
<td>1934 – 1963</td>
</tr>
</tbody>
</table>
Figure 3.3: Probability Density Functions of the likelihood ratio of the combined streamflow reconstruction parameters versus the observed streamflow generation parameters with varying reconstruction length. The mass above and below 1.0 are labeled.
3.4.1: Split-sample Validation

The “observed” and “unobserved” periods during split-sample validation for the sites in this study can be seen in Table 3.2. For each site, the reconstruction model was developed using 40 years of observed streamflow, and the remaining “unobserved” years of streamflow were reconstructed. The correlation between the “observed” and estimated streamflow was calculated to determine the variance of the noise needed to preserve the streamflow variance in MCMC.

During split-sample validation, posterior distributions using observed and reconstructed flows were obtained before storage estimation. Since the likelihood of a set of streamflow generation parameters can be calculated as in Section 3.3.3, likelihood distributions can be obtained by calculating the likelihood of each parameter set in the distribution with the True streamflow record. The likelihood ratio of the distribution using observed streamflow and the distribution using combined observed and reconstructed streamflow can also be calculated using the posterior distributions. This ratio being below or above 1.0 will show whether the observed streamflow or the combined observed and reconstructed streamflow produce streamflow generation parameter distributions more representative of the True streamflow respectively. In the distribution of this likelihood ratio (Figure 3.3), there is more mass above 1.0 meaning the streamflow generation parameters from the combined observed and reconstructed streamflow are more representative of the True streamflow than the parameters from the observed record. Furthermore, the mass above 1.0 increases as the length of the reconstructed record increases for each site. This result
illustrates the argument that combining observed and reconstructed streamflow will reduce uncertainty in stochastic streamflow generation.

These streamflow generation parameters were then used to estimate the required storage using the procedure outlined in Section 3.3.3. Storage estimates obtained by using both the observed and reconstructed streamflow are closer to the estimates from the true streamflow than the estimates from the observed streamflow, especially in the upper quantiles (Figure 3.4). The improvement in the upper quantiles is important because reservoirs are often sized according the 95th quantile, or the required size so that the assumed demand is the 95% reliability yield, meaning the combination of observed and reconstructed streamflow yields a better storage estimate than using only observed streamflow.

Longer reconstruction lengths also produce better storage estimates (Figure 3.4). In Sites 2 and 3, the deviation from the True storage distribution decreases as the length of the used reconstructed streamflow increases for each quantile. This illustrates the argument made in Section 3.1 that including reconstructed streamflow in reservoir design will improve storage estimation, and furthermore, it demonstrates that longer reconstructed streamflow records will result in improved storage estimates. It is important to note that the storage estimate in Site 1 does not improve as the reconstruction length increases as in Sites 2 and 3. One reason for this is the length of the true streamflow record in Site 1. Site 1 has 56 years of streamflow data, which means only 16 years can be reconstructed since 40 years are used for model regression. In Site 1, the longest streamflow record is only 29% of the total streamflow length. When this is compared to Sites 2 and 3, whose longest streamflow record is 43% of the total streamflow record, it can be seen that Sites 2 and 3 have more opportunity...
to improve the storage estimate as more of their streamflow record is “unobserved.” It is important to note however, that including reconstructed streamflow in Site 1 still improved the storage estimates in the upper quantiles, the importance of which was discussed earlier in this section.

3.4.2: Application: Reservoir Storage Estimates

From the findings in Section 3.4.1, including reconstructed streamflow in reservoir sizing estimation will improve the estimate and this improvement increases as the length of the reconstructed streamflow used increases. Therefore, the best storage estimate obtainable would be to combine the entire observed streamflow record with the entire reconstructed streamflow record. Using the reconstruction models in Chapter 2, reconstructed streamflow is available from 1857. The observed and reconstructed periods for each site can be seen in Table 3.2, and the storage estimates using only the observed period and the combination of the entire reconstructed and observed streamflow records are shown in Figure 3.5. In each site, the required storage found by using the only the observed streamflow is higher than the required storage using both the observed and reconstructed streamflow records. If reservoirs were to be sized according to the 95th quantile using the combination of observed and reconstructed streamflow rather than only the observed, the required reservoir size for Sites 1, 2 and 3 would decrease 0.82, 0.70 and 1.06 million acre-feet respectively. To put the size difference in perspective, the volume of 1 million acre-feet is the amount of water 1 foot high over an area of 1 million acres. This means that the size decrease seen in Figure 3.5 will result in a much smaller area being affected by a reservoir.
Figure 3.4: Difference between true storage and estimated storages using different reconstructed streamflow lengths under split-sample validation for the selected sites. Demand is assumed to be 80% of the mean annual streamflow for respective sites.
Figure 3.5: Storage Distribution for the selected sites obtained by the Bayesian combination of observed and reconstructed flow.
It is important to consider how the reservoir sizing would be affected if the reconstruction skill were not as high or if not as many reconstructed years were available. Figure 3.5 shows the reservoir sizing difference using different reconstructed lengths and skills from using the entire observed and reconstructed streamflow set, where the reservoir was sized at the 95th quantile of the storage distribution.

For sites 1 and 3, the storage estimate approaches the true storage estimate as the reconstructed years and reconstruction skill increases (Figure 3.6) which is expected since the true storage estimate is obtained using all of the uncorrupted reconstructed years. However, the number of reconstructed years has a larger effect on the storage estimate than the reconstruction skill. One possible reason for this is the way streamflow is being generated using the mean, standard deviation, and lag-1 correlation. Both the standard deviation and lag-1 correlation are preserved in our methodology, so that they are consistent with the observed streamflow, making the mean the parameter which is updated. Therefore, as long as the skill is high enough to capture the long term mean of the streamflow in the pre-observed period, the reservoir storage estimate will improve.

3.4.3: Generalizing the Findings for Different Reconstruction Skill and Length

In the previous section, it was established that the best storage estimate is obtained by combining all of the observed and reconstructed streamflow records. However, the reconstruction skill and reconstruction length varies from site to site. The sensitivity of the reservoir size estimate with varying reconstruction skill and length was tested by corrupting the reconstructed streamflow values and combining different reconstruction lengths. In the
previous section it was discussed how the storage estimate would change with different skills and reconstruction lengths, but that experiment is limited as the upper bound of the reconstruction skill is the actual skill of the reconstruction model. Another experiment was designed so that there was no upper bound on the reconstruction skill. For this experiment, a reconstructed streamflow time series was divided into two categories, the “observed” and the “unobserved,” while the “true streamflow was assumed to be the entire reconstructed streamflow record. Reconstructed streamflow was used due to the longer length of the time series (about 130 years) when compared to observed streamflow. The skill of the reconstructed streamflow was obtained by corrupting the “unobserved” flow values. The process for finding the estimated storage is the same process outlined earlier in this section.

Unlike the previous experiment, this experiment shows the storage estimates improve as both the skill increases and the number of reconstructed years increases (Figure 3.7) especially in Site 1. One reason for this was the upper bound on the skill in the previous experiment, so it could not be seen how the reservoir estimates behaved with very high skill. Although in Site 2 the reconstructed years have a greater effect on the reservoir estimate performance than the reconstruction skill, it is important to note that the best estimate is obtained by using all of the reconstructed years at the highest skill.
Figure 3.6: Difference between the 95\textsuperscript{th} storage quantile obtained using the combined flows with current skill and the corrupted combined flows having different skill and sample length. Contours denote the difference in million acre-feet.
Figure 3.7: Difference between the 95th storage quantile obtained using the combined flows with perfect skill and the corrupted combined flows having different skill and sample length. Contours denote the difference in million acre-feet
3.5: Discussion

In the previous section it was demonstrated that the combination of observed and reconstructed streamflow can improve stochastic streamflow generation and reservoir storage estimation. Furthermore, both the streamflow generation parameters and reservoir storage estimates improve as the reconstruction skill and length increase. This is expected as insight into the pre-observed period should better explain the long-term streamflow, and higher reconstruction skill will provide more insight. However, it is important to look at the feasibility of this methodology being used in reservoir sizing.

In practice, the reconstruction skill cannot be chosen as it is dependent on the quality of the paleo-data available. Although Figures 3.6 and 3.7 show the reservoir sizing estimate will improve if the correlation between reconstructed and observed streamflow is significant, the actual explained variance of the reconstructed streamflow should be considered. Reconstructed streamflow data are estimated values, so careful consideration is necessary when using these values to design a project as large as a reservoir. In our experiments, the reconstructed streamflow explained almost half of the total variance, which was shown by our experiments to be enough to capture the long term mean. However, in other sites with higher annual streamflow variance, a higher skill may be needed. It is our recommendation that a rigorous validation test be done on the reconstructed streamflow model at an individual site before the proposed reservoir sizing methodology be put into practice.

If the reconstruction skill is able to explain enough variance, the length of reconstructed streamflow used can be chosen. It was shown in Section 3.4 that the reservoir sizing estimates improved as more reconstructed years were added (Figures 3.4, 3.6 and 3.7),
but it is important to note that in each experiment, the length of the observed period was longer than the reconstructed period. Even though a longer reconstructed record should better capture the long term mean, standard deviation and lag-one correlation, the reconstructed values are estimates and have uncertainty. By keeping the length of the reconstructed period shorter than the observed period, the observed values without uncertainty have a greater effect in the MCMC likelihood than the reconstructed values. However, a longer reconstructed streamflow record can be utilized without being given a greater effect than the observed streamflow by having two weighted likelihoods inside MCMC. When the proposed methodology is put into practice, we recommend constructing the MCMC in a way so that the observed streamflow has the greatest effect on the likelihood.

The difference in sizing between using only the observed streamflow and the combined reconstructed and observed streamflow should also be considered when using this methodology in practice. When a reservoir is sized and used, the 95% yield will either be above or below the design demand. If the 95% yield is above the design demand, the reservoir sizing was too large, meaning the cost was higher than necessary. However, if the design demand is greater than the 95% yield, the desired demand will be available less than 95% of the time, stressing the water supply system. When the proposed methodology is used in practice, a water manager should consider whether a higher cost or a lesser water supply efficiency would have a greater negative effect on the system. If the difference between storages from the observed and combined observed and reconstructed streamflow is significant, the more important cost should be considered when choosing between the models.
3.6: Conclusion

Given the streamflow records in the United States are limited to about 80 years, the use of observed streamflow in reservoir sizing does not guarantee that the drought record will be accounted for. It is common to generate streamflow stochastically to obtain a distribution of required reservoir storages. Although this approach does improve reservoir sizing, it is still limited since the streamflow generation parameters are derived from the limited observed record. To overcome this limitation, we considered combining the observed streamflow record with reconstructed streamflow from paleo-data on three virgin basins in the southeastern United States. Combining observed and reconstructed streamflow resulted in streamflow generation parameters that were more representative of the True streamflow record than using only observed streamflow in the split-sample validation. The combination of observed and reconstructed streamflow also yielded storage distributions closer than the observed streamflow to the True storage distribution. The skill of both streamflow generation parameters and storage estimates improved as the length of the reconstruction record increased. When the split-sample validation was generalized, the storage estimates improved as the skill and length of the reconstruction record increased. The improved streamflow generation and reservoir storage estimation from the addition of reconstructed streamflow could potentially add value in reservoir design and increase water supply efficiency.
Chapter 4: Reducing Uncertainty in Stochastic Streamflow Generation and Reservoir Sizing by Combining Observed, Reconstructed and Projected Streamflow

4.1: Introduction

Observed streamflow records are often used in management decisions such as reservoir storages for future events. However, due to the sampling variability due to finite record length, multiple streamflow traces are generated using characteristics (often mean, standard deviation, and lag-1 correlation) of the observed streamflow record, thereby providing a distribution of future decisions. For generating multiple series, a lognormal model is commonly used, due to the positive constraint on streamflow. Multiple generated streamflow traces are then used for analyzing and planning future management decisions. For example, the Sequent Peak Algorithm (SQP) can be used to calculate the required storage for an assumed demand and a given trace of the generated inflow time series [Thomas et al. 1963]. Thus, using multiple streamflow traces with the SQP, one can obtain a distribution of required storages [Vogel et al. 1988]. This is useful because the quantiles of the distribution provide the required storage for the desired reliability. Often, this process is commonly used to size a new reservoir or to reallocate the existing storage of a reservoir for the desired reliability yield.

Although synthetic streamflow generation is used to develop storage estimates, the generations are purely based only on the observed period. In the United States, observed streamflow data is only available for about 80 years. Therefore there is no guarantee that the
pre-historic observations encompass pre-historic droughts. Thus, the generated streamflow will also not be representative of the paleo-records in a location. This limits the effectiveness of using only observed streamflow information to generate synthetic streamflow traces.

Stochastic streamflow generation could be improved by adding information related to the pre-historic record. Streamflow records have been extended beyond the observational period using paleo-proxies such as tree rings and such reconstructed series provide insight into the pre-observational drought record, which could be used to reduce the uncertainty of stochastic streamflow generation [Cook et al. 1983; Gangopadhyay et al. 2009; Woodhouse et al. 2001, 2006, 2010; Devineni et al. 2013]. Flood design estimates have been improved by using paleo-flood records by combining them with observed annual maximum events [Hosking 1986; Stedinger 1986]. However, the limitation in using reconstructed streamflow using tree rings is due to the under estimation of high flow events. Therefore, there is great uncertainty in the values of reconstructed streamflow, which obviously limits their use in reservoir design [Prairie et al. 2008]. This limitation was addressed in Chapter 2 by using the hybrid approach consisting of Sea Surface Temperature (SST) and tree rings.

This dissertation research in Chapter 3 showed that reconstructed streamflow can greatly reduce the uncertainty of hydro-climatic generations by combining reconstructed and observed data. Studies have used Bayesian frameworks to combine observed and reconstructed data in the generation of hydrologic time series [Vicens et al 1975, Valdes et al. 1977]. Resampling observed flows based on reconstruction states has also been used to to reduce the uncertainty in streamflow reconstruction. Reconstructed streamflow tercile categories of flow (above or below normal) were used to generate streamflow of the
Colorado River better representative of the overall drought record [Praire et al 2008]. Still, the reconstructed streamflow was only used to generate above/below normal states, while the observed records were used in the actual streamflow generation conditioned on the generated tercile categories from reconstructed flow. Similarly, Henley et al. [2011] used reconstructed inter decadal Pacific Oscillation and pacific decadal oscillation indices to generate the state of the precipitation (wet/dry), and observed precipitation was used to generate the values in a hierarchical Bayesian framework to generate seasonal rainfall.

In Chapter 3 of this study, it was shown that the addition of reconstructed streamflow values reduced the uncertainty of stochastic streamflow generation and over-year reservoir storage estimation. Unlike previous studies, reconstructed streamflow estimates were combined with observed streamflow values in a Bayesian framework yielding streamflow generation parameters that are more representative of the long term streamflow. It was also shown that streamflow traces from the reconstructed and observed streamflow reduced uncertainty in reservoir sizing. While this study shows that reconstructed streamflow can explain long term streamflow characteristics, large projects such as reservoirs have to be effective under future streamflow conditions. Furthermore, under near-term climate change, there is no guarantee that the future streamflow will be representative of the current and past streamflow characteristics. In fact, several studies have analyzed the impact of climate change especially changes in precipitation and streamflow changes at various spatial scales [Gleick 1987, Lettenmaier et al. 1992, Gleick and Chalecki 1999, McCabe and Wolock 1999, Sankarasubramanian et al. 2001, Sinha and Cherkauer, 2010]. Climate change impact on
streamflow also indicates the need to reallocate reservoir storage for designed uses [Lettenmaier et al. 1999, Hanak and Lund 2012].

For reliable water supply planning and management, stochastic streamflow needs to be representative of future streamflow characteristics. Insight into the future climatic conditions under near-term climate change can be gained by using global circulation models (GCMs). The practicality of GCMs in management has been limited due to uncertainty in projections under varying CO₂ emission scenarios [Hawkins and Sutton 2009]. However, under a given emission scenario, GCMs have similar climate projections, over near-term (10 – 30 years) with the majority of the variability due to different model runs [Singh et al. 2014]. This is primarily because at the decadal time scales (10 – 30 years), very little uncertainty in GCM climate scenarios remains owing to “committed warming” primarily arising from the thermal inertia of the oceans [Hawkins and Sutton 2009, Meehl et al. 2009, Singh et al. 2014]. Given water resources management and planning are usually decadal time scales, the increasing predictability of the climate system on these time scales from initialized ocean states could explain future hydroclimatic conditions [Keenlyside et al. 2008, Milly et al. 2008]

To develop streamflow projections using climate projections, it is common to pursue either dynamical downscaling [Leung et al. 1999; Wood et al. 2004] or statistical downscaling [Sankarasubramanian et al 2008; Oh et al. 2012]. In this study we consider the statistically downscaled monthly climate change projections available at 1/8th degrees from the Bureau of Reclamation Site (http://gdo-dcp.ucar.edu/downscaled_cmip_projections/dcpInterface.html).
The skill of projected streamflow for future periods is also dependent on the chosen watershed model. Water balance models, such as the Variable Infiltration Capacity – ‘abcd’ Model (‘abcd’ Model), can be used to obtain the projected streamflow into the future using climate change projections from GCMs [Thomas, 1981]. It is important to note that since GCMs do not explain interannual variability, but rather capture change in the long term mean and standard deviation, the projected streamflow using climate change projections will also only explain the long term mean and standard deviation. Thus, projected streamflow characteristics obtained from GCMs could be used for sizing for future reservoir storages or for supporting reservoir capacity expansion projects.

Given the limitation of observed streamflow in representing future streamflow changes, the goal of this study is to investigate the role of reconstructed and projected streamflow in reducing the uncertainty in streamflow generation. In order to evaluate the utility of climate change projections for reservoir sizing, we plan to employ a split-sample validation where reconstructed and projected streamflow will be obtained and combined with observed streamflow. A Bayesian framework will then be used to obtain distributions of streamflow generation parameters that consist of paleo, observed and future climate change projections. The likelihood of future streamflow in the split-sample validation will then be calculated based on the distributions of parameters from different streamflow scenario combinations to determine which combination best describes the future validation streamflow. The utility of combining observed, reconstructed and projected streamflow in reducing the uncertainty in reservoir sizing will be tested and then validated with the storage obtained using the streamflow from the validation period.
The chapter is organized and follows: Section 4.2 provides information on the data sources used in this study. Section 4.3 describes the split-sample validation and the methodology considered for deriving reconstructed and projected streamflow along with the validation metrics used. Following that, Section 4.4 provides the results of the experiments which evaluate the utility of adding reconstructed and projected streamflow. Finally, Section 4.5 summarizes the findings and conclusions as well as discusses the utility of various sources of information in reservoir sizing.

4.2: Data

4.2.1: Annual Streamflow

Since we are considering combining observed, reconstructed and projected streamflow, we consider undeveloped basins from the Hydro-Climatic Data Network (HCDN) [Slack et al. 1993]. Streamflow sites and their associated records in the HCDN are nearly void of upstream storage and groundwater pumping. These basins have been identified as virgin basins by the USGS. The locations and summarized characteristics of the streamflow sites can be seen in Figure 4.1 and Table 4.1 respectively. Further, since the tree ring chronologies are only available on the annual time scale, the temporal scale of streamflow will also be annual.

4.2.2: Tree Ring Chronologies

The primary predictor of streamflow reconstructions is tree ring chronologies as discussed in Section 4.1. The tree ring chronologies for the study were obtained from the National Atmospheric and Oceanic Administration (NOAA) International Tree Ring Data
Bank (ITRDB) [available online at http://www.ncdc.noaa.gov/paleo/treering.html]. Tree Ring chronologies are standardized by removing the growth trend in the tree ring measurements due to the natural aging process of a tree. The selection of these chronologies for streamflow reconstruction is often limited to chronologies located within the boundaries of a reconstructed site’s basin [Hidalgo et al 2000; Woodhouse et al 2006]. However, in the Southeast, there is a limited amount of tree ring chronologies, thereby limiting the watersheds for which annual streamflow could be reconstructed. Hence, the selection of chronologies for streamflow reconstruction in the study was extended to chronologies within 200 km of the streamflow site.

Figure 4.1: Map of the streamflow and tree ring chronology sites used for this study

4.2.3: SST Database

Streamflow reconstructions in this study were obtained using the Combined model from Chapter 2, which utilizes the SST-TR and TR-PCA models. The SST-TR model
improves the prediction of high flow values in streamflow reconstruction by adding SSTs to the reconstruction predictors. Since, river basins in the Southeast are greatly influenced by ENSO, we consider Nino-3.4 index as an additional predictor for obtaining annual flows [Trenberth and Stepaniak 2001; Devineni and Sankarasubramanian 2010]. The Nino-3.4 index is an anomalous SST condition over the area of 5°S–5°N and 170°–120°W in the tropical Pacific. In the study, average annual Nino-3.4 calculated using Kaplan’s Analyses of global sea surface temperatures [Kaplan et al 1998] is used [available online at http://iridl.ldeo.columbia.edu/SOURCES/.KAPLAN/.Indices/.NINO34/].

4.2.4: Climate Change Projections

GCMs are used to obtain future climate conditions under near-term climate change. For this study, the Coupled Model Intercomparison Project phase 5 (CMIP5) monthly climate projects were used with the parameters being initialized in 1950. CMIP5 yields monthly precipitation, average surface air temperature, minimum surface air temperature and maximum surface air temperature for every 1/8° degree for the years 1950 - 2010. Data from CMIP5 was obtained from the Bureau of Reclamation’s climate database and May 2013 report (http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/dcpInterface.html). For each streamflow site, the upstream grid points were spatially averaged for each month, yielding a precipitation, average surface air temperature, minimum surface air temperature and maximum surface air temperature time series from 1950 – 2010 for each model run ensemble.
Table 4.1: Details of USGS gauging stations considered for combining paleo and observed data and future climate change projections

<table>
<thead>
<tr>
<th>Site Index</th>
<th>USGS Station Number</th>
<th>Station Name</th>
<th>Mean Annual Streamflow [cfs]</th>
<th>Drainage Area [mi²]</th>
<th>Tree Chronologies Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>02236000</td>
<td>St. Johns River near De Land, FL</td>
<td>3055</td>
<td>3066</td>
<td>FL007, FL008</td>
</tr>
<tr>
<td>2</td>
<td>02246000</td>
<td>North Fork Black Creek near Middleburgh, FL</td>
<td>188</td>
<td>177</td>
<td>FL005, FL007, FL008</td>
</tr>
</tbody>
</table>
There are 22 GCMs in CMIP5 available with each having a different number of ensembles. For each model, the spatially averaged ensembles were temporally averaged to obtain the ensemble mean. This resulted in monthly precipitation, average surface air temperature, minimum surface air temperature and maximum surface air temperature monthly time series from 1950 – 2010 for each CMIP5 GCM and for each of the selected streamflow sites. These time series will be used to obtain the projected streamflow time series.

4.3: Experimental Design

4.3.1: Split-sample Validation

The goal of this study is to reduce the uncertainty in annual streamflow generation by combining observed, reconstructed and projected streamflow. To test the uncertainty, we have constructed a split-sample validation experiment. In this experiment, the validation period will be over 1981 – 2010 and the observational period is assumed to be the observed annual streamflow for the water years before 1981. Essentially, the purpose of the validation experiment is to perform a planning exercise for the period 1981 – 2010 and validate it using the observed flow during that period. The observational and paleo periods considered for the split-sample validation are shown in Table 4.2.

In the split-sample validation, the intent is to use the observed streamflow, tree ring chronologies, precipitation, potential evapotranspiration (PET) and projected CMIP5 precipitation and temperature for obtaining the observed, reconstructed, and projected streamflow (Figure 4.2). The observed streamflow in the split-sample validation will be the
observed streamflow prior to 1981 and the development of reconstructed and projected streamflow is discussed in the following sections. The PET in this experiment will be obtained using the Hargreaves method, which estimates PET using solar radiation, and the minimum, maximum and average temperature.

Table 4.2: Performance of tree-ring based reconstruction model and the ‘abcd’ model in estimating the observed annual streamflow. Table also provides the paleo-information used for combining the three suites of hydroclimate data

<table>
<thead>
<tr>
<th>Site</th>
<th>Observed Period</th>
<th>Paleo Period</th>
<th>‘abcd’ Annual Correlation</th>
<th>Reconstruction Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1934 – 1980</td>
<td>1894 – 1933</td>
<td>0.94</td>
<td>0.74</td>
</tr>
<tr>
<td>2</td>
<td>1932 – 1980</td>
<td>1892 – 1931</td>
<td>0.90</td>
<td>0.55</td>
</tr>
</tbody>
</table>

4.3.1.1: Reconstructed Streamflow

The reconstructed streamflow in the split-sample validation will be obtained using the Combined model described in Chapter 2, which combines the traditional tree ring chronology approach, TR-PCA, with the SST and tree ring hybrid approach, SST-TR, based on the k-nearest neighbors. For each site, regressions (i.e. reconstruction models) between the observed streamflow and observed SST and tree ring chronologies were developed for the observational period (Figure 4.2). Once the reconstruction models were developed, the SST, tree ring chronologies were used to reconstruct streamflow for the pre-instrumental period. For the observational period, the reconstructed and observed streamflow have a correlation of 0.74 and 0.55 for Sites 1 and 2 respectively (Table 4.2). Time series of both observed and reconstructed streamflow are shown in Figure 4.3. The reconstructed streamflow is closely
related to the observed streamflow indicating its usefulness in providing insight into streamflow conditions for the pre-observational period.

4.3.1.2: Projected Streamflow

Reconstructed streamflow will provide insight into past climate conditions. However, under near climate change, there is no guarantee that the future climate will be representative of the past. To reduce uncertainty in future streamflow generation, we incorporate future climate change projections. Using the monthly time series of projected precipitation and temperature data for the period 1950-2010, projected streamflow was obtained using the ‘abcd’ water balance model (‘abcd’ Model). The ‘abcd’ Model uses precipitation and PET flux to estimate the river runoff [Thomas, 1981]. Since the ‘abcd’ Model has PET as its input, PET will need to be calculated using the temperature data available. There are many different ways to calculate PET, but this study uses the Hargreaves Method outlined below [Hargreaves and Samani, 1982]

\[ \text{PET} = 0.0075 \times R_a \times C_t \times \frac{1}{\delta_t} \times T_{avg} \]  (4.1)

The average temperature, minimum temperature and maximum temperature is available for each month starting in 1950. The average temperature is \( T_{avg} \) in Equation 4.1, and the difference between the maximum and minimum temperature is \( \delta_t \) in Equation 4.1. The temperature reduction coefficient, \( C_t \), is a function of relative humidity, but in this study is assumed to be 0.125. The incoming solar radiation, \( R_a \), is a function of the month, and the average monthly values for each site were used.
The ‘abcd’ Model has four parameters which determine the infiltration, runoff and evaporation of the monthly precipitation. ‘a’, ‘b’, ‘c’, and ‘d’ which are obtained by calibration that minimized the sum of squared errors between simulated and observed streamflow. Using these parameters and the projected climatic data from CMIP5, we obtain the projected streamflow for each site. The correlation between the observed and ‘abcd’ Model streamflow was 0.94 and 0.90 for Sites 1 and 2 respectively (Table 4.2), and Figure 4.4 shows the time series of both flows along with the calibrated parameters for each site. The ‘abcd’ Model flows explain most of the variance and identify the high and low flow years (Figure 4.4). This shows that if the projected precipitation and temperature from CMIP5 are representative of the future precipitation and temperature, the ‘abcd’ Model will be able to explain the future streamflow variability.

After the ‘abcd’ Model was calibrated, we calculated monthly projected streamflow for each of the 22 CMIP5 models starting in 1950 (Figure 4.2). As with the observed precipitation and PET, the monthly flows were added together, yielding projected annual streamflow. However, these projected streamflows will be biased due to the uncertainty in near-term climate change projections. To remove the bias, we employ quantile mapping between the observed and projected streamflow in the observational period. Then, we use that regression to remove the bias for the projected streamflow over the 1981 – 2010 validation period (Figure 4.2). Thus, we obtained multimodel monthly streamflow projections for the two sites over the validation period.
Figure 4.2: Modeling framework for obtaining the reconstructed, observed and projected streamflow for the selected two stations
Figure 4.3: Comparison between the reconstructed and the observed streamflow for the two sites.
Table 4.3: Quantiles of the mean, standard deviation and lag-1 correlation of the projected streamflow from the ‘abcd’ model forced with monthly CMIP5 change projections for period 1981 – 2010

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Site 1</th>
<th></th>
<th>Site 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
<td>Correlation</td>
<td>Mean</td>
</tr>
<tr>
<td>Observed</td>
<td>2875</td>
<td>1056</td>
<td>-0.04</td>
<td>174</td>
</tr>
<tr>
<td>0.25</td>
<td>3060</td>
<td>894</td>
<td>0.00</td>
<td>203</td>
</tr>
<tr>
<td>0.50</td>
<td>3187</td>
<td>946</td>
<td>0.11</td>
<td>213</td>
</tr>
<tr>
<td>0.75</td>
<td>3454</td>
<td>1044</td>
<td>0.24</td>
<td>226</td>
</tr>
</tbody>
</table>

The projected streamflow from the multimodel streamflow projection will be 22 projected streamflow time series (1 for each CMIP5 model). The goal of near-term climate change models is not to explain the interannual climate variability, but rather, they attempt to capture the change in the marginal distribution of climate. Table 4.3 compares the quantiles of the mean standard deviation and lag-1 correlation of the 22 projected streamflow time series for the validation period with the mean standard deviation and lag-1 correlation of the observed streamflow for the validation period. We can infer that even after quantile mapping, the projected streamflow overestimates the mean of the validation period in both sites. In Site 1, the projected streamflow underestimates the standard deviation and over estimates the lag-1 correlation. In Site 2, the lag-1 correlation is underestimated, but the projected streamflow captures the standard deviation well.
Figure 4.4: Comparison between observed and ‘abcd’ Model simulated streamflow for the assumed observational period over the selected sites.
Our split-sample validation that arises at sizing the reservoir for future demand intends to use observed streamflow, SST, tree ring chronologies, precipitation and temperature prior to 1981. Observed tree ring chronologies and SST were used to obtain reconstructed streamflow for the prior to 1930 and 1932, and projected precipitation and temperature were used to calculate projected streamflow for the validation period. The observed, reconstructed, and projected streamflow time series will be combined in a Bayesian framework and compared to the observed streamflow from the validation period.

4.3.2: Streamflow Scenarios

Streamflow is synthetically generated from streamflow characteristics (usually mean, standard deviation and lag-1 correlation) which are obtained from the observed time series. Traditionally, these characteristics are derived using only the observed time series. This study aims to reduce the uncertainty in streamflow generation by combining observed, reconstructed and projected streamflow.

As in Chapter 3, we will be generating lognormal streamflow to ensure our flow values are positive (Equation 4.2). Using this approach, a log-transformed streamflow time series, $Y_t$, of length $n$, can be generated using the mean, $\mu$, standard deviation, $\sigma$, and lag-one autocorrelation, $\rho$ of streamflow.

$$Y_{t+1} = \mu + \rho(Y_t - \mu) + \epsilon_t \sigma(1 - \rho^2)^{1/2} \quad t = 1 \ldots n \quad (4.2)$$

However, the observed mean, standard deviation and lag-1 correlation should not be used to generate streamflow as the estimated moments are limited by the small sample size. A Bayesian framework will be used, as in Chapter 3, to capture the parameter uncertainty.
The posterior distributions of the mean, standard deviation and lag-1 correlation (µ, σ, and ρ) from the Bayesian framework will be used to generate streamflow for the validation period.

4.3.2.1: Observed Streamflow Scenario

The observed streamflow scenario (O) is the traditional approach for synthetic streamflow generation. To obtain its posterior distribution, the log of the observed streamflow prior to 1981 is used in the same Bayesian framework from Chapter 3 yielding posterior distributions of µ, σ, and ρ (Figure 4.5). The sets of µ, σ, and ρ are then used to calculate the multivariate lognormal likelihood of the observed flows for the validation period (1981 – 2010) yielding a distribution of likelihoods. These likelihoods are normalized by the likelihood of the climatological flows (i.e. stationary flows) (Section 4.3.3.1). The resulting likelihood ratio distribution will be compared with the likelihood ratio distributions of the streamflow combination scenarios to determine if combining observed, reconstructed and projected streamflow reduces the uncertainty in streamflow generation.

4.3.2.2: Streamflow Combination Scenarios

Since observed streamflow is the best information available, all streamflow combination scenarios will contain the observed streamflow under the split-sample validation. We will consider three streamflow combination scenarios: observed and reconstructed (OR), observed and projected (OP) and observed, reconstructed and projected (ORP) (Figure 4.5). Each combination will be used in a Bayesian combination framework resulting in posterior distributions of µ, σ, and ρ. The Bayesian combination framework for the combinations with reconstructed streamflow is the same framework as discussed in
Chapter 3. In this framework, we add noise to the reconstructed streamflow to account for the unexplained variance of the reconstruction model.
When combining projected streamflow from GCMs in CMIP5, we must consider all 22 projected streamflow time series. During Markov Chain Monte Carlo (MCMC), which estimates the posterior distribution in the Bayesian Framework, the likelihood of a streamflow time series with the set of generation parameters is calculated to obtain the posterior distribution of parameters. Since there are 22 projected streamflow time series, there will be 22 unique likelihoods for each parameter set. In order to run MCMC, a single likelihood value is needed for each parameter set which implies that the 22 likelihoods for the projected streamflow need to be combined.

The likelihoods of the projected streamflow and a parameter set will be combined using the weighted average in Equations 4.3 – 4.4. In these equations, each of the 22 GCMs is denoted by ‘i’. The likelihood of the projected streamflow GCM i for the observational period, $L(P_i^O)$, is calculated by the multivariate lognormal likelihood of the observed streamflow from 1951 – 1980, $O$, and the mean, standard deviation, and lag-1 correlation of the projected streamflow from 1951 -1980, $(\mu(P_i^O), \sigma(P_i^O), \rho(P_i^O))$ (Equation 4.4). The posterior likelihood of the parameter set from the projected streamflow from MCMC, $L([\mu, \sigma, \rho]|P)$, is the weighted average of the likelihood of the projected streamflow from 1981 – 2010 GCM i, $L(P_i^V|[\mu, \sigma, \rho])$, where the weights are determined by the $L(P_i^O)$ terms (Equation 4.3).

$$L([\mu, \sigma, \rho]|P) = \frac{\sum_{i=1}^{22} L(P_i^V|[\mu, \sigma, \rho]) \cdot L(P_i^O)}{\sum_{i=1}^{22} L(P_i^O)}$$ (4.3)

$$= L(O|[\mu(P_i^O), \sigma(P_i^O), \rho(P_i^O)])$$ (4.4)
For each of the 22 projected streamflows, the likelihood of the observed flows from 1951 – 1980 (O) and the mean, standard deviation and lag-1 correlation of the projected flows from the same time period \((\mu(P_t^O), \sigma(P_t^O), \rho(P_t^O))\) is calculated. These likelihoods will be the weights in Equation 4.2, so the CMIP5 models which best explain the observed streamflow during 1951 – 1980 will be given the highest weights for the posterior distribution.

Based on the estimated posterior likelihood of the projected streamflow (Equation 4.3) under MCMC, it will be multiplied by the likelihood from the observed and observed and reconstructed for the OP and ORP streamflow combination scenarios respectfully. Note that we are assuming the projected streamflow is independent of the observed and reconstructed streamflow. Once the overall likelihood is calculated, the MCMC procedure continues as normal, yielding a posterior distribution of \(\mu, \sigma, \text{ and } \rho\).

As with the observed model, the posterior distributions of each streamflow combination scenario will be used to obtain distributions of the likelihood for the validation period (Figure 4.5). These likelihood distributions will also be normalized with the validation likelihood with the validation climatology. This will result in four distributions of likelihood ratios (O, OR, OP, and ORP) to determine which streamflow scenario has the lowest uncertainty.
4.3.3: Performance Metrics

To determine which streamflow scenarios best represent future streamflow conditions, we have used several performance metrics. For each metric, the validation period is 1981 – 2010, and the observed flows during that period are the validation flows.

4.3.3.1: Likelihood Ratio Distribution under Validation

As described in Section 4.3.2, we can obtain a distribution of likelihoods from the posterior distribution of \( \mu, \sigma, \) and \( \rho \). The formulation can be seen in Equation 4.4 for a given posterior distribution, where \( i = 1:n \) and \( n \) is the number of \( \mu, \sigma, \) and \( \rho \) sets in the posterior distribution.

\[
L_i = L(Q_{1981-2010} | [\mu_i, \sigma_i, \rho_i]) \quad (4.5)
\]

The likelihood in Equation 4.5 is used to calculate the likelihood ratio in Equation 4.6.

\[
LR_i = \frac{L_i}{L(Q_{1981-2010} | [\mu(Q_{1981-2010}), \sigma(Q_{1981-2010}), \rho(Q_{1981-2010})])} \quad (4.6)
\]

Note that the denominator is the likelihood of the observed streamflow from 1981 - 2010 with the mean, standard deviation and lag-1 correlation of the observed streamflow from the same time period. Essentially, it is the maximum likelihood for the validation period. This means that the maximum of the likelihood ratio, \( LR \), is 1.0 (the posterior \( \mu, \sigma, \) and \( \rho \) set is the same as the observed \( \mu, \sigma, \) and \( \rho \) set for 1981 – 2010). We will compare the distribution of these likelihood ratios for all four streamflow scenarios.
4.3.3.2: Flow Duration Curve Generation

Using the posterior distributions for each streamflow scenario, we generate streamflow traces. To generate streamflow with information from the validation period, posterior distributions must be obtained using streamflow from 1981 - 2010. This will be done using the same methodology as the observed (O) streamflow scenario in Figure 4.5. A Bayesian framework will be used on the observed streamflow during the validation period, yielding posterior distributions of µ, σ, and ρ.

Posterior distributions from the validation period and streamflow scenarios are used to generate streamflow traces. For a given posterior distribution, a random µ, σ, and ρ will be chosen and a 30 year streamflow trace will be generated as in Equation 4.2. This will be done 10,000 times for each scenario, yielding 10,000 streamflow trances with length 30.

A flow duration curve for a given streamflow time series can be obtained by plotting a flow value (Q_i) vs. its exceedence probability (EP_i) where exceedence probability is shown in Equation 4.7 where n denotes the length of the time series.

\[ EP_i = 1 - \frac{Rank(Q_i)}{n + 1} \quad (4.7) \]

For a given streamflow scenario, there are 10,000 streamflow traces, meaning there are 10,000 flow duration curves. Since all the streamflow traces are the same length, they have the same exceedence probabilities. To obtain the validation flow duration curves for each scenario, the median flow value of the 10,000 flows for each exceedence probability will be plotted against the associated exceedence probability. These flow durations will be
compared with the median flow duration curve from the generated streamflow using information from the validation period.

4.3.3.3: Reservoir Storage Estimation under Validation

As in Chapter 3, storages can be calculated from a given streamflow trace using the SQP (Equations 4.8-4.9) [Thomas et al. 1963]. The storage required at time \( t \) is the sum of the storage required at the previous time step plus the water supply demand (D) and evaporation (E) minus the incoming streamflow at time \( t \). If the inflow is greater than the previous storage, demand and evaporation, the required storage at time \( t \) is zero. The storage required for the entire trace (\( t = 1 \ldots n \), where \( n \) is the length of the streamflow record) is found by taking the maximum of the storage over the entire trace. The assumed demand and evapotranspiration in this study will be 80% of the mean annual streamflow for the observational period.

\[
S = \text{Max}(S_t) \quad (4.8)
\]

\[
S_t = \begin{cases} 
S_{t-1} + D + E - Q_t & \text{if positive} \\
0 & \text{otherwise}
\end{cases} \quad t = 1 \ldots n \quad (4.9)
\]

The SQP will be used on the 30 year streamflow traces from Section 4.3.3.2 from the streamflow scenarios and validation streamflow. This will yield distributions of required storages. We will compare the storage distributions of the streamflow scenarios to the distributions of storages obtained using observed flow under the validation period. This will show how each streamflow scenario reliably delivers future demand, which is how a reservoir would be designed in a real world application. We are particularly interested in the
0.8 – 0.95 quantiles of the distributions because they correspond to the storage ensuring 80 – 95% reliability for the assumed demand.

4.4: Results

The posterior distributions of streamflow generation parameters were obtained using the split-sample validation described in Section 4.3. The likelihood of the observed streamflow from the validation period with each posterior distribution was calculated using the formulation in Section 4.3.3.1. The posterior distributions were also used to generate streamflow and obtain the required storage distribution outlined in Sections 4.3.3.2 and 4.3.3.3.

4.4.1: Likelihood with Streamflow under Validation

To determine which streamflow scenario best explained future streamflow events, the likelihood of the validation streamflow with every set of streamflow generation parameters from each posterior distribution was calculated. These likelihoods were then divided by the likelihood of the streamflow under validation and the climatology under as described in Section 4.3.3.1 yielding a distribution of likelihood ratios for each streamflow scenario. Figure 4.6 shows the cumulative density function (CDF) of likelihood ratio distribution for each streamflow generation scenario. The OR streamflow scenario has the highest likelihood ratio for each cumulative probability in both sites meaning the streamflow generation parameters from observed and reconstructed streamflow are more representative of the streamflow under the validation period in comparison to the other streamflow scenarios. Further, the upper quantiles of the OR likelihood ratio distribution are approximately 1.0, the
maximum likelihood ratio. The findings in Figure 4.6 illustrate that the inclusion of reconstructed streamflow improves streamflow generation. This explains the findings found in Chapter 3 that including reconstructed streamflow would better explain the long term streamflow record. Figure 4.6 shows that reconstructed streamflow can better explain future streamflow events as well.

Figure 4.6 also shows that the OP streamflow scenario had the lowest likelihood ratio for each cumulative probability at each site. We can see that the inclusion of the projected streamflow decreases the streamflow generation skill because the O streamflow scenario has a higher likelihood ratio at each probability. Table 4.3 shows that the projected streamflow from CMIP5 did not capture the streamflow under validation, so GCMs are not guaranteed to capture future climate events. Even though the reconstructed streamflow is an estimate, it is estimated based on observations each year. Hence, the study reveals that the reconstructed streamflow would give more insight than the projected streamflow.
Figure 4.6: The CDF of the ratio of likelihood distribution to the likelihood under climatological streamflow
4.4.2: Streamflow Generation under Validation

Based on the performance of the CMIP5 projected streamflow under the validation period, the OP and ORP streamflow scenarios were not used for stochastic streamflow generation. However, the posterior distributions of the O and OR streamflow scenarios as well as the posterior distribution from the streamflow under validation were used to generate multiple 30 year streamflow traces. The exceedence probabilities of each streamflow trace were combined, and the median of the streamflow was considered to create the flow duration curves described in Section 4.3.3.2. Figure 4.7 shows the median flow duration curve of the streamflow under validation as well as the deviation of the O and OR streamflow scenario flow duration curve for both sites. In Site 1, the OR flow duration curve is closer to the validation flow duration curve except for the higher and lower exceedence probabilities. This further illustrates the finding in Section 4.4.2 that including reconstructed streamflow will result in streamflow traces that are more representative of future events.

The findings in Site 2 are similar, except the flow duration curve from the observed streamflow is the closest to the flow duration curve for the validation period for exceedence probabilities 75% and above. The flow duration under the OR streamflow scenario is still the closest overall though, which supports the finding in Section 4.4.2. However, the higher exceedence probabilities represent the below normal flow values which control the required storage. Therefore, we expect the O streamflow scenario in Site 2 to yield better required storage estimates, especially the higher reliability yields.
Figure 4.7: Difference between the flow values obtained under observed/observed-reconstructed streamflow traces and flow values under validation period for different exceedence probabilities. The secondary y-axis is shown for the actual FDC for the validation period.
Figure 4.8: Difference between validation storage and estimated storages using different streamflow scenarios for the selected sites. Demand is assumed to be 80% of the mean annual streamflow for respective sites.
4.4.3: Reservoir Sizing under Validation

The SQP was used along with the generated streamflow traces from Section 4.4.2 to obtain required storage distributions as described in Section 4.3.3.3. The storage distributions from the O and OR streamflow scenarios are compared to the storage distribution from the validation period in Figure 4.8. Note that the difference between the two distributions appears to be larger in Site 2 only because of the scale of the y-axis. Also, only the difference in distributions between the 0.8 and 0.95 quantiles are shown. This is because the quantile of the storage distribution denotes the reliability of the assumed demand for a given storage.

In Site 1, the OR streamflow scenario has a storage distribution that is closer to the validation storage distribution. This result is expected due to the closeness of the flow duration curve from the OR streamflow scenario and the validation streamflow. However, in Site 2, the O streamflow scenario yields a reservoir storage distribution closer to the validation storage distribution than the OR streamflow scenario. This is due to the finding in Section 4.4.2 that the O streamflow scenario generates lower flows closer to the generated flows under validation in comparison to the OR streamflow scenario. Since the upper quantiles of the storage distribution depend greatly on the lower quantiles of the inflow, it is expected that the O streamflow scenario would have better storage estimation in Site 2. It is our recommendation that if streamflow generation is going to be used for reservoir sizing that the streamflow scenario which best explains the lower flows better be used for estimation.
4.5: Conclusions

Since observed streamflow is limited to around 80 years in the United States, there is no guarantee that the observed drought record captures the pre-historic droughts. This limits the usefulness of observed streamflow in stochastic streamflow generation, especially if large projects such as reservoirs will be based on these generations. Furthermore, to be effective, streamflow generations need to be representative of future streamflow conditions, as indicated by the near-term climate change projections. To address these limitations, we proposed combining observed, reconstructed and projected streamflow to reduce the uncertainty of streamflow generation.

A split-sample validation was set up for two sites in Florida, where only observations prior to 1981 were considered. The performance of different combination of generation schemes developed using observed, reconstructed and projected streamflow time series were compared to the observed streamflow in the validation period (1981 – 2010). A Bayesian framework was employed to combine different streamflow time series for obtaining the distributions of streamflow generation parameters under each scheme. Likelihoods of the observed streamflow under validation based on the prior parameter distributions were calculated yielding a distribution of likelihoods. The combination of observed and reconstructed streamflow had the highest likelihood distribution for both sites illustrating the argument that the addition of reconstructed streamflow yields streamflow generation traces that represent pre-historic flow characteristics. The projected streamflow did not add any skill to the streamflow generation parameters. This is due to the lack of skill in the near-term climate change projections.
The streamflow generation parameters were then used to generate 30 year streamflow traces, which were compared to streamflow traces generated using the observed streamflow from the validation period. The median flow duration curve from the observed and reconstructed streamflow was closer to the flow duration curve under validation overall in both sites. However, in Site 2, the flow duration curve from the observed streamflow was closer to the validation curve for the higher exceedence probabilities. When the SQP calculated the required storage from the streamflow traces, the storage estimates from the observed streamflow were closer to the validation estimates in Site 2. This is due to the observed streamflow better explaining the lower flows at this site. Hence, for reservoir design, it is recommended that the streamflow scenario which better explains lower flows be chosen.
Chapter 5: Conclusions and Future Work

5.1: Conclusions

The major goal of this study is to reduce uncertainty in estimating reservoir storage for the specified reliability. Given that the streamflow in the Southeastern US only has around 80 years of observed records, estimating reservoir storage that can supply the desired demand even under extreme drought (e.g.: once in 100 years) is limited. To overcome this limitation, we considered reconstructed streamflow using tree rings to gain insight into paleo droughts, but such estimates tend to underestimate high flow values due to a metabolic growth limit of trees. Thus, the first objective of this study was to improve annual streamflow reconstruction estimates using the traditional approach (TR-PCA), particularly in reconstructing high flow values. By considering SSTs as an added predictor (SST-TR), we improved high flow estimates in reconstructed streamflow values. By combining the traditional approach with the two SST-TR models, the combined model improved the skill under all flow categories. The reconstructed streamflow in the combination model were then used to reduce the uncertainty in stochastic streamflow generation and reservoir sizing.

A common approach to reservoir sizing is to generate multiple realizations of synthetic streamflows to obtain a distribution of required reservoir storages using the SQP. Such approach has limitations since parameters of the stochastic streamflow generation model are purely derived from the observed streamflow record. By combining the observed streamflow record with the improved reconstructed streamflow records in the Southeastern US, we estimated reservoir storages for two virgin basins. Based on split-sample validation,
we showed that, the streamflow generation from the combination of observed and reconstructed streamflow yielded better storage distributions that are closer to the storage distributions estimated from the “true” streamflow. We also observed that as the length of the reconstruction increased, the skill of both streamflow generation parameters and storage estimates improved. Further, to generalize the findings, we considered reconstructed records with different lengths and skills. Under this, we observed that uncertainty in storage estimates reduced as the skill and length of the reconstruction record increased. Thus, we concluded that estimation from the addition of reconstructed streamflow clearly reduces the uncertainty in reservoir storage estimates.

On limitation in using observed and reconstructed records for reservoir sizing is that it assumes future flows are representative of the past. Under near-term climate change, there is no guarantee that the future streamflow conditions will have the same moments as that of the observed flows. To accommodate future streamflow characteristics, we considered future streamflow from near-term climate change projections. The projected streamflow estimates were combined with the observed and reconstructed streamflow data to reduce uncertainty in future streamflow generation and reservoir sizing. Based on a split-sample validation, we considered only observations prior to 1981 and evaluated the performance of observed, reconstructed and future streamflow in estimating observed streamflow over the validation period (1981 – 2010). A Bayesian framework was employed to combine the three – observed, reconstructed and future – streamflow time series and yield distributions of streamflow generation parameters. The performance of the combined stochastic models was tested by calculating the likelihood of the observed streamflow from the validation period
with the generation parameters from the distributions. The distribution from the combined observed and reconstructed time series (OR) yielded the highest likelihoods with the validation streamflow of all the models, meaning the addition of reconstructed streamflow yielded better insight into future streamflow characteristics and reduced the uncertainty of future streamflow generation. Distributions with the addition of projected streamflow had the lowest likelihood with the validation streamflow. Thus, we concluded that projected streamflow would not reduce the uncertainty of streamflow generation due to the lack of skill in near-term climate change projections.

To further analyze the performance of the observed and observed and reconstructed streamflow models, 30 year streamflow traces generated from parameter distributions were compared to 30 year streamflow traces using the observed streamflow from the validation period. Quantiles of 10,000 streamflow realizations were combined to obtain a median flow duration curve for both models and the validation streamflow. The flow duration curve from the observed and reconstructed streamflow combination was closer to the curve from the validation streamflow, illustrating that the addition of reconstructed streamflow reduces future streamflow generation uncertainty. However, in Site 2 of Chapter 4 one of the two sites, the observed model’s flow duration curve better explained the higher exceedence probabilities.

The application of these streamflow generations in reservoir sizing was tested using the SQP, yielding a distribution of required reservoir storages. For the upper quantiles (higher reliabilities) of the storage distribution, the observed and reconstructed distribution better explained the validation storage in Site 1 and the observed distribution better explained
the validation storage in Site 2. Due to the models’ median flow duration curves performance at lower flows at these sites, we concluded that the model which best explains future low flow values will yield better reservoir estimates.

5.2: Future Work

The next step in this research is to apply the methodology across different sites in the US, since this study only focuses on streamflow in the Southeastern US. While, the reconstructed and observed streamflow combination methodology can be tested anywhere with valid tree ring chronologies, the hybrid SST and tree ring reconstruction approach is only applicable in areas with a high correlation with one or more SST anomalies. In the US, these areas are the Sunbelt and Pacific Northwest. For the sites in the Southeastern US in this study, annual streamflow is highly correlated with ENSO. There is a high relationship between ENSO in the Sunbelt and Pacific Northwest. However, streamflow sites in the Western US are also highly affected by the Pacific Decadal Oscillation (PDO), which has a periodicity of around 10 years. Application of the hybrid SST and tree ring chronology reconstruction methodology in the Western US would have to identify both the ENSO and PDO streamflow components to reduce reconstruction uncertainty.

The reconstructed and observed streamflow combination methodology would also be practical in the Western US, because most reservoirs operate as over-year reservoirs as in this study due to the high variability of annual streamflow. Due to the large size, reservoirs in the Western US also provide hydropower to an area. Another future project using work from this study could be combining observed, reconstructed and projected streamflow to reduce
uncertainty in reservoir sizing for water supply and hydropower. For this project the SQP would still be used for water supply sizing and a hydropower model would be needed for the hydropower demand. Another future research project could take into account the variability in future water demand. For this study, we assumed the demand and evaporation to be constant, but in reality, populations will increase, straining the water supply. Furthermore, temperatures are increasing due to near-term climate change, meaning the evaporation will increase in the future. It is important to note though that evaporation is dependent on the reservoir surface area, so an elevation-area relationship would be needed for variable evaporation.

While over-year reservoirs are common in the Western US, reservoirs in the Southeastern US behave as within-year reservoirs where a reservoir only has enough storage capacity for one year and is designed to fill up during the year. Unlike the Western US, the streamflow variability in the Southeastern US is lower, making within-year reservoirs functional. Over-year reservoirs were considered in this study to the annual temporal scale constraint of tree ring chronologies used for streamflow reconstruction. Within-year reservoirs must be designed with at least seasonal or monthly inflows therefore tree ring chronologies alone could not be used to obtain reconstructed streamflow. However, monthly ENSO data is available from 1857, meaning monthly reconstructed streamflow could be estimated. The ENSO state alone could be used as a predictor, or the annual tree ring chronologies could be combined with the monthly ENSO state in a hierarchical Bayesian framework. The resulting monthly reconstructed streamflow could be combined with the
observed monthly streamflow to reduce streamflow generation and reservoir sizing for within-year reservoirs.
REFERENCES


Appendix A: SSA Methodology Description

To separate the periodic and non-periodic components of streamflow and tree-ring chronologies, we employ SSA developed by Ghil et al. [2002]. SSA is a time series approach that quantifies the lag dependence in streamflow through estimates of lagged cross-covariance [Shun and Duffy, 1999; Ghil et al. 2002]. Like PCA, SSA produces orthogonal PCs, the variance of the PCs (eigen values) and the eigenvectors corresponding to each PC. Similarly, SSA also brings each PC obtained from the lagged cross-covariance matrix back to the original space creating reconstructed components (RCs). If one uses all the PCs for reconstruction, it is expected to produce the original time series. Given that SSA employs PCA on lagged time series of tree rings, year-to-year storage characteristics or persistence of annual streamflow would also be incorporated in the reconstructed streamflow. For complete details on the SSA, see Ghil et al. [1992].

Consider $X$ denotes the time series of tree-ring chronology at a given site or annual average Nino3.4 over the tropical Pacific with length $N$. To begin with, SSA considers an embedding dimension, $M$, which denotes the number of lagged time series to be created. The embedding dimension is usually chosen within $1/4$ to $1/3$ of $N$. The lagged matrix, $\tilde{X}$, shown in (A-1), has a length of $N' (N-M+1)$.

\[
\tilde{X} = \begin{bmatrix}
X(1) & X(2) & \cdots & X(M) \\
X(2) & X(3) & \cdots & X(M+1) \\
\vdots & \vdots & \ddots & \vdots \\
X(N') & X(N'+1) & \cdots & X(N)
\end{bmatrix} \quad (A - 1)
\]
A covariance matrix, $C_X$, of dimension $M \times M$ is estimated and then decomposed into the respective eigenvectors and eigenvalues, $E_X$ and $\Lambda_X$, using orthogonal decomposition based on equation (A-2).

$$E_X^T C_X E_X = \Lambda_X \quad (A - 2)$$

Similar to PCA, principal components, $A^k$, are found by multiplying the eigenvectors by the original values of $X_t$ using equation (A-3).

$$A^k(t) = \sum_{j=1}^{M} X(t+j-1)E^k(j), \quad 1 \leq t \leq N' \quad (A - 3)$$

There will be a PC for each column of the lagged matrix totaling $M$. The orthogonal PCs for SSA are brought back to the original time space by convolving and normalizing the corresponding PCs and eigenvectors creating Reconstructed Components (RCs), $R^k(t)$, based on equation (A-4).

$$R^k(t) = \frac{1}{M_t} \sum_{j=L_t}^{U_t} A^k(t+j-1)E^k(j), \quad 1 \leq t \leq N \quad (A - 4)$$

The $M_t$, $L_t$, and $U_t$ terms in the equation above represent the normalizing term, lower limit and upper limit for each time step in $X$. The need for these terms arises from bringing each PC back to the original space since the eigen elements are calculated using the lagged matrix. For instance, $X(2)$ is located in elements $(2,1)$ and $(1,2)$ of the lagged matrix. Thus, the values of $M_t$, $L_t$, and $U_t$ could be obtained based on equation (A-5).
Thus, RCs can be added together to recreate the original time series. Each RC will also share the same periodicity as its associated eigenvector. Monte Carlo Singular Spectrum Analysis (MC-SSA) can be used to identify periodic components of a time series [Shun et al. 1999]. Periodic components of a time series can also be identified based on Signal to Noise (S/N) separation of oscillatory pairs of eigen elements [Ghil et al. 2002]. Geophysical time series, such as ENSO, often have greater power at lower frequencies, so a red noise “null hypothesis” [Allen, 1992] is often used for periodic component detection [Ghil et al. 2002].

The test for periodic detection of RCs against red noise devised by Allen [1992] is known as MC-SSA. MC-SSA generates simulated red noise data from the original time series and computes a covariance matrix $C_R$ (subscript $R$ denotes red noise). The eigenvalues $\Lambda_R$ of the red noise data is calculated by projecting $C_R$ onto the eigenvectors $E_X$ from SSA (equation A-6).

$$\Lambda_R = E_X^T C_R E_X. \quad (A - 6)$$

The generation of red noise data and subsequent calculation of $\Lambda_R$ are performed a large number of times giving an ensemble of eigen values $\Lambda_E$. The 95% confidence limits are calculated by finding the interval between the 2.5th and 97.5th percentile of a component of the eigenvalue ensemble $\Lambda_E^k$. If a corresponding eigenvalue from SSA $\Lambda_E^k$ is outside the 95% confidence range of $\Lambda_E^k$, then the eigenvalue is not due to the red noise. Thus, the RCs
associated with an eigenvalue outside the confidence range can be said to be “periodic.”

Since ENSO exhibits periodicity of three to seven years, the RCs associated with ENSO will have a similar periodicity.