2-19-2015

Predicting Flood Severity for New England Basins with a Semi-Distributed Hydrologic Model

Eric W. Buckley
University of Connecticut - Storrs, eric.buckley@uconn.edu

Predicting Flood Severity for New England Basins with a Semi-Distributed Hydrologic Model

Eric William Buckley

B.S., University of Connecticut, 2002

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

at the

University of Connecticut

2015

APPROVAL PAGE

Master of Science Thesis

Predicting Flood Severity for New England Basins with a Semi-Distributed Hydrologic Model

Presented by

Eric William Buckley, B.S.

Major Advisor ____
Emmanouil N. Anagnostou

Associate Advisor ____
Amvrossios C. Bagtzoglou

Associate Advisor ____
David M. Bjerklie

University of Connecticut

2015

Acknowledgements

I would like to start by thanking my Major Advisor, Professor Emmanouil N. Anagnostou, whose enthusiasm for his work is contagious and inspired me to have a better appreciation for what it was that I was doing in my career, as opposed to just knowing how to get the answer; and for always making the time for me throughout. I would also like to thank Dr. David M. Bjerklie for not only allowing me use his hydrologic model for my research, but also for taking the time out of his day(s) to meet with me and help me get this work off of the ground. Additionally, I would like to thank Professor Amvrossios C. Bagtzoglou for his support in serving on my defense committee.

Further, I would like to thank Madan and Manish Gupta of GM2 Associates, Inc. who supported me in my graduate studies. I would especially like to thank my wife, Carrie, and my sons Micah and Brody for always being there for me and understanding that Daddy's busy but can't wait to play with you soon. And to my parents; Mom, Dad, Karen and Leon, and Joe and Debbi Oparowski…you were right, I could do it! Thank you all for being there…this is for you.

Abstract

In an effort to modernize the state of practice of flash flood forecasting, recent research has shown promise in utilizing regionalized, continuous, distributed hydrologic models. Additional avenues of refining the forecasting methods have included attempting to forecast event frequencies in lieu of relying on flood magnitudes generated by the models. It is anticipated that this additional post processing of the distributed modeling results can alleviate some modeling errors inherent with trying to represent any natural process. This study examines the application of a regional distributed hydrologic model of lower New England, specifically the results calculated at internal sub-basins by comparing those results to historical gauged data. Then, through frequency transform methods, the forecasting potential is reassessed to determine if frequency predictions can increase confidence in predicting a flash flood. This study further addresses the sensitivity of the spatial scale of the subject catchments as well as the temporal scale in determining the effectiveness of both the distributed model and the frequency prediction. It was found that by post processing the predicted data, the bias in the forecasted events was greatly reduced as compared to the raw output from the modeling. This bias was also sensitive to the resolution of the time step, with the error directly related to that resolution. On the spatial scale, it was shown that the variation in the catchment size did not have a significant impact on the results. Overall, it is shown that there is value to post processing hydrologic modeling results from a continuous, distributed model in order to predict a probability of exceedance as opposed to basing flood warnings on raw flow magnitude calculations.

1. Introduction

Hydrologic models are useful tools to evaluate climate induced or anthropogenic changes in parameters to a study basin. An example would be land development in a study watershed; a hydrologic model would be useful in determining the impact to assumed increases in impervious surface. Similarly, a hydrologic model would be useful in determining the runoff relationship to variable rainfall events. Taking the modeling a step further, in the event that calibration of a model to a known (or series of known) events were possible, the tool becomes more useful in watershed management, and even river stage forecasting to predict flood events and impacts.

Flooding provides for 50% of all water related natural disasters and 15% of all deaths related to all natural disasters (WMO, 2011). Nondiscriminatory, flooding can occur anywhere and under any condition. Flooding can occur under low intensity rainfall events if the event follows strong drought conditions, or inversely, if the soil has been previously saturated following a preceding rainfall event. Flooding will also occur following high intensity rainfall events, regardless of the condition of the catchment, though the hydrologic conditions of the catchment will prescribe the severity of the resultant flooding. Destructive flooding will occur both in wide, flat floodplains, as well as mountainous, high relief areas in the form of flash flood. According to the National Flood Insurance Program, floods and flash floods are considered the number 1 natural disaster in the United States and resulted in almost eight billion dollars in approved flood insurance claims in 2012 alone. The World Meteorological Organization recognizes that there is an increasing trend in flood related socio-economic damages which are due to natural changes such as increases in rainfall intensity and duration as well as anthropogenic influences such as changes in land use and the increasing population and concentrations of that population migrating to flood prone areas. For this reason, maintaining a robust flood and flash flood forecasting system is essential to protecting life and property. At the heart of that forecasting system is the hydrologic model.

The current state of the industry for flash flood forecasting in the United States is based on a rainfall-runoff threshold approach established by the National Weather Service (NWS). Flash floods as defined by the NWS are considered as those flood events that occur within 6 hours of the causal rainfall event and usually occur in basins less than 100 square miles (Reed et al., 2007). The current framework for flash flood forecasting established by the NWS is two tiered, where regional River Forecast Centers (RFC) prepare hydrologic data for use by local Weather Forecast Offices (WFO). This system is comprised of meteorologists working at the local level with hydrologists at the regional RFC's. The RFC's (12 for the lower 48 states plus 1 for Alaska) are tasked with monitoring antecedent moisture conditions within their respective regions and determining the amount of rainfall that will produce runoff to a previously determined flood threshold (RFC Development Management Team, 2003). Carpenter et al. (1999) presents the threshold runoff or the flooding flow as that which meets the representative bankfull discharge, and further indicates the conservatism inherent in this definition as the bank full discharge would generally have to be exceeded to produce damaging flood conditions. Bankfull discharge is commonly considered the 1.5-year flood having a 67% chance of exceedance in a given year (Rosgen, 1994). This amount of precipitation per given time is termed the Flash Flood Guidance (FFG) product . Rates per time are produced on various time scales with RFC's typically producing 1-, 3-, and 6-hour products or even products up to inches per 24-hour (Reed et al., 2007).

Development of the FFG national product is detailed in Georgakakos et al. (1993) and further developed in Carpenter et al. (1999). The underlying parameter that develops the FFG is the amount of rainfall that becomes runoff, accounting for infiltration, evapotranspiration and local detention potential for a given watershed. This initial abstraction potential is maintained through continuous water and energy balance modeling at the RFC's. Additionally, an underlying assumption facilitating the threshold runoff approach is that the relationship between runoff in a watershed and the excess rainfall is linear. Once the effective rainfall is determined with respect to potential initial abstraction, that effective rainfall depth is compared to previously defined threshold runoff values for given catchments. Mathematically, with the flooding flow known for a catchment (Q_p) and the developed synthetic unit hydrograph for a duration of interest (q_{pR}) and the known catchment area (A), the effective runoff, or threshold runoff (R) is compared by

$$\boldsymbol{R = Q_p / q_{pR} A} \tag{1}$$

If the forecasted rainfall is equal to or greater than the required rainfall or computed threshold runoff to meet the peak flood flow for the duration specified, than a flood watch or warning should be announced by the local WFO.

In 2004, Smith et al. published the motivation and methods of the Distributed Model Intercomparison Project (DMIP) funded by the NWS. This project was initiated based on the NWS desire to maintain state of the art hydrologic and meteorologic practice in its flood forecasting capabilities. It was recognized that the evolution of hydrologic sciences, modeling technologies and remote sensing techniques had far surpassed the requirements of the current threshold runoff techniques as described above. Further, with the passage of time and work performed in the hydrologic field, there was an abundance of developed data on record with which to expand on forecasting capabilities available to the RFC's responsible for providing data to be used in the National Weather Service River Forecasting System (NWSRFS). Reed et al. (2004) presents the results of the DMIP. It appears that distributed models overall do not significantly outperform the lumped model currently in use for predicted peak flows. This is whether calibration is performed on the models or not. It is noted though that there are a "significant" number of cases where the distributed modeling will perform equal to or better than the lumped model, though this is not general over all test basins. In fact, it was shown that the performance of many of the distributed models were basin specific; performing well in one basin but not another. There are numerous observational differences between the two modeling approaches noted. For instance, in a distributed model with a hydraulically based routing routine, the timing of the peak flow is dependent on the volumetric flow rate computed. This is in contrast to the unit hydrograph method applied in the lumped parameter model, where time to peak is constant. For these models where the routing is more physically based, the result is more sensitive to the antecedent soil moisture budget. It is further shown that the calibrated distributed models with physically based routing schemes performed consistently better than those that were uncalibrated, showing the need for a rigorous calibration process. This may not be directly considered a benefit over lumped modeling in that the calibrated lumped model still outperformed the calibrated distributed models overall and calibration of a distributed model is typically more difficult than that of a lumped model. Another relatively obvious difference between the two modeling approaches is the potential to predict the volumetric flow rate of a distributed model at any choice interior point of the modeled basin. This would be a benefit to the modeler by eliminating the need to set up multiple basins and rather rely on the distributed hydrologic solution. In general, it is presented that the results of the DMIP still requires additional research to determine the benefits of a, or many, distributed modeling platforms over the traditionally utilized lumped modeling approach employed by the NWSRFS.

Reed went on to further the study area in 2007, presenting “A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations” (Reed et al., 2007). Proposed in that research was the application of a distributed hydrologic model results presented in the frequency realm as another method of preparing FFG products for WFO’s and flood forecasting. The driving force of this research is minimizing prediction bias through application of the Log Pearson Type III frequency transform. The research was conducted on gauged basins in Eastern Oklahoma and Western Arkansas and 7 of the 10 basins presented fit the definition of being flash flood basins. The approach presented is somewhat of a combination of the threshold runoff approach and the continuous distributed modeling studied as part of the DMIP in that the threshold frequency is predetermined for a catchment and the continuous distributed model develops forecast peaks that are then statistically post processed to develop a predicted frequency event. It is offered in Reed et al. (2007) that continuous distributed modeling inherently offers benefits over lumped parameter models through affording high resolution in both time and space which would better approximate those catchments considered flash flood basins. Further, the results of the post processed frequency based prediction helps with bias correction over the resultant flow magnitudes produced by the modeling.

The objective of this research is to evaluate the performance of a semi-distributed regional hydrologic model applied to the Long Island Sound Watershed for the prediction of flash floods. Evaluation will include comparing the flow predictions of the model at internal computation points of sub-basins to observed historical data. In addition, the research provides support in validation of the frequency prediction approach to flash flood forecasting (as described in Reed et al. (2007)) over the raw flow magnitude output from the semi-distributed hydrologic model by demonstrating the efficiency of predicting the probability of exceedance of flooding over attempting to determine the magnitude of flooding in relation to flood forecasting; effectively reducing modeling bias. The sensitivity of the spatial scale of study basins is presented, as is the temporal scale for both the flow and frequency results. Finally, the success rate of the forecast frequency is presented comparing not only the size of the basins but the time scale of the basins.

Following this introduction, section 2 of this paper gives a brief description of the study site including basin locations and regional climate overview. Section 3 presents the distributed hydrologic model platform used and the model of the Long Island Sound Watershed prepared. Section 4 presents the study methodology, including selection of the study catchments, applied error metrics, frequency transform and calculation of the post processed frequency success index. Results and conclusions from this research are presented in section 5 and 6.

2. Description of study site

The study area encompasses portions of southern New England including basins in south central and south western Massachusetts and throughout Connecticut. Elevations of the selected basins range from +2000 feet in the Berkshires in Western Massachusetts to 1000 feet in the eastern uplands of Connecticut/south central Massachusetts to 200 feet and less in the south central coastal lowlands of Connecticut.

The climate of Southern New England is moderate with average annual temperatures ranging from 46-48°F in Southwestern Massachusetts to 50-52 °F in Southwestern Connecticut based on the normal (30-year) period from 1981 to 2010. Seasonal averages are more extreme with the January average temperature in Southwestern Massachusetts between 20 and 24 °F and 27-30 °F along the shore in Connecticut. In July, average temperatures reach 68-70 °F in Southwestern Massachusetts and 72-74°F in Southwestern Connecticut. Average annual precipitation for the same normal period is between 47 and 50 inches per year, and tends to be evenly distributed with most months averaging 4-5 inches per month, and the winter months of December to February averaging slightly less at 3-4 inches per month. Yearly snowfall accumulation is more stratified throughout the region with the upper inland elevations in the Berkshires and northwestern hills of Connecticut averaging 50-70 inches per year, Central Connecticut and Southeastern Massachusetts averaging 40-50 inches per year and the coastal areas of Connecticut averaging much less at 20-40 inches per year. (Northeast Regional Climate Center, 2009).

3. Description of hydrologic platform and model used

The modeling system evaluated in this study is the Precipitation Runoff Modeling System (PRMS) developed by the United States Geological Survey (USGS). PRMS was developed as a distributed hydrologic modeling platform used to determine watershed hydrology and various aspects thereof. Watershed response to normal and extreme precipitation can be estimated and used to calibrate the model to observed events.

The basic construction of a PRMS hydrologic model is to subdivide the study watershed into "Homogeneous Response Units" (HRU[1]) which are areas in which hydrologic parameters are computed based on a given time step (typically a 24-hour period). Parameters include solar radiation, precipitation, base flow, land cover, topology, subsurface conditions etc., providing for a representation of an energy and water balance for the study area and for a given time step. Each HRU is interconnected through routing procedures taking into account overland flow and stream reach and reservoir or depression routing (Leavesley et al., 2006).

The PRMS model used in this study was developed by Bjerklie, Trombley & Viger (2011) herein referred to as Bjerklie Model, as part of an investigation assessing the potential for climate change impacts on groundwater recharge and snowfall in the Long Island Sound Watershed. The model simulates hydrologic processes of approximately 15,800 square miles of New England that drains to Long Island Sound, and includes the entire Connecticut River Watershed, the Thames River Watershed, the Housatonic River watershed, and the Southwest, South Central and Southeast coastal

[1] HRU is also commonly used as an acronym for "Hydrologic Response Unit". The meaning remains the same, i.e. a discrete area of homogeneous hydrologic parameters including land use and cover, soils, solar radiation, rainfall, etc.

watersheds of Connecticut. By understanding the distribution of various hydrologic parameters and simulating a water budget of the study area over a number of years with observed streamflow data for comparison and calibration, the goal of this research was to determine future trends in hydrologic processes taking into account estimates of climate change.

In preparing the Bjerklie Model, a variety of parameters for use in calibration were considered, such as subsurface flow, surface runoff, snowmelt, groundwater recharge, however these processes are difficult to observe on a regional scale and a comprehensive historical data set does not exist (Bjerklie et al., 2011). Instead, the assumption used for calibration was that if streamflow was accurately estimated (based on gauge records at 73 stations over a 46 year period) resulting from known precipitation, the evapotranspiration of the watershed(s) modeled was accurately represented. This approach does not discount the necessity of representing the sub-surface and groundwater runoff/storage processes in addition to the surface runoff and storage processes, rather it relies heavily on previously developed mathematical representations of these water balance components that are included in the modeling platform.

Bjerklie et al. (2011) demonstrate that the Long Island Sound Watershed model has a mean annual daily streamflow error of 4.4% for the 73 gauging stations observed, with maximum errors reported as high as ±30%. The greatest errors are associated with smaller watersheds and are likely due to the resolution of the model in general; local physical features and meteorological factors in the smaller watersheds, which are influential in the modeling results not being represented with enough precision. For the

Connecticut River at Thompsonville (the primary basin of the analysis) the mean annual streamflow has an overall error of 6%.

It is noted in Bjerklie et al. (2011) that the errors of simulated to observed data for the various gauges used for calibration tend to vary spatially and seasonally for both average annual flows and (more so) for the mean monthly flows. This observation indicates that there is an inherent problem with single parameter calibration to address overall model error. Again, local features of smaller watersheds may not be represented with enough resolution. There is a benefit to the single parameter calibration technique; other than simplifying the calibration process, evaluating results across the distributed model is more consistent, knowing the process utilized as opposed to attempting to qualify results from various calibration techniques across a regional model. The results of the Bjerklie modeling and study include also the observation that the model is capturing well the trends in streamflow both in time and space. This observation plays a key role in the study presented in this paper.

4. Study methodology

4.1 Study catchment selection

Preparing the research for this analysis included selecting USGS gauges and corresponding watersheds that fit a number of requirements. The gauge had to be within the limits of the Bjerklie Model and have a record of at least 10 years (following industry standard requirements for frequency analysis (IACOW, 1982)), encompassed by the time window used in the Bjerklie Model; water years 1961 to 2007[2]. In following with the calibration and evaluation of the Bjerklie Model, gauges with watersheds less than 20 square miles were not included in this study. Additionally, those gauges with watersheds which exhibited substantial flood storage availability were not included due to the potential for additional error associated with flood routing models. Substantial available flood storage is considered as those watersheds with 103 acre-feet of usable storage volume per square mile (Benson, 1963). Watersheds with less than 103-acre-feet of usable storage per square mile were found to impact peak runoff rates by less than 10%.

The gauged watershed divide was also a factor in selecting the stream gauge. Since the Bjerklie model was spatially set up with Hydrologic Response Units (HRU's) which may or may not be based on the actual watershed divides or known gauged locations, gauges and the corresponding watersheds were compared to the geometry of the HRU's used to ensure that the gauged watershed selected was reflected by even numbers of HRU's, which could then be combined to define the computed runoff for the given watershed at a corresponding outlet or node. Figure 1 illustrates an example of this.

In Figure 1 the black polygons represent the HRU's that make up the hydrologic model. The color shaded polygons are the watersheds that contribute flows to stream gauges recording the observed data. The Salmon River Watershed, Yantic River Watershed and the East Branch Eight Mile Watershed have one or more HRU's that encompass the entire watershed at the stream gauge. In contrast, the watershed of the

[2] A water year is defined as the period between October 1st and September 30th with the designating year as that calendar year of the ending year; i.e. water year 2007 begins October 1 2006 and ends September 30, 2007.

Eight Mile River stream gauge is much less than the associated HRU defined for the area, and as such, this stream gauge has not been included in this study.

In all 13 stream gauges were selected for use in this study; all of them meeting the study established requirements, and representing the spatial variability of Southern New England from Long Island Sound to Southern Massachusetts. The selected watersheds varied in size from 20 square miles to 210 square miles, and had an average record length of 40 years (minimum record length 12, maximum 46 years). Figure 2 illustrates the final selected watersheds and their locations within the study area.

4.2 Data evaluation metrics

Evaluation of the predicted against observed streamflow data for this study included a measurement of modeling bias, determining modeling skill and evaluation of the correlation between the simulated and observed data.

4.2.1 Bias

Modeling bias gives a measure of how close the prediction is to the observed event. To represent this relationship, the **M**ean **R**elative **E**rror is presented through a ratio of the difference between the observed data and the simulated data and the observed data, or:

$$\boldsymbol{MRE} = \frac{\sum_{t=1}^{n}(Q_{pr}(t) - Q_{ob}(t))}{\sum_{t-1}^{n} Q_{ob}(t)} \quad (2)$$

Where Q_{pr} is the predicted flow rate, Q_{ob} is the observed or measured flow rate, t is the time step and n is the number of time steps reviewed. A negative MRE indicates that the model under-predicts the event, a positive MRE over-predicts the event, and an MRE score of 0 would be perfect agreement of the prediction with the observed data.

4.2.2 Skill

The skill of the model in capturing individual events was evaluated through the relative Root Mean Square Error metric (${}_{r}RMSE$).

$$ {}_{r}\boldsymbol{RMSE} = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(Q_{pr}(t)-Q_{ob}(t))^2}}{\frac{1}{n}\sum_{t=1}^{n}(Q_{ob}(t))} \tag{3} $$

This error evaluation determines the departure of the observed data from the simulated data in relation to the average observed measurement. In this case, the lower the ${}_{r}RMSE$ value, the better the events are captured by the prediction, with a perfect fit expressed as ${}_{r}RMSE = 0$.

4.2.3 Correlation

Finally, the linear relationship between the observed and predicted values is presented by the Correlation Coefficient.

$$ \boldsymbol{r} = \frac{n\cdot\sum_{t=1}^{n}\left(Q_{pr}(t)\cdot Q_{ob}(t)\right)-(\sum_{t=1}^{n}Q_{pr}(t))\cdot(\sum_{t=1}^{n}Q_{ob}(t))}{\sqrt{n\cdot(\sum_{t=1}^{n}Q_{pr}(t)^2)-(\sum_{t=1}^{n}Q_{pr}(t))^2}\cdot\sqrt{n\cdot(\sum_{t=1}^{n}Q_{ob}(t)^2)-(\sum_{t=1}^{n}Q_{ob}(t))^2}} \tag{4} $$

This evaluation metric presents the relationship between the predicted and the measured values. The value of r will range between -1 and 1. An r value of 1 represents a perfect positive relationship where for every increase predicted, an equal increase is observed. Conversely, an r value of -1 indicates a perfect negative relationship, where for every increase predicted, an equal but opposite decrease is observed. An r value of 0 would indicate no relationship between the predicted and observed data points.

4.3 Data sets evaluated

4.3.1 Individual catchment evaluation

Evaluation or comparison of the modeling results to the observed data was completed on a range of spatial and temporal scales. Initially, each basin was evaluated independently. This provided insight on the correctness of the results at interior points of the distributed model. The predicted daily flow magnitudes or computed daily discharge rates were compared to the record of observed daily discharge rates.

4.3.2 Regionalized daily data

Moving from a localized or geographically specific analysis of data to a regional evaluation of the model, the interior observation nodes were lumped to evaluate the Bjerklie Model on a regional basis. This regionalizing approach was further refined to determine the impact of watershed size and streamflow magnitude on the modeling results. To evaluate the influence of watershed size, basins included in the study were qualitatively grouped into three classes; small (20-50 square miles), medium (50-100 square miles) and large (100+ square miles). To evaluate the influence of streamflow magnitude, that grouped population was further classified into quintiles (0-20%, 20-50%, 50-70% 70-90% and 90-100%) based on the amount of data to try and capture/evaluate the daily flow magnitude influence. This regionalized predicted daily flow dataset was then evaluated against the observed daily data to determine the modeling bias and skill as well as the relationship between the model results and historical flow measurements.

To evaluate the performance of the modeling results in a frequency domain, this original data set was prepared and assessed through a frequency distribution to assign recurrence intervals for each observed and predicted data point. For each stream gauge and corresponding model node, the yearly peak flows were determined and a frequency of occurrence was determined through a Log-Pearson Type III distribution analysis. The Log-Pearson Type III distribution is the method of flow frequency analysis recommended by the Interagency Advisory Committee on Water Data (IACOW, 1982) and used extensively in hydrologic design.

This data frequency transform method determines the probability of exceedance of flows by

$$\log Q_p = \overline{Q_L} + KS_L \tag{5}$$

where:

$\overline{Q_L}$ = Mean of the logarithms of annual peak discharges

Q_p = Flow magnitude

K = Frequency factor based on data skew determined for the data set

S_L = Standard deviation of the logarithms of the annual peak discharges

The frequency factor, K, is determined by the skew coefficient. This factor is dependent on the desired probability of exceedance and the computed skew of the data set given as

$$G = \frac{N^2(\sum X^3) - 3N(\sum X)(\sum X^2) + 2(\sum X)^3}{N(N-1)(N-2)S_L^3} \tag{6}$$

where:

G = Skew coefficient

N = Number of observations

X = Logarithm of the annual peak discharges

This frequency factor provides for the shape, or curvature of the resultant flow frequency distribution. With a calculated data skew coefficient of 0, the Log-Pearson Type III distribution is analogous to the Log Normal distribution. When the skew coefficient is negative, the flows outside of the 0.15 and 0.85 exceedance probability will be lower when compared to a log-normal distribution and with a positive skew the discharges for the same range will be larger as compared to the log-normal curve (Oberg and Mades, 1987).

The Log-Pearson III data transform was completed using the Army Corps of Engineers' Hydraulic Engineering Center's Statistical Software Package (HEC-SSP, version 2.0). This statistical software was developed to process hydrologic data, and as part of its function, can be used to determine flood flow frequency curves from given annual peak flow data.

Once a frequency distribution curve was developed for each data set based on the calculated probability of exceedance values, the original observed daily gauged data and the Bjerklie Model predicted daily flow rates were transformed to the frequency domain based on the curve developed for each data set. The individual daily values of observed and predicted flows were linearly interpolated between data points of the developed frequency curve to assign each observation and prediction a probability of exceedance.

The comparison between observed and predicted *frequencies* was completed just as with the daily flow magnitude analysis; determining the mean relative error to understand the bias between the predicted frequencies and the observed frequencies, then determining the skill of the model in capturing the individual events, and understanding the correlation between the two data sets to see if the natural trends of the observed data are reflected in the model.

A comparison between the magnitude domain and the frequency domain can now be conducted to determine the usefulness of modeling recurrence intervals in lieu of absolute flow values for the purposes of flood forecasting.

4.3.3 Temporal resolution of the regionalized data

In Bjerklie et al. (2011) it was indicated that the temporal resolution of the Long Island Sound watershed afford data evaluation on a monthly time step. In light of this potential modeling constraint, additional data manipulation in this study included evaluating the data on a 10-day averaged flow and a monthly averaged flow. This was completed by simply averaging the daily flow data sets, both observed and predicted, in consecutive 10-day increments and monthly increments. All other analysis procedures remained as before, with performance metrics applied evaluating the predicted averaged flows and re-defined frequencies to the observed averaged flows and re-defined frequencies, respectively.

4.4 Success rate of the frequency forecast

With the frequency transform of the mean daily, 10-day averaged and monthly averaged data sets, a flood prediction success rate could be assessed using the Critical Success Index (CSI) method. The CSI method is discussed in Schaefer (1990) as a forecasting success rate indicator and is used extensively by the National Weather Service to verify and score results of forecasting models. This success indicator statistic is rooted in the premise that the prediction can have either a yes or no alternative, or a hit or miss scenario. An additional assumption inherent in this metric is that when an event was not expected or predicted, it has no influence in the forecasting success result.

To apply this statistic, the predicted and observed events are compared under a governing threshold value, say for instance the 5-year event or 20 percent probability of exceedance. A hit, or positive success event, would be if the prediction and observation were to meet or exceed the threshold of a 5-year event. A miss, or negative success event, would be if an observed event exceeded the desired threshold that the forecast did not predict, and a false alarm would be if the forecast predicted an event at or above the threshold which did not have an associated observed event. A fourth category in this data examination would be a “no hit” group, which those events that did not exceed the threshold of interest, nor was an event predicted.

Once each observation/prediction couple is categorized as above, the success rate of the forecast model for the threshold of interest is a ratio of the number of hits (H) to the sum of hits, misses (M) and false alarms (FA). Success of the forecasting is directly related to the number of hits and is negatively impacted by the number of misses and false alarms.

$$\boldsymbol{CSI = \frac{H}{H+M+FA}} \tag{7}$$

To evaluate the success of the frequency forecasting of this study, the data sets were analyzed for each relative watershed size, taking into account the entire population.

In addition, a fourth success test was conducted to determine if or what impact the relative size of the watershed has on the success of the forecast by eliminating the relative size sub-grouping. Thresholds for success were set following standard industry design events; 1-, 2-, 5-, 10-, 20-, 25-, and 50-year. The thresholds were truncated at the 50-year because as of this research, there was no success above this return period.

5. Results

5.1 Individual catchment evaluation

The first step in evaluating the performance of the internal solutions from the Bjerklie Model as compared to the observed stream gauge data included comparing the mean daily discharge for each gauge as reported by the USGS with the results of the Bjerklie model on a daily time step. This was a basin specific comparison which provided insight on the correctness of the distributed modeling results. At this point of the study, only computed and observed discharge rates were compared. The average mean relative error of the thirteen basins is -0.13, indicating that overall the model results were slightly biased to be less than that of the observed results. The maximum departure from observed was indicated by a relative error of -1.79, and the minimum departure from the observed was a relative error of 0.03. From this it appears that the model does a sufficient job of predicting the overall magnitude of daily stream flow rates for water years 1961 – 2007.

The accuracy or skill of the model in representing the flow rate on a daily time step was also reviewed through the use of the Relative Root Mean Square Error ($_r$RMSE). This statistic allows for the determination of the accuracy of the model in predicting the magnitude of the flow rate for a specific time step of the model and the gauge, respectively. The average $_{r}$RMSE for the 13 gauged basins is 1.30 (130%), indicating that in modeling the basins, the model is not providing results better than the average of the total population of observed values for the daily time step. The range of the skill statistic for the thirteen basins was 74% to 291%.

For this basin specific comparison, the correlation coefficient averaged for the thirteen basins is 0.59, indicating that the model does show a definite positive correlation between the observed and predicted discharge rates; the trend of the observed data is reflected in the prediction.

While the magnitude of flows generated by individual basins may agree with those observed at the gauges, the model does not do well in predicting the phasing of the observed and predicted flow values for the daily time step at the individual internal observation points selected for this study.

5.2 Regionalized daily data

Evaluating the model on a regional scale was completed through lumping and categorizing the results of the model based on the corresponding measurement at the observed time step. Figures 3 through 5 present the mean relative error, relative root mean square error and correlation coefficient of the predicted flow data to the observed, respectively. The figures are categorized to separate the small, medium and large watersheds (relative size) as well as break up the population into quintiles of 0-20%, 20-50%, 50-70%, 70-90% and 90-100% of the data set.

It can be seen that the bias in the model (represented by the MRE) is greatly reduced for the greater flow rates within the 50-90% population percentages, and the bias does not seem to be greatly affected by watershed size. The lower flows in the 0-20% population sub-set have the greatest relative error, which is likely due to the high amount of lower flow events that occur in the data. Additionally, it is shown that the greatest flows in the upper 10% of the population begin to increase in relative error, likely to the limited number of extreme events available for comparison.

Similar to the mean relative error, the accuracy of the daily flow modeling results (reflected in the $_{r}$RMSE) indicate that the higher population sub-set; the 50-90% flows are better captured by the modeling than the more frequently occurring subset of the predicted data.

Figure 5 depicts the correlation of the predicted data to the observed data. From this figure, it is clear that the strength of the correlation between data sets increases with the relative size of the watershed. This shows that the Bjerklie Model is more likely to capture the actual streamflows with the larger watersheds.

Figure 6 shows the results of the Mean Relative Error for the streamflow frequency data set. It is clear from this chart that in the frequency domain, the bias of the model is reduced as compared to the magnitude of flows determined prior and presented in Figure 3. This reduction is noticeable for the data as a whole, but for the population between 50-and 100% of the data, there is a marked decrease in bias.

In difference to the bias comparison of Figure 6, Figure 7 shows that the frequency data does not do as good of a job representing events, and thus the skill of the frequency data is decreased as compared to the flow data.

Figure 8 shows the correlation of the predicted frequencies to the observed. Similar to the average daily flow magnitude prediction analysis, this figure still indicates that the correlation between observed and predicted frequencies increases with increasing watershed size, however, there is a decrease in correlation in the medium sized watersheds that is not reflected in the flows magnitude analysis presented earlier. Overall, it does appear that correlation is much weaker for the daily streamflow frequency dataset as compared to the prediction of the flow magnitude, and in fact a large portion of the medium sized watersheds population is shown to be negatively correlated.

5.3 Temporal resolution of the regionalized data

The daily data sets, both predicted and observed were refined to create two new temporal domains; the 10-day averaged and monthly averaged flow data sets. This was completed in order to determine if the bias, skill and correlation of the model results could be improved. For the flow or magnitude domain, there was only a marginal improvement to the bias as represented by the mean relative error, when comparing across the three presented resolutions. This is shown in Figure 9.

The skill as represented by $RMSE_r$ is also influenced by the temporal resolution of the analyzed data set, however, similar to the mean relative error, the skill is only marginally improved with the decreasing time resolution as shown in Figure 10. The same marginal improvement can be seen in the correlation coefficient shown in Figure 11.

To evaluate the data sets in the frequency domain, each averaged set was transformed independently through frequency analysis of the yearly peak flows. The observed and predicted frequency data for both the 10-day averaged and the monthly averaged data sets were then compared to the daily averaged data to determine if a coarser time resolution increased the efficiency of the frequency predictions.

It is clear in comparing Figure 12 to Figure 9 that by decreasing the temporal resolution of the data sets and transforming the data to the frequency domain, the modeling bias is greatly reduced. Interestingly, between the daily, 10-day and monthly averaged data sets in the frequency domain, the 10-day averaged predictions appear to have the lowest mean relative error, indicating the least modeling bias.

Similar to the mean relative error comparison, Figure 13 shows that by decreasing the temporal resolution of the frequency data, the skill of the model appears to increase overall, with the greatest increase being apparent for the 10-day averaged data set. In difference to the observations of the prediction bias, when comparing the $RMSE_r$ of the predicted flows (chart 8) to the predicted frequencies, it appears that the flow magnitude results of the modeling better reflect the observations, though not drastically.

Figure 14 continues to show the trend of increasing reliability in the frequency prediction by decreasing the temporal resolution of the data, and that the correlation of the predicted frequency to the observed frequency is greater for the larger watersheds. Overall correlation of the predicted frequencies to the observed appears to be similar for the 10-day averaged and monthly averaged data sets. Additionally, the correlation of the predicted frequencies is weaker than that shown with the flow magnitude data sets of Figure 11.

The results above appear to show that the temporal resolution of the modeling has a great impact on the reliability of the predictions. As the resolution in the data set is decreased, the modeling better predicts the observations. Further, in transforming the flow magnitudes to a frequency domain through a standard Log-Pearson Type III distribution, inherent errors in the mathematical model due to assumptions and initial conditions of the rainfall/runoff relationship can be effectively “washed out”, bringing the observed peaks and the predicted peaks closer together, decreasing the bias of the predictions. The results further show that the selected method of transform from flow to frequency may have a detrimental, though not significant impact on the reliability of the predicted frequencies. The skill of the magnitude predictions is marginally better than that of the frequency predictions. One additional noteworthy observation in the results of the skill assessment of the model and frequency analysis results is that though the errors do seem high on the surface, this is not unique to this study. In Reed et al (2007), skill errors were also reported high (between 100 and 300%) for the results of the distributed modeling. This is similar to the results presented herein, though a magnitude to frequency error comparison was not made in the Reed et al (2007) study, as that portion of the research was focused on the bias reduction from frequency transform.

5.4 Success rate of the frequency forecast

Figures 15-17 illustrate the results of the CSI for the frequency predictions of the extreme events based on threshold for the daily, 10-day averaged and monthly averaged data sets. As can be expected, the success of predicting events correctly decreases with lower exceedance probability/higher recurrence interval. This can be seen for all three temporal resolutions. It can also be seen through the sequence of figures above that success increases with lower temporal resolution; the monthly averaged success rate is greater than the 10-day averaged success rate is greater than the daily averaged success rate. As far as relative watershed size in relation to prediction exceedance success, it appears that for the daily and 10-day averaged data sets, there is a small difference in success rates. For the monthly data, the difference in relative basin size is more marked with the large modeled basins appearing to more correctly predict threshold events.

6. Conclusions

Refining and improving flood forecasting techniques is a focus of the World Meteorological Organization (WMO), recognizing that socio-economic shifts as well as a changing climate are placing more people and property in harm's way. In recent decades, flood risk mitigation has shifted from localized physical flood mitigation to regional practices and management of watersheds as a whole. A large part of that effort includes developing a flood forecasting system to help with the watershed management and life and property protection strategies (WMO, 2011). Improving on flood forecasting is also a focus of the National Oceanic and Atmospheric Administration National Weather Service. It has been recognized that in recent history, hydrologic science, computational technologies and remote sensing techniques have been developed to a point that facilitates higher space and time resolution in hydrologic and meteorologic modeling which, when applied appropriately, could increase the success and even lead time in flood prediction (Smith et al, 2004).

A strategy of flood forecasting was presented in this paper, attempting to use frequency analysis to improve the results of a hydrologic model and determine the impact of temporal resolution of the model. The hydrologic model presented was loosely calibrated to streamflow using evapotranspiration as a calibration parameter and assuming that the other hydrologic processes (e.g. groundwater interchange, sub-surface flow) are adequately represented through mathematical models developed through past empirical studies.

Basins selected for this analysis included those with more than 10-years of observed streamflow within the same time window used in the Bjerklie Model and those that are not appreciably impacted through regulation and/or flood protection measures to minimize possible sources of error. At the onset, the model was run on a daily time step and the results compared to the observations of the same time frame. It was found through the daily time step analysis that in transforming the flow magnitudes to a frequency domain, the prediction bias was reduced as compared to the observed data, with the error reduction greater for the higher magnitude subsets, and that this bias was relatively standard across the relative watershed sizes presented. Similar results of bias reduction are also shown in Reed et al (2007).

While the bias of the modeling tends to be reduced when transformed to the frequency domain, the skill of the model, as represented by the Relative Root Mean Square Error ($RMSE_r$), was shown to decrease for the daily time step. This indicates that the individual events predicted by the flow magnitudes are not captured as well when the observed and predicted data is analyzed in the frequency domain. Similar to the $RMSE_r$, the correlation of the predicted to the observed daily data is weaker in the frequency domain as compared to the modeling results of flow magnitudes.

Considering the temporal resolution of the data set, the results showed an increase in modeling effectiveness when the resolution is decreased in the frequency domain. Modeling bias showed a marked improvement in the 10-day averaged frequency results, and again in general, for the daily and monthly averaged frequency results. Comparing the model skill over the three time step resolutions also showed that the 10-day averaged frequencies perform better than the daily and monthly results, though the $RMSE_r$ error in the frequency domain is slightly greater than that of the skill in the magnitude predictions. Similar to the trends shown for the $RMSE_r$, the correlation of the predicted frequencies to the observed increases with decreasing time step resolution, but is weaker than the correlation of the predicted and observed magnitudes.

A future consideration for expanding upon this work would be to reassess the frequency transform method selected. The frequency transform method used on the observed data was based on the Log-Pearson type III distribution. This distribution was applied to each selected model node/gauge location taking into account only the data set produced at that point. The data forced through that distribution was used to develop a skew coefficient termed the "station skew" which provides for the degree of curvature for the resultant flow frequency curve. This station skew is highly sensitive to extreme events contained in the data set. In "Guidelines for Determining Flood Flow Frequency" commonly referred to as "Bulletin 17B" prepared by the Interagency Advisory Committee on Water Data (IACOW, 1982), an additional skew coefficient has been developed taking into account regional conditions to better refine the curvature of the distribution. This "generalized skew coefficient" is used to weight the station skew developed from the peak flood data set. This approach of weighting the station skew based on a regionalized skew coefficient is a more rigorous approach of frequency curve determination than that of only addressing the data at the observation point. However, it has been found to decrease the influence of the extreme events at the station and provide additional confidence in the resultant distribution curve. Building on the work presented in this paper, the frequency transform method could incorporate the weighting of the station skew with the generalized skew coefficient as recommended in “Bulletin 17B”. Metrics could then be performed again to determine the success of modeling in the frequency domain and re-assess the efficiency for flood prediction as compared to the flow magnitude results of the hydrologic model.

7. References -

Benson, M.A., (1962), Factors influencing the occurrence of floods in a humid region of diverse terrain: U.S. Geological Survey Water-Supply Paper 1580-B, 64 p.

Bjerklie, D. M., Trombley, T. J., & Viger, R. J. (2011). Simulations of historical and future trends in snowfall and groundwater recharge for basins draining to Long Island Sound. *Earth Interactions*, *15*(34), 1-35.

Carpenter, T. M., Sperfslage, J. A., Georgakakos, K. P., Sweeney, T., & Fread, D. L. (1999). National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *Journal of Hydrology*, *224*(1), 21-44.

Georgakakos, K.P., Carpenter, T.M., Cramer, J.A., Sperfslage, J.A., Sweeney, T.L., & Fread, D.L., (1993): A national system for threshold runoff estimation. *Proceedings of International Symposium on Engineering Hydrology, San Francisco, California*, *25-30 July 1993,* American Society of Civil Engineers, New York, 952-957

IACOW, (1982). Guidelines for determining flood flow frequency: Bulletin 17B of the Hydrology Subcommittee, US Geological Survey, Reston, VA

Leavesley, G. H., Markstrom, S. L., & Viger, R. J. (2006). USGS modular modeling system (MMS)-precipitation-runoff modeling system (PRMS). *Watershed models*, 159-177.

Northeast Regional Climate Center (2009) climate normals mapping retrieved October 12, 2014, from http://www.nrcc.cornell.edu/page_northeast.html

Oberg, K. A., & Mades, D. M. (1987). Estimating generalized skew of the log-Pearson Type III distribution for annual peak floods in Illinois (No. WRI-86-4008). United States Geological Survey.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Se, D. J., & DMIP Participants (2004). Overall distributed model intercomparison project results. *Journal of Hydrology*, *298*(1), 27-60.

Reed, S., Schaake, J., & Zhang, Z. (2007). A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of Hydrology*, *337*(3), 402-420.

RFC Development Management Team, 2003: Flash Flood Guidance Improvement Team—Final report. River Forecast Center Development Management Team Rep. to the Operations Subcommittee of the NWS Corporate Board, 47 pp. [Available online at http://www.nws.noaa.gov/oh/rfcdev/docs/ffgitreport.pdf.]

Rosgen, D. L. (1994). A classification of natural rivers. *Catena*, *22*(3), 169-199.

Schaefer, J. T. (1990). The critical success index as an indicator of warning skill. *Weather and Forecasting*, 5(4), 570-575.

Smith, M. B., Seo, D. J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q.-Y., Moreda, F. & Cong, S. (2004). The distributed model intercomparison project (DMIP): motivation and experiment design. *Journal of Hydrology*, *298*(1), 4

World Meteorological Organization (WMO), 2011. Manual on flood forecasting and warning. Geneva, Switzerland.

List of Figures

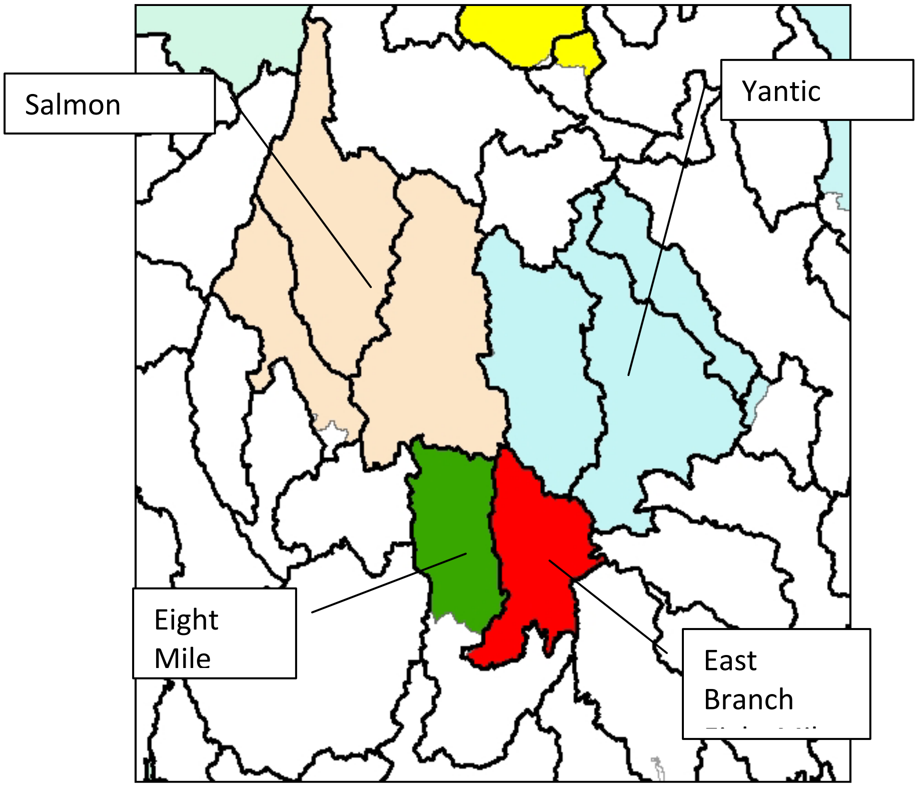

Figure 1. **A comparison of HRU delineations (black polygons) with the respective watershed boundaries (color shaded).**

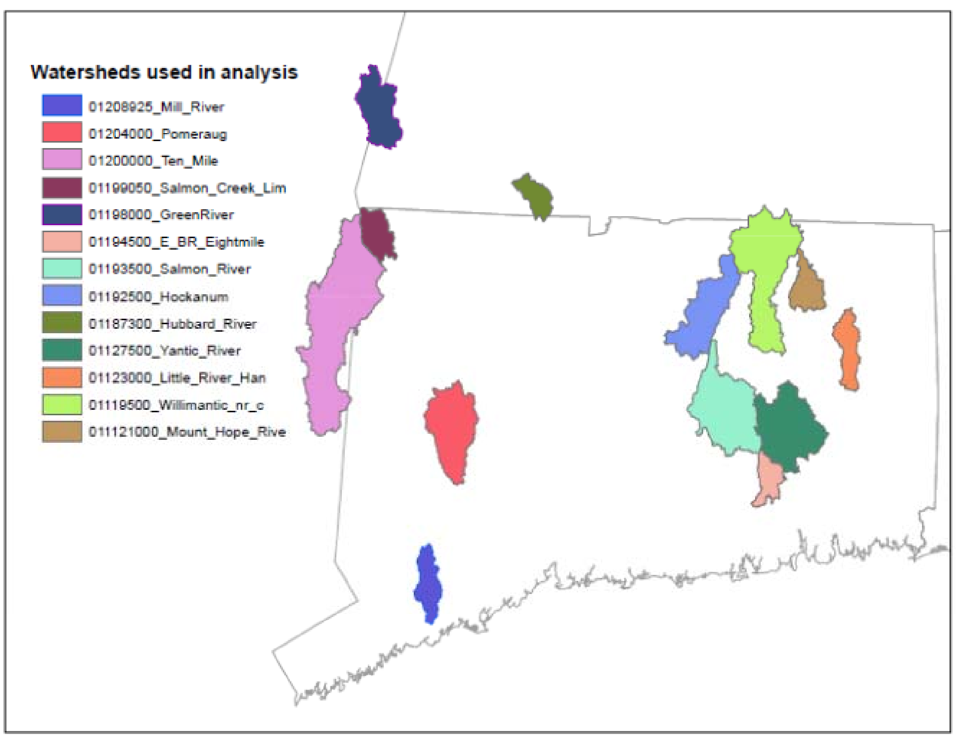

Figure 2. Watersheds used in analysis

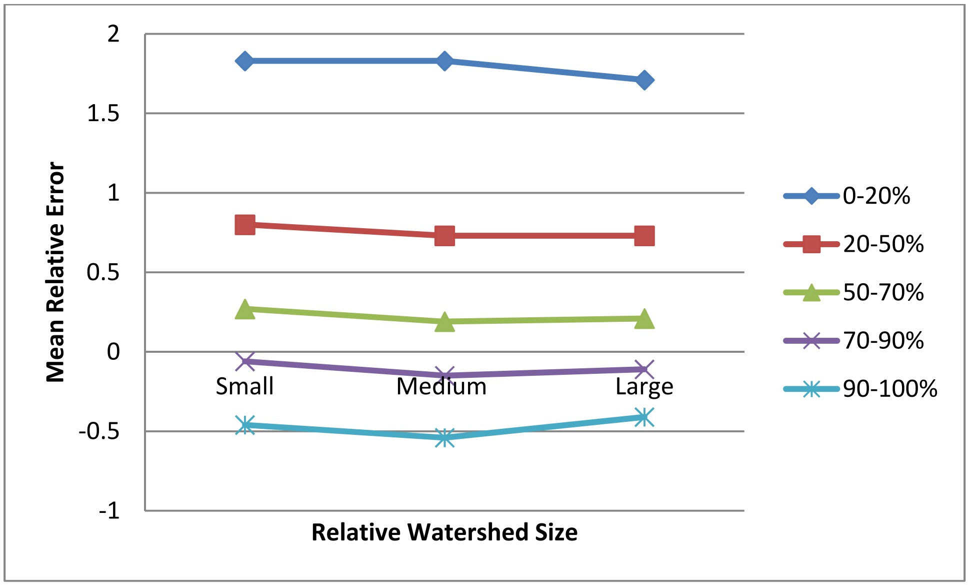

Figure 3. Mean Relative Error by watershed size and proportions of population for daily streamflow

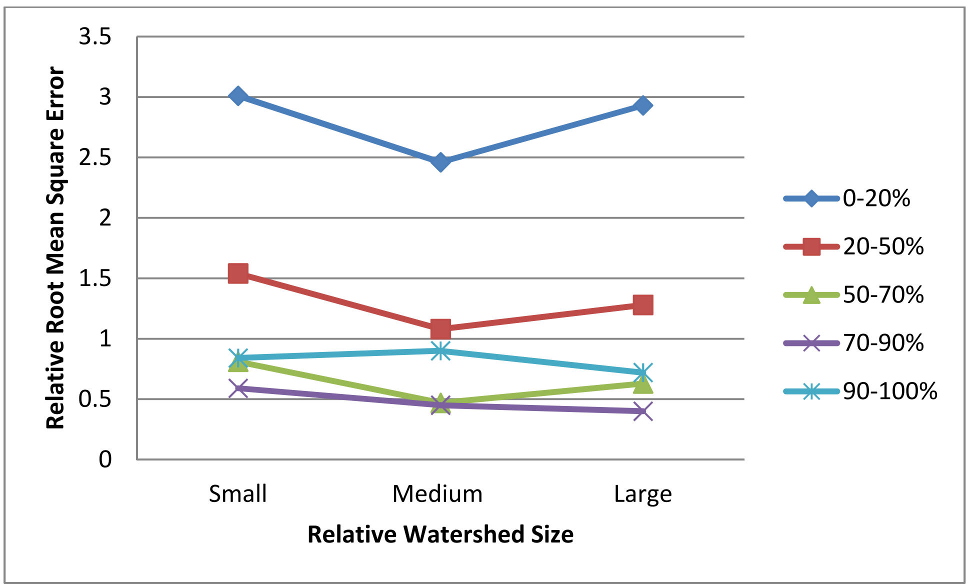

Figure 4. Relative Root Mean Square Error by watershed size and proportions of population for daily streamflow.

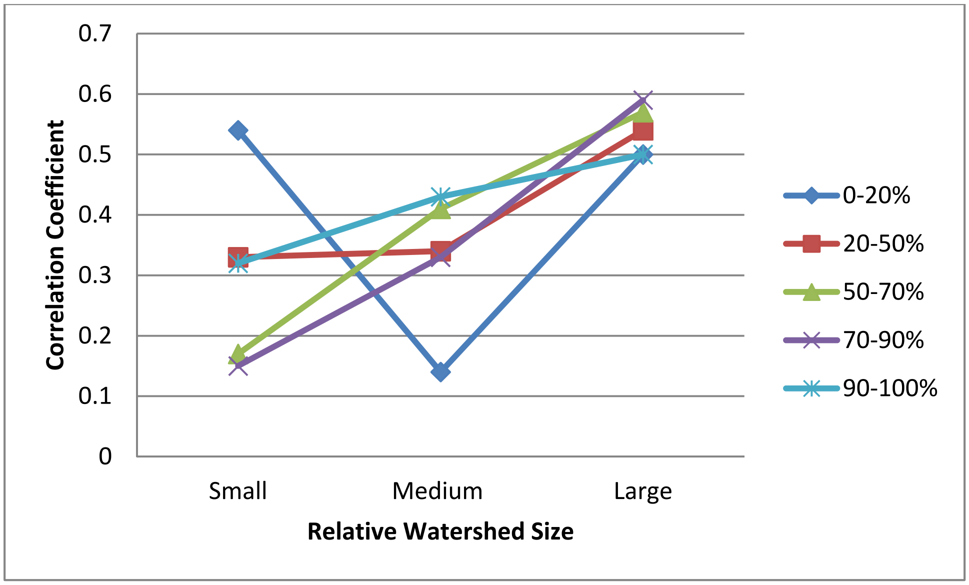

Figure 5. Correlation Coefficient by watershed size and proportions of population for daily streamflow.

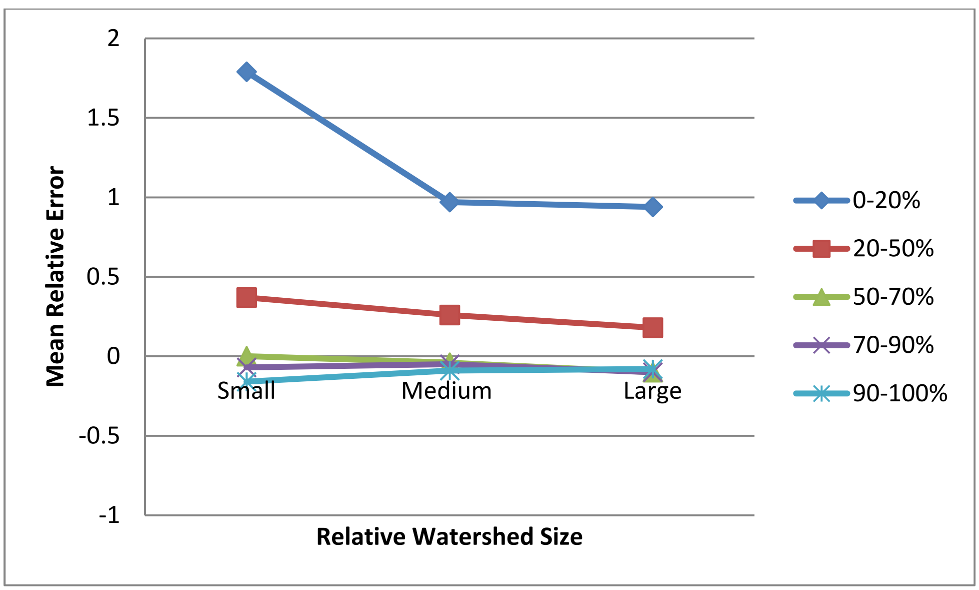

Figure 6. Mean Relative Error by watershed size and proportions of population for daily streamflow frequency

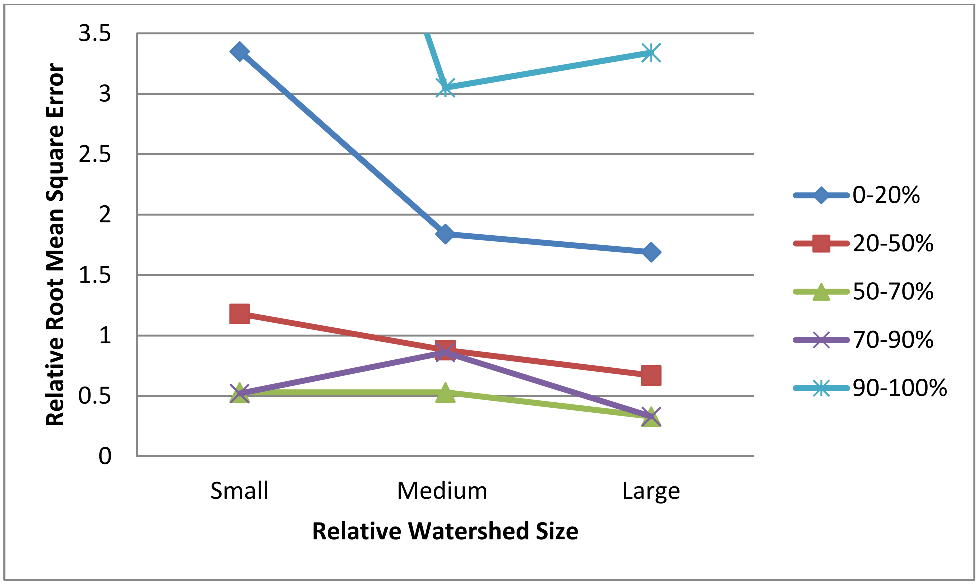

Figure 7. Relative Root Mean Square Error by watershed size and proportions of population for daily streamflow frequency

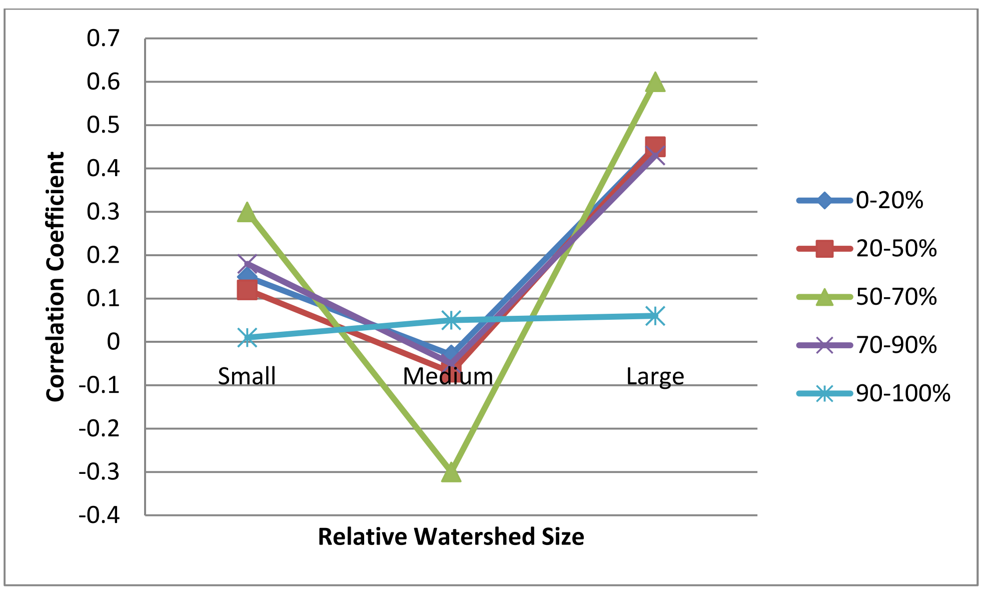

Figure 8. Correlation Coefficient by watershed Size and proportions of population for daily streamflow frequency

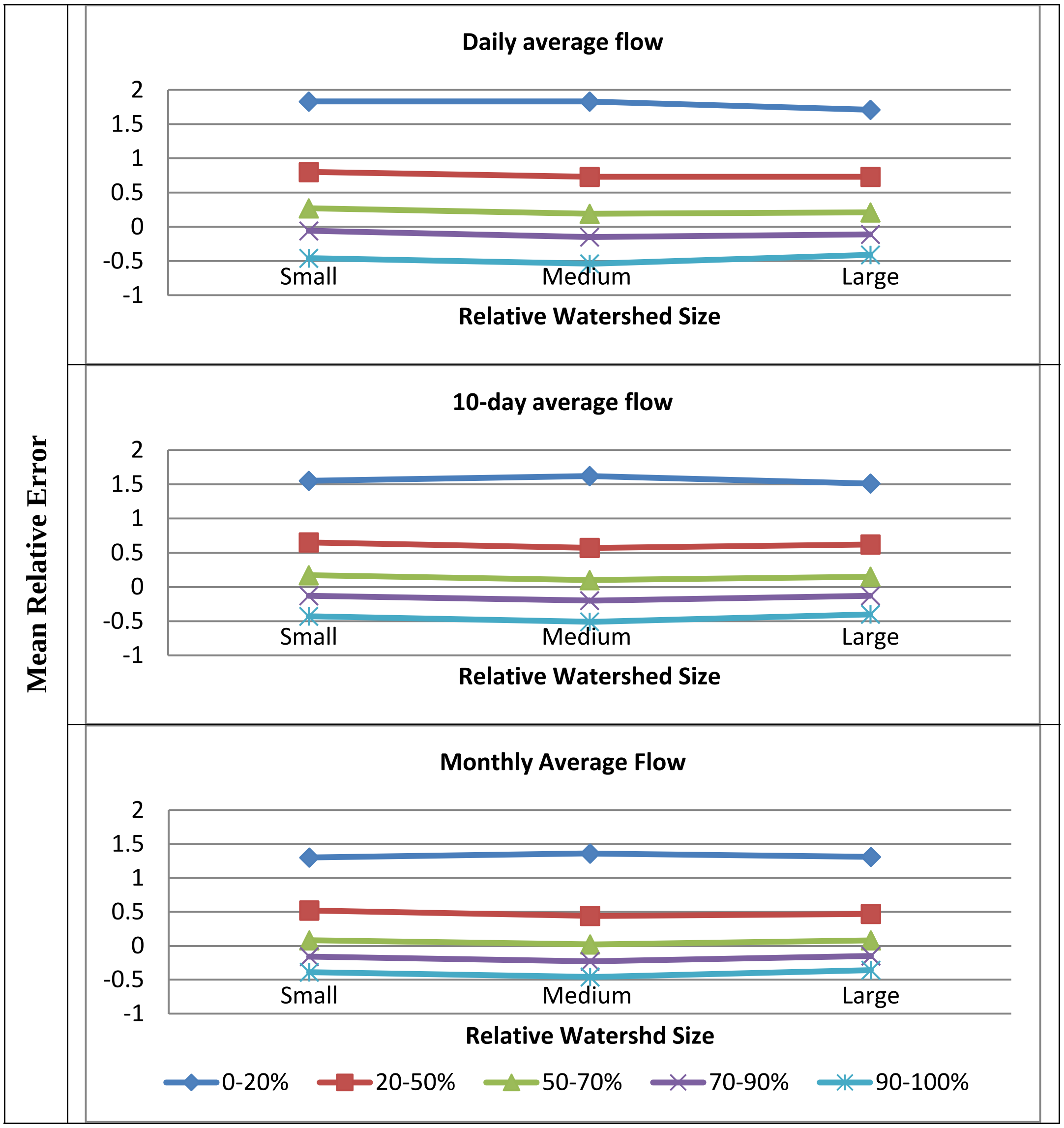

Figure 9. Mean Relative Error by watershed size and proportions of population for daily, 10-day averaged and monthly averaged streamflow.

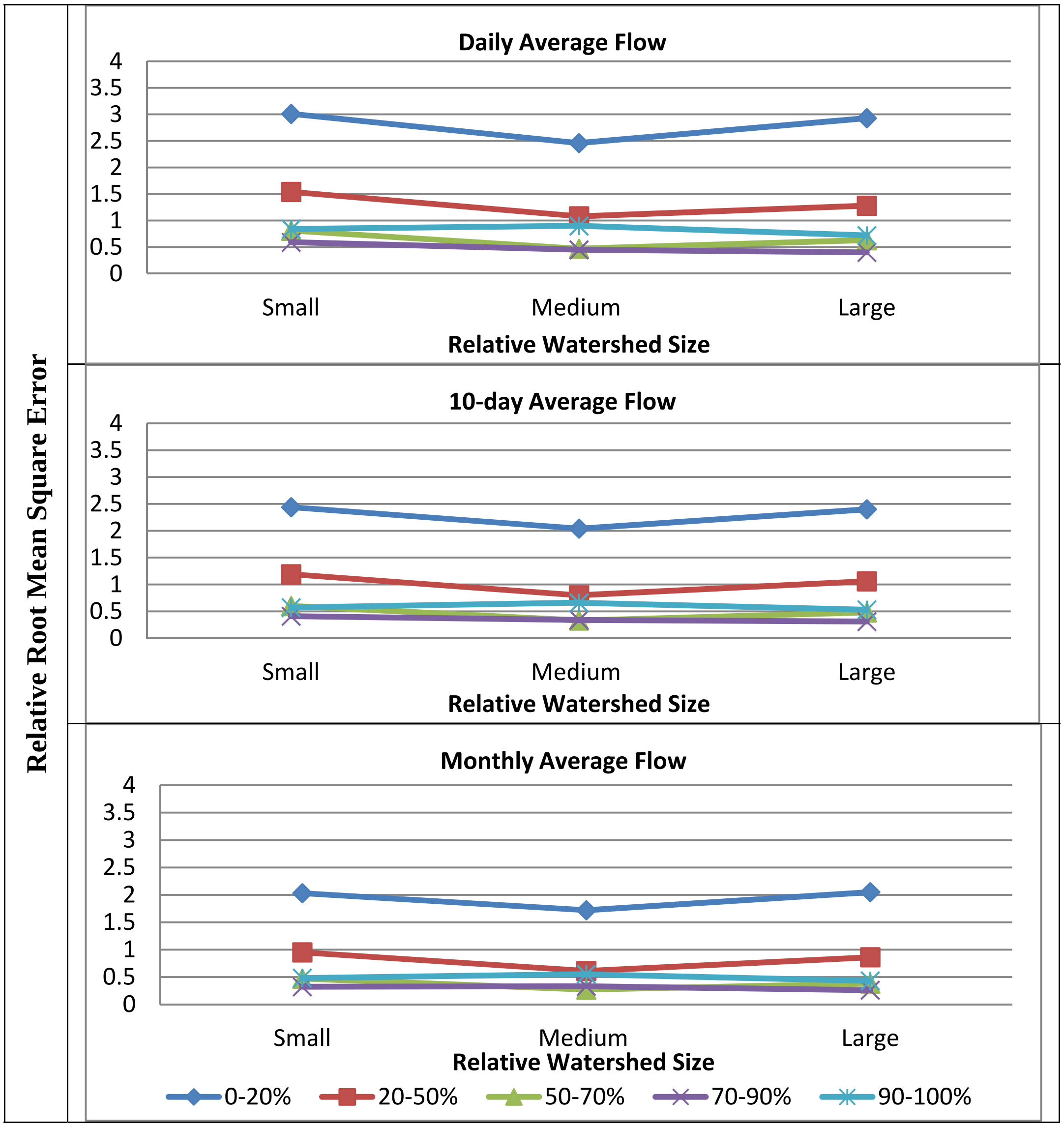

Figure 10. Relative Root Mean Square Error by watershed size and proportions of population for daily, 10-day averaged and monthly averaged streamflow.

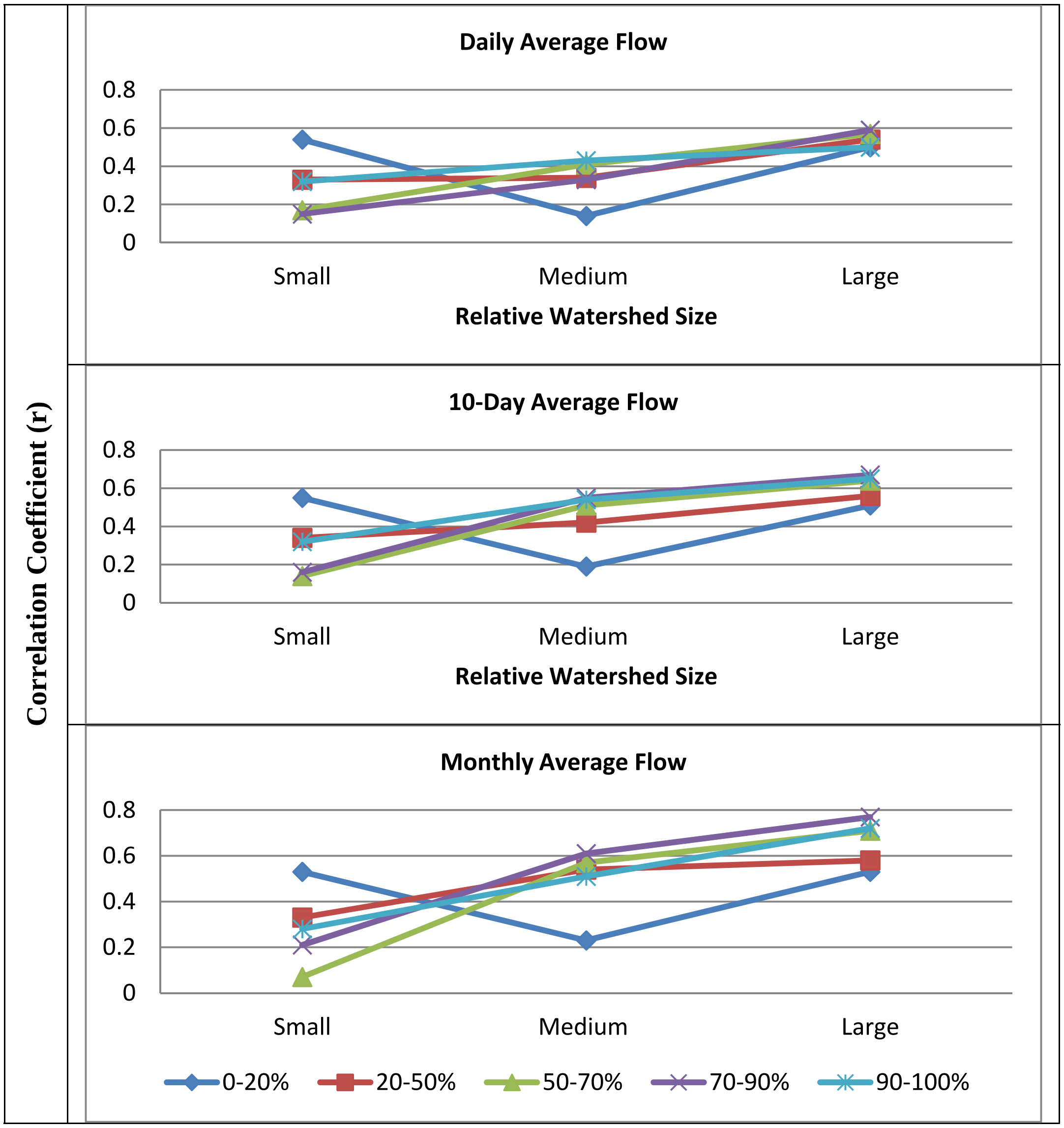

Figure 11. Correlation Coefficient by watershed size and proportions of population for daily, 10-day averaged and monthly averaged streamflow.

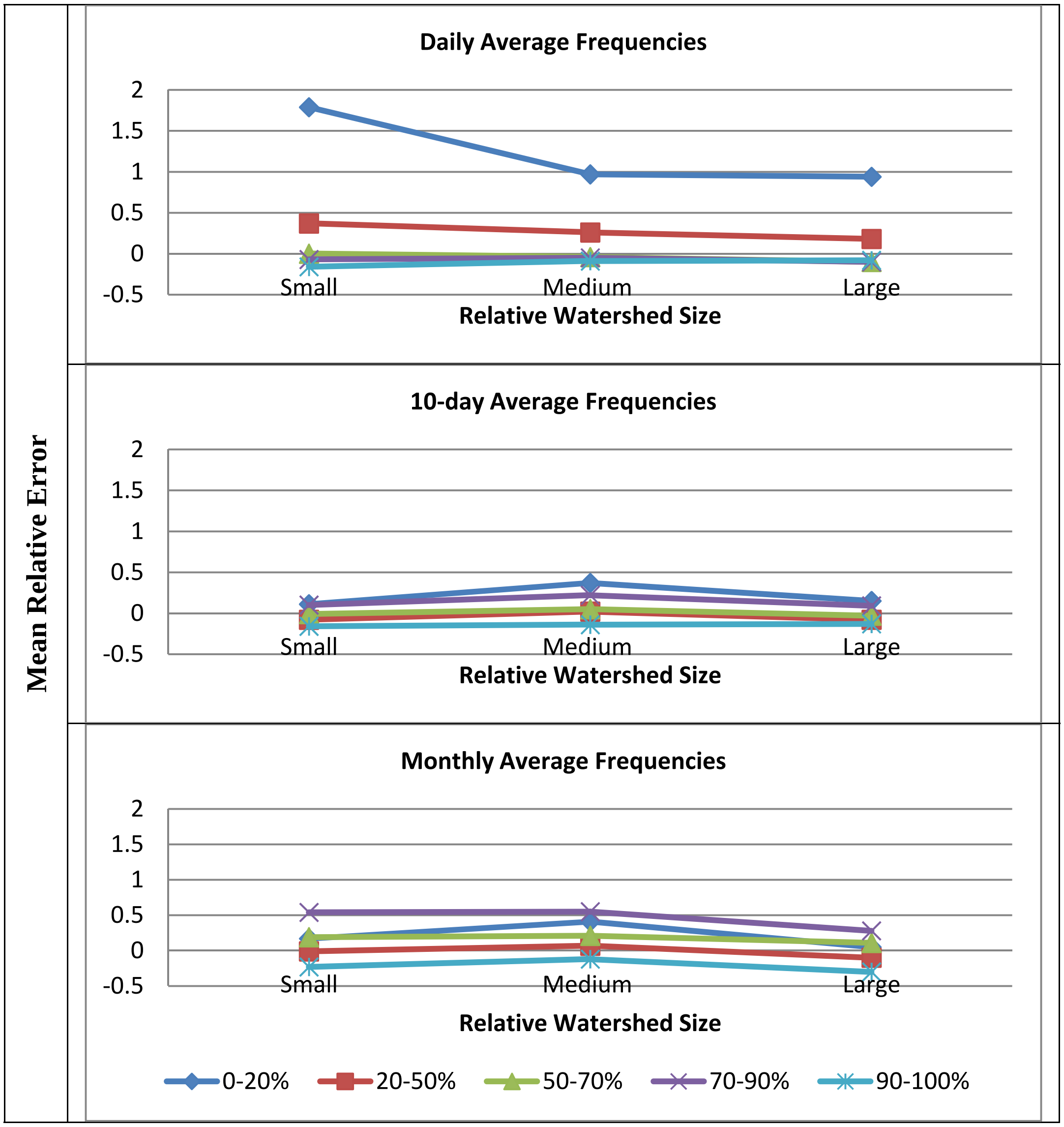

Figure 12. Mean Relative Error by watershed size and proportions of population for daily, 10-day averaged and monthly averaged streamflow frequency.

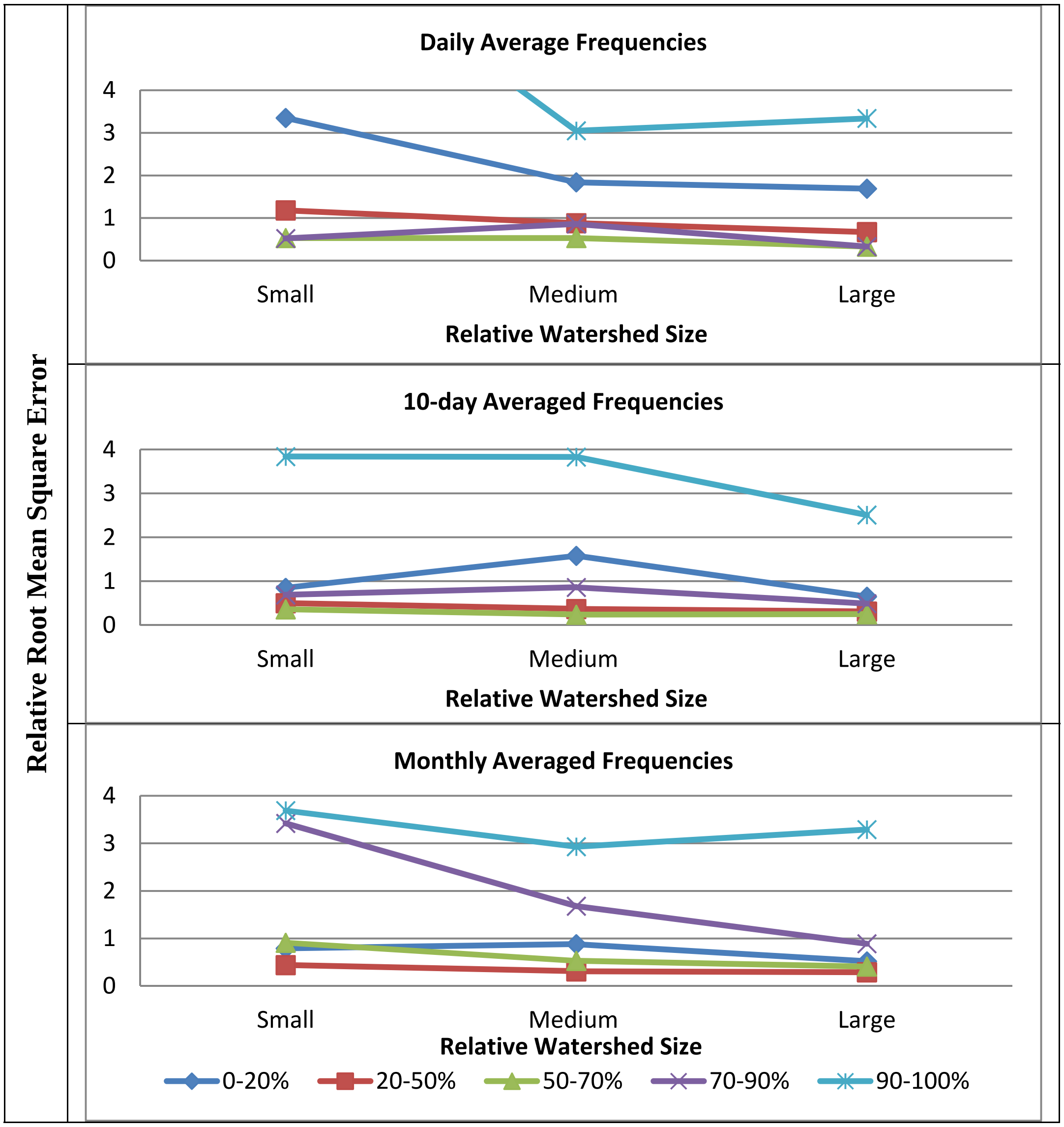

Figure 13. Relative Root Mean Square Error by watershed size and proportions of population for daily, 10-day averaged and monthly averaged streamflow frequency.

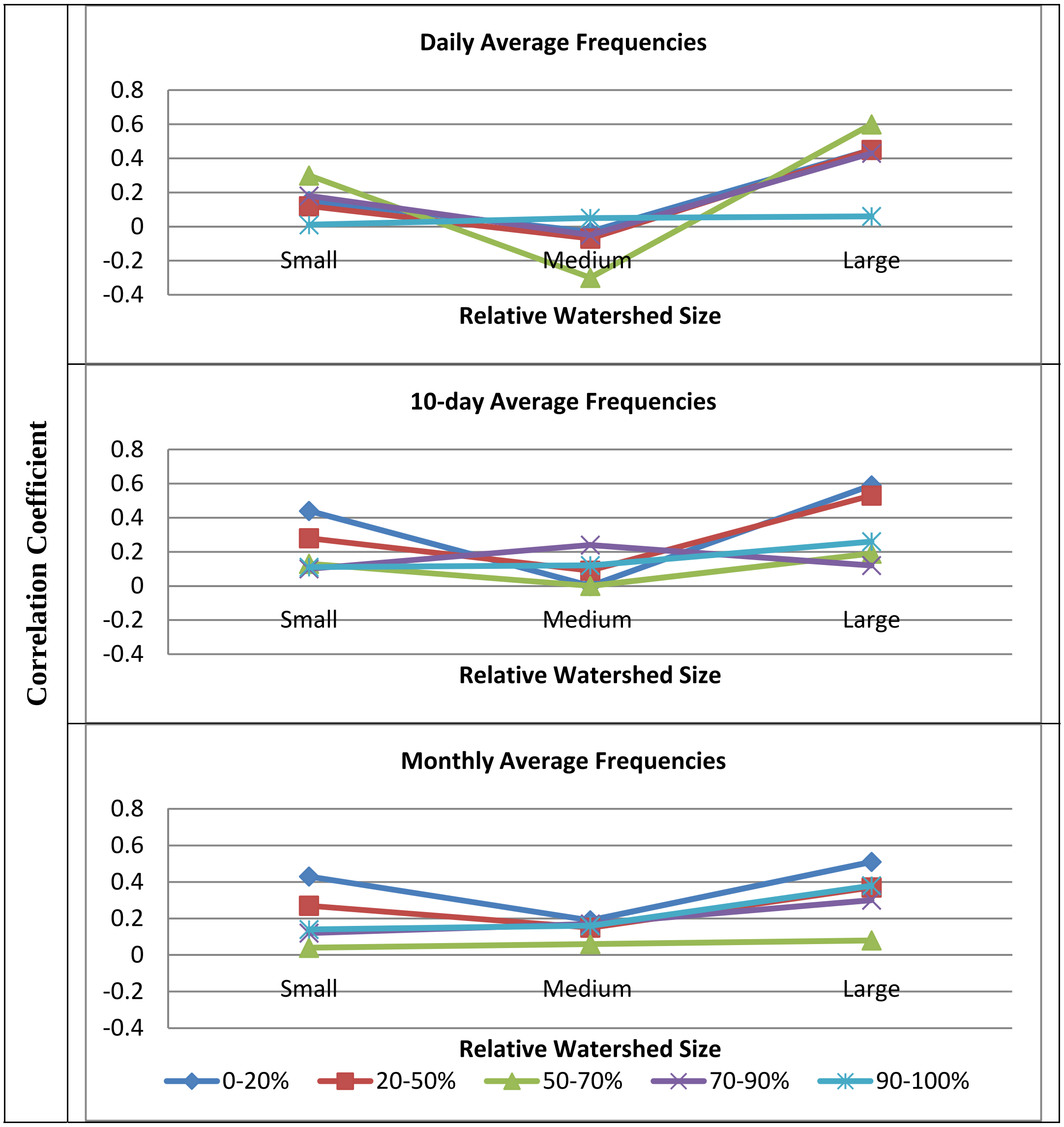

Figure 14. Correlation Coefficient by watershed size and proportions of population for daily, 10-day averaged and monthly averaged streamflow frequency.

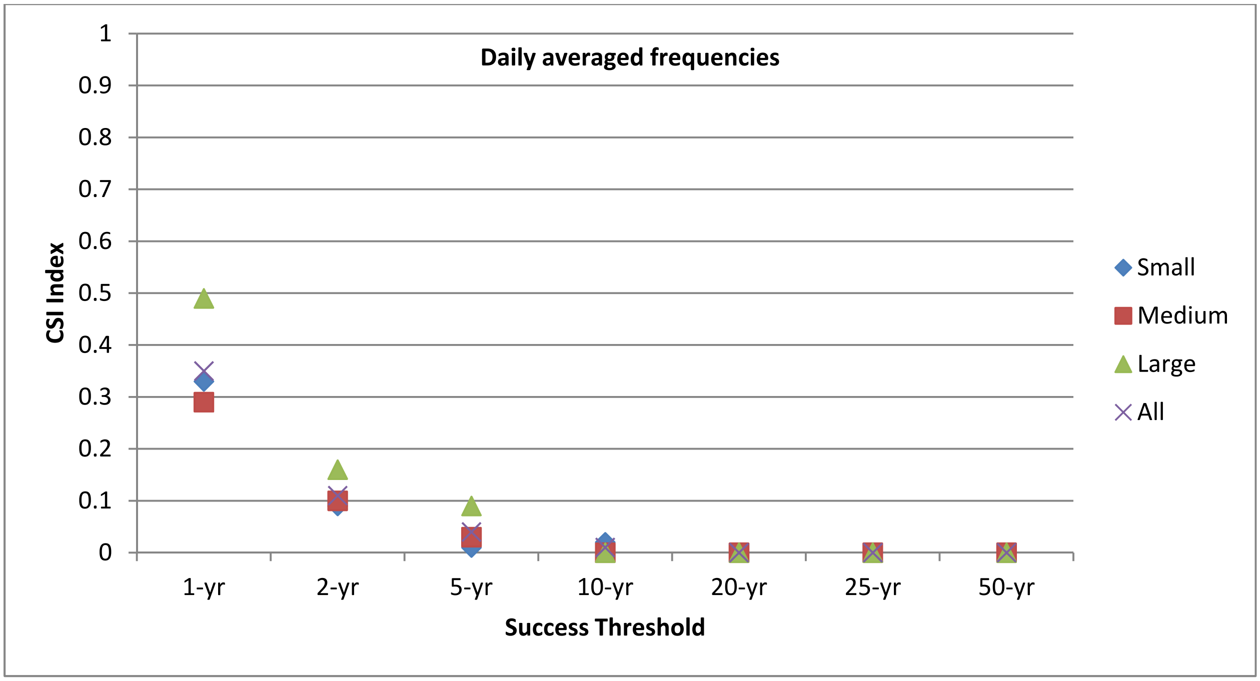

Figure 15 – Success rate of event prediction for the daily averaged frequency results.

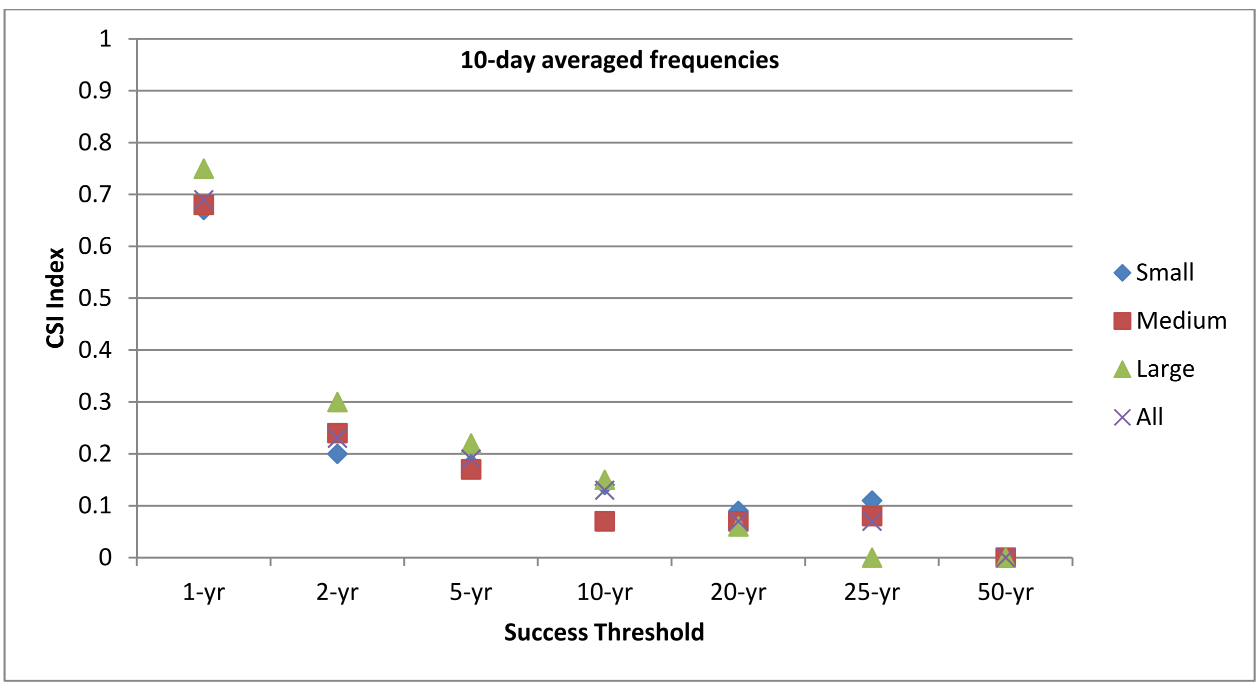

Figure 16 – Success rate of event prediction for the 10-day averaged frequency results.

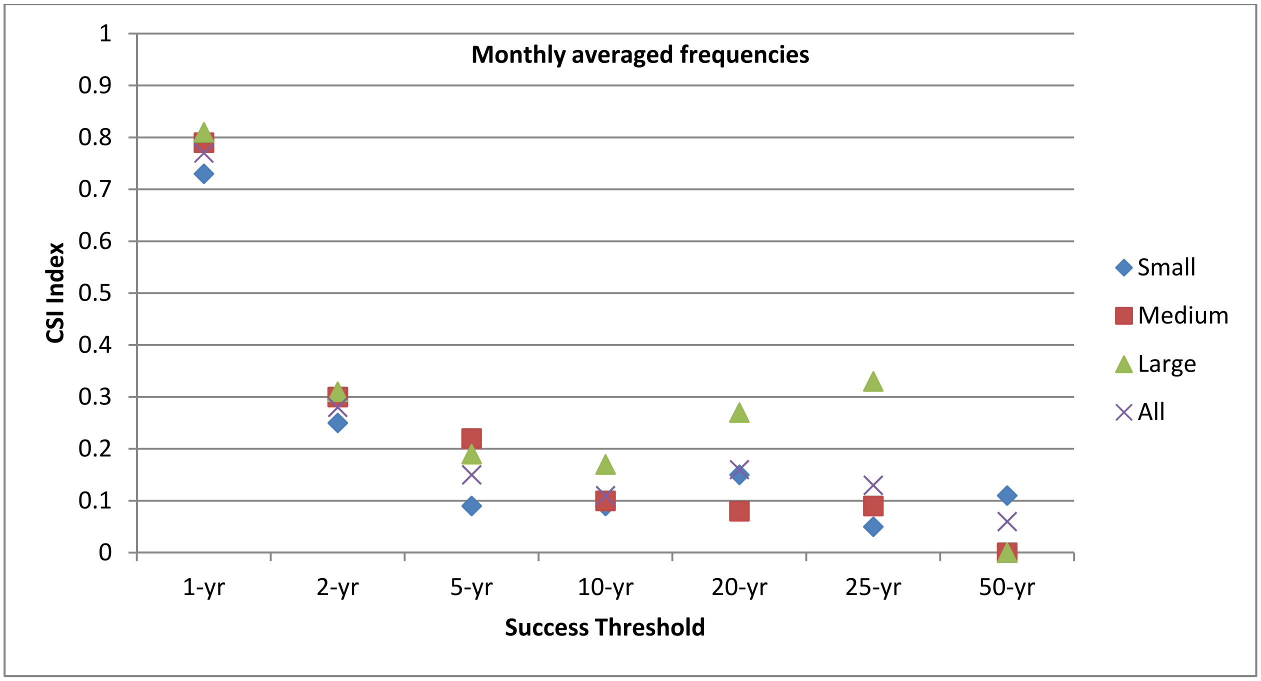

Figure 17 - Success rate of event prediction for the monthly averaged frequency results.