Extreme value theory: Applications to estimation of stochastic traffic capacity and statistical downscaling of precipitation extremes

Eric Matthew Laflamme

Follow this and additional works at: https://scholars.unh.edu/dissertation

Recommended Citation
https://scholars.unh.edu/dissertation/744

This Dissertation is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.
EXTREME VALUE THEORY: APPLICATIONS TO ESTIMATION OF STOCHASTIC TRAFFIC CAPACITY AND STATISTICAL DOWNSCALING OF PRECIPITATION EXTREMES

BY

ERIC MATTHEW LAFLAMME

BA, Bucknell University, 1998
MPS, Cornell University, 2004

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy
in
Mathematics
September, 2013
This dissertation has been examined and approved.

Ernst Linder, Professor of Mathematics and Statistics

Paul J. Assenbruggen, Professor of Civil Engineering (Emeritus)

Jennifer M. Jacobs, Professor of Environmental Engineering and Civil Engineering

Linyuan Li, Associate Professor of Mathematics and Statistics

Philip J. Ramsey, Instructor of Mathematics and Statistics

August 2, 2013
ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to both Dr. Ernst Linder and Dr. Paul J. Ossenbruggen for their support, assistance, patience, and availability throughout the entirety of this process. Next, I would like to thank Yibin Pan, a fellow graduate student and research partner whose help and input were invaluable. Lastly, the traffic analysis, the first part of this work, would not have been possible without the help and support of the New Hampshire Department of Transportation, most notably Denise Markow for her assistance in data acquisition.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ iii
LIST OF TABLES ...................................................................................................................... vii
LIST OF FIGURES.................................................................................................................... viii
ABSTRACT ................................................................................................................................. xii
PART I: ESTIMATION OF STOCHASTIC TRAFFIC CAPACITY ................................................ 1
INTRODUCTION (Part I) ......................................................................................................... 2
CHAPTER I: DATA COLLECTION ......................................................................................... 4
  Aggregation ....................................................................................................................... 5
CHAPTER II: CAPACITY DATA .............................................................................................. 7
  Daily Flow Maxima as Capacity ..................................................................................... 7
  Censored Values .............................................................................................................. 9
  Related Work/Literature Review ................................................................................... 11
  Data Traces and Breakdown ......................................................................................... 11
  Notation ............................................................................................................................. 13
  Illustration ........................................................................................................................... 13
CHAPTER III: GENERALIZED EXTREME VALUE DISTRIBUTION FOR BLOCK MAXIMA... 15
  Model Assumptions .......................................................................................................... 17
CHAPTER IV: PREPROCESSING .......................................................................................... 19
  Spillback ........................................................................................................................... 22
  Comparison of Maxima to Breakdown Flows .............................................................. 24
CHAPTER V: BAYESIAN APPROACH ................................................................................. 26
  Procedure and Results ................................................................................................... 29
  Validation .......................................................................................................................... 33
  Comparison to Other Estimates of Capacity ............................................................... 39
CHAPTER VI: STOCHASTIC CAPACITY IN PRACTICE ....................................................... 46
CHAPTER VII: HIERARCHICAL MODEL ............................................................................... 51
  Procedure and Results ................................................................................................... 54
LIST OF TABLES

Table 1: Posterior estimates and credibility intervals for Bayesian (hierarchical and non-hierarchical) models. ..........................................................59
Table 2: RCM summary information...............................................................104
LIST OF FIGURES

Figure 1: Illustration of correspondence between daily flow maxima and breakdown event .................................................................9
Figure 2: Illustration of functional data trace for cases of censored maxima (left) and un-censored maxima (right) .................................................................14
Figure 3: Auto- (left) and partial auto-correlation function (right) plots of daily flow maxima by location .................................................................18
Figure 4: Example of a plot of FDA flow trace versus FDA speed trace for November 1, 2010 ...........................................................................20
Figure 5: Comparison of daily flow maxima to breakdown flows for both location q02 (left) and q03 (right) ........................................................................................................25
Figure 6: Bayesian fitted densities (left), cumulative distributions (right), median and upper/lower 5% curves (thin black lines), and MLE fit (thick black line) ........31
Figure 7: Histograms of validation data, overlaid fitted training densities (grey curves), and median and upper/lower 5% curves (thin black lines) ..........34
Figure 8: QQ-plots for GEV model fit to training data (x-axis) and validation data (y-axis) .......................................................................................35
Figure 9: QQ-plot for GEV model fit to training data, ignoring censoring designation (x-axis), and validation data (y-axis) .......................................................................................36
Figure 10: Comparison of GEV-derived capacity estimates (red line) and observed flows (black line) at location q03 (Two days of observations) ......37
Figure 11: Ten day sample of predicted breakdowns derived from fitted GEV distribution (top) versus observed breakdowns (bottom) for location q02. Red dots indicate breakdown, both predicted and observed ..........................38
Figure 12: Ten day sample of predicted breakdowns derived from fitted GEV distribution (top) versus observed breakdowns (bottom) for location q03. Red dots indicate breakdown, both predicted and observed ..........................39
Figure 13: Ten day sample of predicted breakdowns derived from a fixed capacity of 5400 vph (top), stochastic capacity using GEV (middle), and observed breakdowns (bottom) for location q02. Red dots indicate breakdown. .........................................................................................................................41
Figure 14: Ten day sample of predicted breakdowns derived from a fixed capacity of 5400 vph (top), stochastic capacity using GEV (middle), and observed breakdowns (bottom) for location q03. Red dots indicate breakdown.

Figure 15: Comparison of capacity distribution functions derived from Weibull models fitted to breakdown flows and GEV models fitted to daily flow maxima for locations q02 (left) and q03 (right). Results from K-S tests of equal distribution are included for each location (See inset).

Figure 16: Ten day sample of predicted breakdowns derived from a stochastic Weibull model fitted to breakdown flows (top) and observed breakdowns (bottom) for location q02. Red dots indicate breakdown.

Figure 17: Ten day sample of predicted breakdowns derived from a stochastic Weibull model fitted to breakdown flows (top) and observed breakdowns (bottom) for location q03. Red dots indicate breakdown.

Figure 18: Increased, average, and decreased (blue, black, and red lines, respectively) simulated flow derived from one daily observed flow pattern at location q03.

Figure 19: Predicted congestion based on comparison of fitted capacity distribution to decreased (top), average (middle) and increased (bottom) simulated flows (red marks indicate predicted breakdown).

Figure 20: Flow-chart of hierarchical structure of parameters.

Figure 21: Hierarchical (right) and non-hierarchical (left) densities (top) and cumulative distributions (bottom) for location q02 with median curves (thick black lines) and upper/lower 5% curves (thin black lines).

Figure 22: Hierarchical (right) and non-hierarchical (left) densities (top) and cumulative distributions (bottom) for location q03 with median curves (thick black lines) and upper/lower 5% curves (thin black lines).

Figure 23: Hierarchical capacity distribution functions by location with 50th percentile (red lines).

Figure 24: Plot of observed flows versus speeds at location q03. Red lines separate distinct regimes of congested and freely flowing traffic while blue line
represents midpoint between these boundaries used as a breakdown threshold.

Figure 25: Layout of I-93 at locations q02-q04 illustrating number of lanes, exit/entrance ramps, and approximate distances between structures. ...........74

Figure 26: Cyclic trend observed in traffic flows at location q03. .................79

Figure 27: K-S test for assumption of GPD model form for each New England location (Red dots indicate significant differences) ...........111

Figure 28: 25-year return level estimates for predicted series, \( Y_P \), for all 58 New England locations. .................................................................117

Figure 29: Difference in 25-year return level estimates for \( X_P \) and \( X_C \) ..........119

Figure 30: Difference in 25-year return level estimates for \( Y_P \) and \( Y_C \) ..........120

Figure 31: 90% bootstrap confidence interval for downscaled (\( Y_P \)) 25-year return level estimates for New Hampshire locations only. .........................122

Figure 32: Difference in 90% confidence interval lower bound for \( X_P \) and \( X_C \) 25-year return level estimates. .................................................................124

Figure 33: Difference in 90% confidence interval lower bound for \( Y_P \) and \( Y_C \) 25-year return level estimates. .................................................................125

Figure 34: Comparison of BCa and mean corrected 25-year return level estimates .................................................................129

Figure 35: Difference in 90% confidence interval lower bound for \( Y_P \) and \( Y_C \) BCa 25-year return level estimates. .................................................................130

Figure 36: Comparison (for all 8 RCM/GCM combinations and 8 New Hampshire locations) of downscaled (predicted) distribution and distribution of exceedances of validation period. Black lines represent distribution of validation set exceedances and red lines represent predicted distributions .................132

Figure 37: p-values for one-sided Kolmogorov-Smirnov tests by location. Red dots provide evidence to reject the claim that predicted distribution fit to training data is equal to distribution of validation data. .................................134

Figure 38: Illustration of non-parametric extension for one location ..................143

Figure 39: Comparison of Bayesian and bootstrap 25-year return level estimates. ...............................................................................................145

Figure 40: Difference in \( Y_P \) and \( Y_C \) Bayesian 25-year return level estimates. ....146
Figure 41: Difference in 90% confidence interval lower bound for $Y_P$ and $Y_C$
Bayesian 25-year return level estimates. ................................................................. 148
ABSTRACT

EXTREME VALUE THEORY: APPLICATIONS TO ESTIMATION OF STOCHASTIC TRAFFIC CAPACITY AND STATISTICAL DOWNSCALING OF PRECIPITATION EXTREMES

by

Eric Matthew Laflamme
University of New Hampshire, September, 2013

This work explores two applications of extreme value analysis. First, we apply EV techniques to traffic stream data to develop an accurate distribution of capacity. Data were collected by the NHDOT along Interstate I-93, and two adjacent locations in Salem, NH were examined. Daily flow maxima were used to estimate capacity, and data not associated with daily breakdown were deemed censored values. Under this definition, capacity values are approximated by the generalized extreme value (GEV) distribution for block maxima. To address small sample sizes and the presence of censoring, a Bayesian framework using semi-informative priors was implemented. A simple cross validation procedure reveals the GEV model, using both censored and observed capacity data, is suitable for probabilistic prediction. To overcome the uncertainty associated with a high number of censored values at one location, a hierarchical model was developed to share information between locations and generally improve fitted results.

Next, we perform a statistical downscaling by applying a CDF transformation function to local-level daily precipitation extremes (from NCDC
station data) and corresponding NARCCAP regional climate model (RCM) output to derive local-scale projections. These high-resolution projections are essential in assessing the impacts of projected climate change. The downscaling method is performed on 58 locations throughout New England, and from the projected distribution of extreme precipitation local-level 25-year return levels are calculated. To obtain uncertainty estimates for future return levels, both a parametric bootstrap and Bayesian procedure are implemented. The Bayesian method consists of a semi-parametric mixture model for daily precipitation where extremes are modeled parametrically using generalized Pareto distributions, and non-extremes are modeled non-parametrically using quantiles. We find that these Bayesian credibility intervals are generally larger than those obtained from a previously applied parametric Bootstrap procedure, indicating that projected trends in New England precipitation tend to be less significant than is hinted at in many studies.
PART I: ESTIMATION OF STOCHASTIC TRAFFIC CAPACITY
INTRODUCTION (Part I)

When studying traffic flow and volume, a prominent, yet not so easy concept to define, is that of roadway 'capacity' (Hyde and Wright, 1986). The Highway Capacity Manual (HCM) defines capacity as 'the maximum hourly rate at which persons or vehicles can be reasonably expected to traverse a point or a uniform section of a lane or roadway during a given time period, under prevailing roadway, traffic and control conditions' (HCM, 2000). As this definition includes the term 'expected,' the capacity of a freeway facility is likely not a constant value. However, the HCM currently provides a single set of capacity values for basic freeway segments as a function of free-flow speed under ideal conditions. These values presented in the HCM are considered to be reasonably representative values for freeways located throughout the U.S. and are typically based on the analysis of speed-flow diagrams. By using these values, and by using conventional static and dynamic traffic assignment methods, road capacities are deterministic, or fixed.

It has been shown that capacity varies widely on a daily basis for the same facility and under the same geometric and traffic conditions (Lorenz and Elefteriadou, 2001; Elefteriadou et al., 1995; Persaud et al., 1998; Persaud et al., 2001; Brilon, 2005; Cassidy and Bertini, 1999; Kuhne et al., 2006). Moreover, breakdown does not necessarily occur at the same demand levels, but can occur when flows are lower or higher than the numerical value traditionally accepted as capacity (Elefteriadou et al., 1995). Thus, a single value of the
capacity value for a freeway facility does not reflect real-world observations and capacity should be considered a random variable that is \textit{stochastic} in nature (See, for example, Brilon et al., 2005; Brilon and Geistefeldt, 2007; and Geistefeldt, 2008). By considering capacity in this way, by rejecting the deterministic view of the HCM, one is left to identify/estimate capacity as a probability distribution across a range of values (See, for example, Brilon et al., 2005; Lorenz and Elefteriadou, 2001). The corresponding cumulative distribution of capacity values has become known as the ‘capacity distribution function,’ $F_c$ (Brilon et al., 2005), and is considered a valuable tool for evaluating roadway performance and efficiency. For further details regarding the derivation, theoretical basis, estimation, and interpretation of the capacity distribution function, the reader is directed to \textit{APPENDIX A: STOCHASTIC CAPACITY}.

In this analysis, we pursue the use of \textit{daily flow maxima} as estimates of capacity. After outlining our theoretical justification for using such values, the generalized extreme value (GEV) distribution will be used for their approximation. Based on data collected from two adjacent locations from Interstate 93 in New Hampshire, we will use the GEV model form to define a capacity distribution function, a measure of stochastic capacity, for the freeway segments. In addition to extreme value theory, this work will apply aspects of lifetime analysis and Bayesian model-fitting based on Markov Chain Monte Carlo (MCMC) methods. The approach will later be extended to a hierarchical platform.
CHAPTER I: DATA COLLECTION

Data was collected by the State of New Hampshire (NH) Department of Transportation along northbound lanes of interstate I-93 at eleven locations between Northern Massachusetts (MA) and Manchester, NH. These locations along I-93, denoted as q01, q02, ..., q11, begin in MA and proceed northerly in roughly one-mile increments. Depending on location and demand, I-93 is either a two- or three-lane freeway segment, but with the same uniform design, width, etc. Additionally, a variety of secondary roadway characteristics along the route account for further differences between collection sites. Data collection was made possible by side-fire radar units secured to portable, but semi-permanent, platforms along I-93. At each location, the radar devices were scheduled to intermittently measure traffic at irregular but frequent time periods about 1 minute apart. The data collection period occurred between April 1 and November 30, 2010, a total of 242 days. Data quality is highly variable as measurements from some locations are high-quality (extremely accurate and precise) and nearly complete, while those at other locations exhibit unexplained, sudden shifts in flow and speed measurements, are hindered by missing data, or contain a large number of spillback events (See Spillback in CHAPTER IV: PREPROCESSING). For each device, observations (raw data) consist of the following measurements: vehicle counts, average speed, occupancy, and speed of individual vehicles observed over the interval.
Aggregation

Because radar data are collected over very short, irregular time intervals, these measurements were aggregated into uniform intervals. The HCM recommends that transportation engineers use aggregates (intervals) no shorter than 15-minutes in order to ensure 'stable' traffic flow rates. Longer intervals are especially suitable for macroscopic/speed-flow analyses and intervals shorter than 5-minutes should be avoided (Highway Capacity Manual, 2000). That is, for shorter intervals, it is possible to observe speeds and flows that simply cannot be sustained over longer periods, and these measurements are not accurate representations of traffic conditions. As stated by Greenwood et al. (2007), analyzing capacity is a macroscopic endeavor related to management, policy, design and regional speed/flow comparisons, thus stable, longer interval durations are appropriate. While exact lengths vary among analyses, 15-minute intervals are commonly used for capacity studies (See, for example, Agyemang-Duah and Hall, 1991; Yang and Zhang, 2005; Lorenz and Elefteriadou, 2001; Elefteriadou and Lertworawanich, 2002; Minderhoud et al., 1997). For these reasons, and to produce results that are directly comparable to those put forth by the HCM, observations were binned into 15-minute aggregates.

Raw traffic flow measurements were reported as total volume, or number of vehicles, observed during a precisely defined interval. From each of these vehicle counts and the corresponding interval length during which they were observed, a flow rate, or equivalent hourly rate of vehicles passing a point (the 'traffic flow rate,' by definition), was simply calculated. This then yields flow rates
that are expressed in the customary scale of vehicles per hour (vph), but based on short, unstable time intervals. Next, these flow rates were binned into non-overlapping, sequential 15-minute intervals based on original observation times. Lastly, for each bin, the component flow rates were averaged to produce aggregated flow \((q)\) rates for each 15-minute interval during the collection period. Addressing traffic speed \((u)\) is a simpler task as precise speed (in mph) and time measurements are recorded for every vehicle observed during the collection period. Speed aggregates are then calculated by first separating all observations into 15-minute bins, and then simply taking their average. Clearly, both speed and flow aggregates are sensitive to missing data as averages based on few measurements can be erratic.
CHAPTER II: CAPACITY DATA

Daily Flow Maxima as Capacity

While researchers have come to acknowledge capacity as a stochastic process, no universally accepted measure of capacity has been established. Breakdown flows, however, those flows measured immediately before the onset of congestion, have been widely adopted as estimators of capacity (see, for example, Elefteriadou and Lertworawanich, 2002; Brilon et al., 2005; Lorenz and Elefteriadou, 2001; Minderhoud et al., 1996). However, one concern when using breakdown flows is that higher flows or daily flow maxima often occur prior to congested conditions. In such cases, as density increases, flow peaks and then begins to drop as the freeway section enters a state of congestion. Such behavior has been documented by Banks (2009) who, in an investigation of freeway bottlenecks, commonly observed periods of decreasing flow before breakdown. These cases suggest that breakdown flows are actually underestimating capacity since higher flows were observed just prior to breakdown. By instead considering daily flow maxima, we would capture these higher flows prior to breakdown and ultimately obtain more representative measures of capacity, the maximum flow that a roadway can sustain. Another concern when using breakdown flows is that they are highly dependent on subjective, and often arbitrary, breakdown identification criteria. For example, if breakdowns are identified based on a fixed speed threshold, as many are, changing this threshold can significantly affect the breakdown flows extracted.
from the traffic stream data. Daily flow maxima, of course, have the advantage of being independent of the methods used to identify breakdowns.

Another commonly used estimate of freeway capacity is the maximum pre-breakdown flow (See, for example, Hall and Agyemang-Duah, 1991; Hall et al., 1992), or the maximum sustained flow prior to breakdown. Based on aggregated traffic data, Elefteriadou and Lertworawinich (2002) provide evidence that distributions of breakdown flows and maximum pre-breakdown flows are statistically similar and, by extension, that maximum pre-breakdown flows are suitable estimates for capacity. Because daily flow maxima typically occur prior to breakdowns, in practice, daily flow maxima closely resemble maximum pre-breakdown flows, although the two are conceptually different. Thus, using daily flow maxima to estimate capacity has many of the same benefits as using pre-breakdown flows. For one, both maxima and pre-breakdown flows account for cases (as described above) where flows decrease prior to congestion, a so-called 'lingering' or lagged effect, and are better representations of the true maximum throughput of the roadway. However, while maxima and pre-breakdown flows are somewhat similar, maxima have the advantage of being consistently available. That is, extraction of pre-breakdown flows is completely dependent on breakdowns, while daily maxima, on the other hand, have no such dependence and may be obtained on a regular basis regardless of breakdown occurrence. Using maxima, then, is especially beneficial when breakdowns are rare as such cases would yield relatively few pre-breakdown flows. Of course, this becomes less of an issue when breakdowns occur very frequently and thus pre-breakdown flows become more abundant.
Censored Values

As stated by the HCM, the capacity for a given facility is the flow rate that can be achieved repeatedly for peak periods of sufficient demand (Highway Capacity Manual, 2000). Therefore, not every maximum flow is a suitable estimate of freeway capacity, and only those daily maxima associated with 'sufficiently high demand', demand typically resulting in breakdown, should be considered as such. On days where demand is insufficient, when breakdowns are not observed, daily conditions are not adequate (sufficiently extreme) to assess the true capacity of the roadway. In these cases, under a survival analysis premise, the corresponding maxima are deemed censored (actually right-
censored) capacity values as the roadway can surely service higher demands. That is, breakdowns would occur at some higher flow rates, and the resulting capacity, the maximum daily flows, would necessarily be larger than the observed value. Alternatively, we may simply consider these cases as incomplete data records where the true maximum, the capacity, is simply missing. By introducing censoring into our definition of capacity, our data are a sort of compromise between pre-breakdowns flows and simple daily maximum flows. Also, our censoring designation establishes a correspondence between extreme flows and breakdown which we intuitively know exists.

Despite the incompleteness associated with censored values, they still contain valuable information and will therefore be considered in the calibration of our capacity distribution (See, for example, Geistefeldt, 2010). Non-parametrically, the capacity distribution function, $F_c(q)$, has been estimated using the Kaplan-Meier/Product Limit method (PLM) based on samples that include both censored and uncensored values, a survival analysis approach (See APPENDIX A: STOCHASTIC CAPACITY). Minderhoud et al. (1996) would later formulate the PLM for the analysis of freeway capacity. Parametrically, assuming a known model form, capacity distributions have been estimated based on both censored and uncensored values as well. It has been shown that, under this survival analysis framework, the Weibull model is well-suited and accurately approximates capacity data based on breakdown flows (Brilon and Zurlinden, 2003; and Brilon et al., 2005). In a comparison of capacity distributions approaches, Geistefeldt and Brilon (2009) found that using censored data achieves significantly more precise estimates, especially at higher quantiles.
Related Work/Literature Review

In Minderhoud et al. (1997), extreme values are discussed, and the authors state that observed maximum volumes (collected over days, for example) and corresponding extreme value statistics can be used to estimate capacity distributions. Next, and probably most applicable to our study, is the work of Hyde and Wright (1986) who use a variety of flow maxima to calibrate a capacity distributions based on direct probability methods and asymptotic theory. Their asymptotic approach relies on the theory (extreme value theory) that large values/maxima can be approximated by a simple, known statistical model. These works provide evidence that maximum flows are appropriate estimates of capacity and lend credibility to our approach.

Data Traces and Breakdown

For the analysis described above, the general behavior of the traffic stream data must be identified. Although the 15-minute aggregated data is an accurate representation of speed and flow, this data is still very ‘noisy.’ Irregularities in the data collection process, or missing measurements, may result in aggregates calculated from few observations. Because these raw traffic observations can be erratic, because short intervals may identify traffic estimates that can only be sustained over the short term, aggregates based on few of these raw observations may be similarly noisy. To address this issue, functional data analysis (FDA) is employed to produce summaries, or smooth representations, of the data (both speed and flow) over a fine time scale. These resulting interpolating smooths will be herein referred to as ‘traces’ (For further
details regarding our FDA models and theory, please see APPENDIX D: FUCNTIONAL DATA MODELS. Regarding flow, traces are especially beneficial as they represent trends of already aggregated data, and the resulting peaks necessarily correspond to sustained maxima, natural estimates of capacity. Therefore, going forward, daily flow maxima will be extracted from fitted FDA flow traces.

Because the censoring designation of daily flow maxima is dependent on daily breakdowns, and because these values directly relate to fitted capacity distributions, accurate assessment of breakdowns is essential. Like most traffic analyses, we define breakdown to be the transition between freely flowing traffic and congested conditions. This transition, or breakdown, occurs when persistent speeds above a fixed threshold are immediately followed by persistent speeds below the same threshold. This speed threshold was set to 48 mph based on a visually distinction between congested and freely flowing traffic regimes (See APPENDIX B: SPEED THRESHOLD SELECTION). This means that breakdown is only identified for cases when at least 5 minutes of freely flowing travel (average speed above 48 mph) is followed by at least 5 minutes of congested conditions (average speeds below 48 mph). However, when using 15-minute aggregates, measurements may be difficult to decipher, and exact moments of breakdown are impossible to identify. Consequently, some shorter, ‘true’ breakdowns could be overlooked if an analysis is based solely on coarse speed aggregates. By applying FDA, we are able to extract interpolated, intermediate values based on the smooth representations of the data. Thus, traces address the two issues with our data: they provide precise (in time) estimates while maintaining smoothness.
For this reason, going forward, breakdown identification, sustained drop in traffic speeds, is based on speed traces rather than aggregated data.

**Notation**

Let the random variables $C$ and $Q_{\text{max}}$ be respective capacity and the maximum daily flow of the roadway where, for a given day, $q_{\text{max}}$ is the maximum observed flow. Also, let $Y_t = \{0, 1\}$, where 1 corresponds to a breakdown/transition to a congested state at time $t$, and 0 corresponds to freely flowing traffic at time $t$. In practice, for each day, two categories may be observed.

1) If $Y_t = 0$ (i.e., when no breakdown is observed) for every $t$, then $q_{\text{max}} < c$.

2) If $Y_t = 1$ for some $t$ (i.e., when breakdown is observed), then $q_{\text{max}} = c$.

**Illustration**

To illustrate our procedure, we observe typically traffic traces that result from the application of FDA. In all cases, the $x$- and $y$-axis correspond to flow and speed, respectively, and time proceeds from the upper endpoint to the lower endpoint. First, for the majority of speed/flow/time traces, breakdowns are observed and daily flow maxima typically occur within a relatively short time prior to them (Below, right). For such a case, the daily demand is deemed sufficiently high to warrant capacity estimation and the maximum daily flow is recorded as the capacity, or $q_{\text{max}} = c$. Next, for some cases, speed traces never fall below our set threshold and breakdowns never occur (Below, left). In such a case, since breakdown is never realized, the corresponding capacity of the
roadway is greater than the highest observed daily flow and the maximum is recorded as a censored value, or \( q_{\text{max}} < c \).

Figure 2: Illustration of functional data trace for cases of censored maxima (left) and un-censored maxima (right).
CHAPTER III: GENERALIZED EXTREME VALUE DISTRIBUTION FOR

BLOCK MAXIMA

Consider an underlying, and often unobservable, i.i.d. sequence of independent random variables, \(X_1, X_2, \ldots\) with common distribution function \(F\), where \(F(x) = \Pr(X \leq x)\). Let \(M_n = \max(X_1, \ldots, X_n)\), the maximum of the process over \(n\) time units of observation. Thus, \(M_n\) is the \(n\)-sample maximum. Since \(\Pr(M_n \leq x) = F(x)^n\), then \(\Pr(M_n \leq x) \to 0\) as \(n\) gets large and the distribution of \(M_n\) degenerates to a point mass. However, if we can find appropriate standardizations for the maximum, this probability will achieve a non-degenerative limiting distribution. That is,

\[
\Pr\left\{ \frac{M_n - b_n}{a_n} \leq x \right\} = F(a_n x + b_n)^n \to G(x)
\]

where \(a_n\) and \(b_n\) are normalizing sequences and \(G(x)\) is a non-degenerative distribution. Furthermore, by theorem, if normalizing sequences exist and a distributional limit exists on \(M_n\), then \(G\) must be one of three families of distributions. These three families are known as the Gumbel, Frechet, and Weibull (strictly, negative Weibull) and correspond to tail behavior that is exponential, bounded below, and bounded above, respectively. This is the 'three types theorem' that is the backbone of classical extreme value theory. However, working with three distinct distribution families may become inconvenient (Coles, 2001), and it has become the norm to work with a general distribution that includes the Gumbel, Frechet, and Weibull distributions as special cases. This
A subclass of distributions is known as the generalized extreme value distribution, or GEV, and may approximate a variety of tail distribution shapes. Direct use of the GEV rather than the three types separately allows for flexible modeling and avoids having to determine which family is most appropriate and simply allows the data to decide. The GEV is given by the following form:

\[ G(x) = \exp \left\{ -\left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \]

defined on \( \{ x: 1 + \xi (x - \mu)/\sigma > 0 \} \) with \( \mu \) and \( \sigma \) the respective location and scale parameters. The shape parameter of the GEV, \( \xi \), characterizes the rate of tail decay, where \( \xi > 0 \), \( \xi = 0 \), and \( \xi < 0 \) correspond to data with heavy tails, light tails, and short tails, respectively. For our data, in the expression above, we replace the random variable \( X \) with capacity data \( C \) as both censored and uncensored (observed) daily flow maxima, by theory, can be suitably and accurately approximated by the GEV.

In most extreme value analyses, blocks are typically set to a length of one year to limit bias in estimation. However, this recommendation assumes data are collected relatively sparsely, which is not the case for our data, and thus daily flow maxima will suffice. Another concern when considering block maxima is that too few data (maxima) will obviously lead to larger estimation variance. With potentially hundreds of daily maxima collected over several months, we have ample data to adequately estimate GEV model parameters.
Model Assumptions

First, as above, the GEV model assumes a series of independent random variables in defining block maxima. Clearly, this is not the cases as traffic flows are highly dependent and follow a daily, cyclic pattern. In such cases, despite violating model assumptions, the conclusion that the block maxima have a GEV distribution may still be reasonable (Coles, 2001). Quantile plots for daily maxima (both q02 and q03) show strong correspondence between observed and predicted values, which illustrates the appropriateness of the model class and justifies the use of the GEV for our data. Next, it is critical that block maxima themselves be independent, a requirement, according to Coles (2001), which is likely even if the original data constitute a dependent series, such as we have. In our case, a daily pattern is observed somewhat comparably each day, and there is expected to be little influence on one day’s maximum on any other (day-to-day dependence). Autocorrelation (ACF) and partial autocorrelation (PACF) plots for both locations show no evidence of autocorrelation or dependence, and daily maxima are appropriately assumed independent (Figure 3). Lastly, the approach assumes daily maxima be identically distributed. Given the 8 months of data, we observe no true seasonality (‘heavy’ season), so there is homogeneity in the daily traffic flow. That is, daily maxima exhibit no trend, no oscillations or systematic patterns in maximal values and are thus approximately identically distributed.
Figure 3: Auto- (left) and partial auto-correlation function (right) plots of daily flow maxima by location.
CHAPTER IV: PREPROCESSING

The entire collection period spans a period of 9 months, or 242 days, and details regarding the data can be found in CHAPTER I: DATA COLLECTION. Of the 11 northbound radar locations, only a handful yielded continuous traffic measurements for the entire collection period (only about 40 measurements per hour were recorded at the best locations). Of these, data from only two locations, q02 and q03, located in Salem, NH, were retained as the vast majority of breakdowns at the others were attributable to spillback (See Spillback in CHAPTER IV: PREPROCESSING) and deemed unsuitable. As data from q02 and q03 are deemed sufficiently reliable, our flow maxima analysis is based solely on traffic measurements from these two locations. Following the procedure outlined in Aggregation of CHAPTER I: DATA COLLECTION, data are aggregated into 15-minute intervals for both speed and flow. From these speed and flow aggregates, and their corresponding unique time identifiers, FDA models were fitted (See APPENDIX D: FUNCTIONAL DATA MODELS) and resulting traces were used to estimate daily maxima, breakdown, and times of breakdown.

At location 3, q03, due to scheduled maintenance of the measurement equipment, there were typically two or three days at the end of the month that were missing measurements. In addition to this, at location q03 only, there were short stretches of interruptions in August, September, and November during which no traffic measurements were recorded. For the remaining days, FDA flow and speed models, the traces, were fitted to the data. Very rarely, in four cases,
shorter, intra-day gaps in the data resulted in unreliable traces or complete failure of the FDA fitting process. In the end, a total of 206 daily maxima were retained from q03 for the analysis. In order to conduct a spillback (tailback) analysis (discussed in the following section) at q02, reliable downstream data, from q03, is required. Thus, data from q02 was used for only those days where FDA (speed-flow, speed-time) models at q03 exist. For these 206 days, FDA models were fit to q02 traffic stream data without incident. Following the procedure outlined above, speed-flow and speed-time plots for q02 and q03 were compared, carefully evaluated, and checked for outliers or unusual behavior. Below is an example of an FDA trace at q03 for November 11, 2010. We notice that breakdown is not observed as the speed trace never drops below 48 mph, and thus the maximum recorded flow is a censored value.

Figure 4: Example of a plot of FDA flow trace versus FDA speed trace for November 1, 2010.

Both traffic locations, q02 and q03, are designed to the highest standards, classified by the HCM as 'ideal,' and both locations are subject to relatively high
daily traffic flow conditions (for further details of the collection sites, see

APPENDIX C: COLLECTION SITES). It suffices to say that both q02 and q03 are
suitable sites for the collection of capacity data. However, location q03, a
physical bottleneck where traffic merges and discretionary weaving is observed,
experiences congestion (breakdown) on a daily basis, whereas location q02
serves traffic mostly free of interruption and experiences breakdowns at a much
lower rate. Because of this disproportionate number of breakdowns, our
procedure yields a disparate number of capacity data from the q02 and q03
traffic traces. For q03, from the 206 days with FDA traces, 145 capacity data
were identified where daily maxima were associated with sustained traffic
breakdown. Incidentally, for these 145 cases, daily maxima were typically
identified just a few minutes prior to breakdowns (median = 12 minutes). The
remaining 61 days did not observe breakdowns, so corresponding daily maxima
were deemed censored values. At location q02, results are quite dissimilar as
FDA analyses of speed, flow, and time result in only 60 capacity data and 146
censored data. Also, like location q03, when congested conditions are
observed at q02, maxima are typically found immediately prior to breakdown
(median = 7.5 minutes).

Analyses based on maxima (or minima) are typically wasteful of data as
only one measurement per block (day) is considered. However, our process
does retain a fair number of capacity data, and supplementing these estimates
with censored information yields datasets sufficient for model-fitting and
inference.

21
Spillback

Spillback (or tailback, spillage) occurs when a queue originating from a downstream location spreads upstream. Spillback assessment/identification is especially critical for capacity analyses based on breakdown flows (previously defined) as breakdowns caused by oncoming queues are not naturally occurring and will result in atypical estimates of capacity. Breakdown flows attributable to spillback are commonly omitted from capacity studies. In our case, flow maxima are used as capacity estimates, and these measurements are not directly defined by breakdown. However, because censoring is based solely on the observance of daily breakdowns, capacity data (flow maxima) extracted from days where spillback is observed should rightly be considered censored values. Un-censored capacity data, then, should only correspond to days where 'true,' non-spillback breakdowns occur.

Based on the results of a previous analysis, and based on our knowledge of the freeway segments, data from q03 was not scrutinized for cases of potential spillback. That is, because a structural bottleneck is located immediately downstream of device q03, and because no bottleneck is located near q04 (downstream of q03), no spillage is expected to corrupt data obtained from q03. In our experience, the few breakdowns at q04 occur concurrently to or immediately after breakdowns upstream at q03, in the vast majority of cases. Thus, we have no evidence to suggest that queues at q04 corrupt capacity data at q03. Spillback is, however, examined at location q02.
From FDA traces, precise breakdown times at q02 (upstream) were compared to those at q03 (downstream) to assess potential spillback. We observe a relatively strong correlation between these times, with both upstream and downstream breakdowns typically occurring around 12 hours (12PM) or between 15 and 18 hours (3PM – 6PM). In the vast majority of cases, these values occur concurrently, or maxima at q02 slightly precede breakdowns at q03. Thus, for these cases, one of two scenarios is observed, neither of which allows for the possibility of spillback. First, and most likely, traffic flow mimics a ‘rising tide’ where demand is increased across the network simultaneously and breakdowns/maxima occur at roughly the same moment. Or, second, high flows originating from an upstream location (the south, upstream of q02) are progressively spreading through the network (q02 and q03). In this scenario, the same elevated flow that creates a bottleneck at q03 is also responsible for high flows at q02 at an earlier time.

For those few cases where maxima at q02 occur after breakdowns at q03, a shock-wave speed analysis was applied. Shock-wave speed is defined to be the speed of propagation of a disturbance in a traffic system, or, in this case, the speed of the upstream traveling queue. Based on observations and admittedly crude estimates of congested speeds, we calculate the shock-wave speed to be roughly 5 mph. In a study of freeway bottlenecks, Bertini and Leal (2004) found shock-wave speeds (queue speeds traveling upstream) at a three-lane bottleneck to be between 3 and 4 mph and nearly constant, strong support for our estimate. So, based on a shock-wave speed of 5 mph, it would take roughly 12 minutes for a queue to reach q02 from q03, a distance of about one mile.
Therefore, if spillback is occurring, we would expect to see breakdowns at q02 between 12 and 15 (5 and 4 mph shock-wave speeds, respectively) minutes after breakdowns downstream at q03. Such cases, however, are rarely observed as breakdowns at q02 are found either a very short (a few minutes) or very long time (more than an hour, say) after breakdowns at q03.

In general, based on the assumption of consistent shock-wave speeds (Bertini and Leal, 2004), if spillback was the source of congestion at q02, we would expect to see similarly consistent differences in breakdown times between upstream and downstream locations, around 12-15 minutes (as discussed). However, these differences are not consistent, are randomly dispersed in terms of magnitude and sign, and do not follow any discernible pattern. Without more concrete, confirmatory evidence, congestion is not believed to spillback from q03 to q02. Thus, breakdowns at q02 and q03 are assumed to be coincidental and independent. We suspect that the exit ramp at q03 plays a significant role in efficiently accommodating the local traffic flow and minimizing the chance a queue reaching a length of one mile.

**Comparison of Maxima to Breakdown Flows**

To compare the capacity values resulting from our procedure (above) to those obtained via traditional methods (i.e. breakdown flows), overlaying densities were produced by location (See below). As expected, our procedure generally yields higher values of capacity as breakdown flow techniques are believed to misidentify capacity in cases where flows drop prior to congestion (Banks (2009), for example, addresses the occurrence of such flow drops).
approach is robust to this phenomenon and ultimately leads to a more diffuse distribution of capacity. We note, however, that these distributions are not extremely dissimilar.

Figure 5: Comparison of daily flow maxima to breakdown flows for both location q02 (left) and q03 (right).
CHAPTER V: BAYESIAN APPROACH

In general, daily maxima collected from q02 and q03 have two dominant characteristics: small sample sizes and a high number of censored values. Because of this, and because capacity model-fitting need not be done in real-time, a computational Bayesian approach was employed. We will provide a brief justification for the Bayesian approach as well as a brief overview of the procedure and theoretical basis. This, however, is not a rigorous treatment of Bayesian methodology. For a more detailed treatment, the reader is directed to any number of sources including Carlin and Louis (2008) and Gelman et al. (2003).

For small samples, Bayesian analysis can have significant advantages over a classical/MLE approach. First, Bayesian results do not depend on asymptotic theory. That is, inference based on maximum-likelihood estimators relies on an asymptotic distribution which may not be appropriate for small samples, although we suspect our datasets are large enough to justify MLE. Second, if priors are particularly well-informed, the Bayesian approach can combine this information with the sparse data. In this way, the use of subjective priors can add value to the analysis through increased accuracy and efficiency. That said, however, poorly chosen priors will negatively affect the resulting posteriors and lead to biased inference.

For censored values, Bayesian methods offer very natural, proper approach for addressing these 'incomplete' data (See, for example, Gelman et
al., 2003 or Ibrahim et al., 2005). As stated, in Bayesian statistics, prior knowledge along with a given set of current observations is incorporated in statistical inferences. This prior information could come from operational or observational data, from previous experiments (empirical Bayes), or from engineering knowledge. Under the computational Bayesian format, plausible values are generated from the posterior predictive distribution of the censored observations conditional on the observed data. That is, censored values are treated as missing data and imputed from draws of a random variable from a truncated distribution based on Markov chain Monte Carlo techniques. Furthermore, MCMC sampling allows for credibility intervals, the Bayesian analog to confidence intervals, to be established which, in turn, allows for clear interpretation of the effects of censored data.

For these primary reasons, results from the Bayesian/MCMC estimation are preferable to MLE, although MLE results are useful references in determining whether or not MCMC models, including their prior distributions, are correctly specified. Bayesian approaches are similarly preferred over non-parametric techniques of capacity estimation (Kaplan Meier approach, for example) as these estimators are not defined for all values of possible traffic flow. That is, in non-parametric approaches, the last observation defines the terminal point of the estimated distribution, and a complete curve is rarely obtained. Resulting distribution functions are thus discontinuous and non-smooth, and estimates are generally unreliable. In a related work, Ozguven and Ozbay (2008) concluded that Bayesian estimation is far superior (more efficient) to non-parametric
techniques for survival analyses with small samples and substantial amounts of censoring.

Under the Bayesian framework, identifying the capacity distribution function \( F_C(q) \) is simply identifying the cumulative distribution that most accurately estimates the capacity data. By assuming a model form for this distribution function, the GEV, the objective of the analysis is then simply the estimation of the model parameters. Thus, Bayesian estimation, which ultimately yields distributions for model parameters, seems a logical approach. The OpenBUGS statistically software was used for all Bayesian applications, and the analysis and manipulation of all OpenBUGS output, the data containing the Bayesian samples of the parameters, was then performed with the R statistical software. Limited details of Bayesian theory and MCMC techniques are presented in Appendix E: MCMC Sampling as a thorough treatment is beyond the scope of this analysis.

Because Bayesian/MCMC techniques are computationally demanding, they are still considered relatively new (the last 20 years). However, because Bayesian analyses can overcome difficulties in complicated data collection schemes, Bayesian literature is extensive and widespread applications can be found across disciplines. However, there is limited literature relating such methods to the estimation of capacity. Mentioned above, Ozguven and Ozbay (2008) introduce a non-parametric Bayesian estimation used to estimate capacity. To our knowledge, and to date, this is the most relevant work to our analysis. Zheng et al. (2006) used a combination of Bayesian and neural
networks approaches to develop short-term traffic flow predictions for freeway data. Tebaldi et al. (2002) used hierarchical regression models to capture and predict short-term changes in traffic flow for a freeway network.

**Procedure and Results**

Using asymptotic distributions, Hyde and Wright (1986) found flow maxima are approximated most accurately by short-tailed distributions. This agrees with our experience with extreme traffic flows and makes intuitive sense as there is an absolute limit to the number of vehicles a road may carry. Thus, finite upper bounds were assumed for the capacity data, and, consequently, GEV shape parameters were assumed to be negative. The corresponding prior distribution on the shape parameter reflects this limitation. A somewhat diffuse prior distribution for the GEV scale parameter was chosen, but within a realistic range based on previous model-fitting. Similarly, a semi-informative prior distribution was chosen for the location parameter as the center of the distribution can be estimated within a reasonable range. Thus, the scale and shape priors were not over-specified, and we are truly allowing the data to guide these posterior analyses. Specific prior distribution for the GEV form is defined below.

For both upstream and downstream locations, q02 and q03, freeway capacities, C, are assumed to be generalized extreme value distributions (GEV)

\[ C \sim GEV(\mu, \sigma, \xi) \]

with location, scale, and shape parameters \( \mu, \sigma, \) and \( \xi \), respectively. Collected data, \( q_{j,max} \), are maximum daily traffic flows for day \( j = 1, \ldots, 206 \), each of which is
classified as a censored or un-censored, capacity \( c_j \) value based on the previously established definition (See Data Traces and Breakdown in CHAPTER II: CAPACITY DATA). For the GEV model, the shape \( \xi \), scale \( \sigma \), and location \( \mu \) parameters are assumed to follow prior distributions with the following forms.

\[
\begin{align*}
\xi &\sim \text{Unif.} \left( a_\xi, b_\xi \right), \\
\sigma &\sim \text{Unif.} \left( a_\sigma, b_\sigma \right), \\
\mu &\sim \text{Unif.} \left( a_\mu, b_\mu \right)
\end{align*}
\]

where \( (a_\xi, b_\xi) \), \( (a_\sigma, b_\sigma) \), \( (a_\mu, b_\mu) \) are specified as \( (-.75, 0) \), \( (0, 10) \), and \( (3, 10) \), respectively.

The following results, the posterior analysis, are based on the output of 5,000 MCMC iterations, the first 2,000 discarded as a 'burn-in' period. Convergence and independence from the starting values were checked by CODA (distributions, traces, etc.), the standard tools in such cases. Also, in all cases, starting values for the sampling scheme were generated from the defined prior distributions. Below we have fitted distributions (densities and CDFs) for several thousand sets of parameter estimates, the median fitted distribution (as well as upper and lower 5% fitted distributions), and the distribution fitted via MLE to only uncensored values.
Figure 6: Bayesian fitted densities (left), cumulative distributions (right), median and upper/lower 5% curves (thin black lines), and MLE fit (thick black line).

From the plots above, we observe the effect of the inclusion of censored values on the fitted distributions, the capacity distribution functions, $F_c$s. For q03 (bottom), compared to the distribution fitted via MLE to uncensored values only (thicker curve), including censored values in our model-fitting results in right shift of the distribution (grey curves). Of course, because most data from q03 are
uncensored, the effect of including the censored value is more subtle (yet still visible). On the other hand, this effect, the effect of including censored values in model-fitting, is much more noticeable at q02 where substantially fewer capacity data are observed. Here, including a large number of censored values results in a very significant right-shift of the distribution, as expected. Also, because so few capacity data are available, the variability of the fitted curves at q02 is quite substantial, especially at the upper tail and compared to those of q03. For the same reason, we feel the MLE-based distribution, based solely on 60 capacity data, is unreliable and the Bayesian approach offers a sensible alternative for such a small sample. That is, informed (or semi-informed) priors and information contained in censored values improve our estimates and increase our confidence in fitted results.

Specifically, the MLE-derived capacity distribution for q02 corresponds to a segment that is very sensitive to demand. That is, based on the MLE result, q02 appears to be a segment that cannot ably handle high flows and has a high probability of congestion at relatively low flow values (See above). The Bayesian result, however, corresponds to a much more efficient roadway with comparably lower probabilities of congestion. Since both high demands and relatively few breakdowns are observed at q02, we have ample evidence to suggest that the segment can service high demands/flows without becoming overly congested. Because of this, we know the Bayesian result (capacity distribution function) to be realistic and highly credible. Results from q02 most clearly illustrate the effect, and potential advantage, of considering censored values in the model-fitting procedure.
Validation

First, we pursue a validation procedure that will allow us to evaluate the plausibility of our GEV model choice in predicting the distribution of capacity data not used in the fitting process, to assess the model's probabilistic predictive ability. The cross-validation (CV) technique used here is a repeated random sub-sampling procedure where the data is tested against itself. Under the premise of exchangeability, a training set is first created by randomly selecting 70% of both the censored and un-censored values. Of the remaining 30% of the data, only the non-censored capacity values were designated as a validation set. Since the training and validation sets are non-overlapping, the capacity validation set may be considered truly unobserved data suitable for evaluating the procedure. We note that because so few breakdowns were observed at location q02, validation was not a realistic option. That is, resulting training and validation sets were too sparse to provide reliable models and evaluation.

Using the Bayesian approach described previously, a GEV model is fit to the q03 training set and then compared to the validation set. This process is repeated numerous times to ensure consistency, and results from two of these replications are presented below. From these results, the histograms represent the distribution of validation capacity values, those capacity values not used in model fitting (the 30% withheld from model fitting). The grey and black lines represent the GEV distribution (density) fitted via Bayesian methods to the training data. We observe the fitted GEV model visually captures the shape of the validation data, and therefore appears to successfully predict the distribution.
of capacity data. This suggests that the GEV is an appropriate model form for our capacity data, and fitted results appear to benefit from the inclusion of censored values.

Figure 7: Histograms of validation data, overlaid fitted training densities (grey curves), and median and upper/lower 5% curves (thin black lines).

Additionally, quantile plots serve as a diagnostic tool for the evaluation of the fitted model. These plots compare theoretical quantiles from the model calibrated on the training set against the observed quantiles of the validation set. Results directly indicate probabilistic predictive ability for unobserved data, where adherence to the $y = x$ transverse indicates successful prediction of quantiles. From the plots below, despite small datasets, the GEV model class is a successful predictor of capacity data, especially at the upper tail. Additionally, as above, the procedure supports the use of censored values in model-fitting as their inclusion yields accurate probabilistic predictions.
To further illustrate the benefit of censoring, a validation was performed ignoring the designation and considering all daily flow maxima as capacity data. This set of 206 data was then divided into non-overlapping training and validation sets as above. A GEV model was fitted to the training data via MLE, and the resulting theoretical quantiles were compared to the observed quantiles of the validation data (See Figure 9 for one such quantile plot). As we can see, when censoring is not introduced into our flow maxima estimates, the model form significantly underestimates capacity, by more than 1,000 vph in some cases. This result provides evidence that our capacity data, the daily flow maxima, are most accurately approximated when introducing censored values. We note that this process was repeated numerous times to ensure consistency, and in all cases the model underestimated the validation data as below.
While the previous result validates the use of the GEV model form with censoring, it does not assess the ability of daily flow maxima to approximate freeway capacity. To do this, a validation procedure was performed in which 'true', observed breakdowns were compared to breakdowns predicted by the fitted capacity distribution function. To perform this validation, random capacity estimates were first generated from the fitted Bayesian GEV model for each 15-minute interval. Also, in the case of congestion, a modest and constant capacity drop of 6% is assumed as per Hall and A-Duah (1991) and Cassidy and Bertini (1999), for example. When observed demands exceed randomly generated capacities, a breakdown was predicted. Figure 10 below illustrates the procedure for two days of data at location q03.
Next, these predicted breakdowns were compared to true breakdowns, or breakdowns based on sustained speeds below a threshold of 48 mph. Zurlinden (2003), Brilon et al. (2007), and Geistefeldt and Brilon (2009) performed similar procedures to test the consistency of capacity distributions. From the figures below, we find the validation procedure is highly successful as the predicted capacities yield congestion profiles that are remarkably similar to observed cases (Figures 11 and 12 below). For q02, the randomly generated capacities tend to be large, on average, which results in predicted breakdowns that are sporadic and comparable to true, observed breakdowns. For q03, the generated capacities are comparably smaller, and the predicted breakdowns are therefore more frequent. Thus, for most days at q03, we observe the occurrence of both predicted and true breakdowns. Overall, from a visual assessment, the results validate our procedure, support the use of daily flow
maxima as an estimate of capacity, and provide evidence that capacity
distributions derived from flow maxima are suitable for prediction. We note,
however, that although the images below exhibit near day-to-day
correspondence between predicted and observed breakdowns, such a
relationship is not required as our model is expected to provide a probabilistic
prediction, or a prediction of the distribution of capacity.

Figure 11: Ten day sample of predicted breakdowns derived from fitted GEV
distribution (top) versus observed breakdowns (bottom) for location q02. Red
dots indicate breakdown, both predicted and observed.
Lastly, to further assess our GEV model predictions, the procedure above was replicated 100 times, and for each replication the number of predicted breakdowns was recorded. After averaging these 100 predicted totals, the GEV procedure predicted 553 breakdown intervals at location q02 as compared to 598 observed cases. At location q03, on average, 1578 breakdown intervals were predicted as compared to 1669 observed cases. Agreement between predicted and observed totals provides additional evidence that our model is a suitable probabilistic predictor of breakdown.

Comparison to Other Estimates of Capacity

To further illustrate the usefulness of our approach, we compare the predictive ability of a fixed (deterministic) capacity to that of our GEV
procedure. The NH DOT assumes a baseline capacity of 1800 vphpl for q02 and q03, so a fixed, three-lane capacity of 5400 vph was assumed for both locations. Predicted breakdowns, those occurring when observed flows exceed a flow rate of 5400 vph, are then compared to true breakdowns based on observed, sustained speeds below a threshold of 48 mph. Figure 13 presents results for location q02: predicted breakdowns based on a fixed capacity (top), observed breakdowns (bottom), and, as a reference, predicted breakdowns based on fitted GEV models, the herein developed stochastic approach (middle). We find that the fixed capacity is a poor estimate that yields remarkably few predicted breakdowns at location q02. For such segments that observe breakdowns at relatively low flows, a fixed capacity virtually precludes the possibility of breakdown. Next, the same comparison was performed at location q03, and, from Figure 14, we find better, but still unrealistic, results. Since flows at q03 tend to be high, a modest number of breakdowns are predicted, yet this number is very low compared to the true, observed breakdowns. Based on predictions at both locations and comparisons to the result from the GEV-derived stochastic capacity, we confidently conclude that our approach yields a more accurate representation of capacity than does a simple, fixed value.
Figure 13: Ten day sample of predicted breakdowns derived from a fixed capacity of 5400 vph (top), stochastic capacity using GEV (middle), and observed breakdowns (bottom) for location q02. Red dots indicate breakdown.
Next, we assess the predictive capability of the Weibull model approach of Brilon et al. (2005) as this has become the 'standard' approach to the estimation of stochastic capacity (Details of the approach are given in \textbf{APPENDIX A: STOCHASTIC CAPACITY}). Based on both breakdown flows (capacity estimates) and censored values, a Weibull model was fitted to data from both q02 and q03. The fitted model, the capacity distribution function ($F_C$), was then compared to a GEV distribution fit to daily maxima following the previously presented procedure. Figure 15 presents these results in addition to results from Kolmogorov-Smirnov tests of equality of distributions. For both locations, the GEV and Weibull models are found to be significantly different based on a K-S test p-value analysis.
Figure 15: Comparison of capacity distribution functions derived from Weibull models fitted to breakdown flows and GEV models fitted to daily flow maxima for locations q02 (left) and q03 (right). Results from K-S tests of equal distribution are included for each location (See inset).

Finally, we assess the predictive ability of the Weibull approach for each location by randomly generating 15-minute capacities from the corresponding fitted models. As before, in the case of congestion, a modest and constant capacity drop of 6% was assumed. When observed demands exceed randomly generated capacities, a breakdown was predicted. These predicted breakdowns were compared to true breakdowns, or breakdowns based on sustained speeds below a threshold of 48 mph. Figures 16 and 17 present these results for locations q02 and q03, respectively. Although the predicted and observed breakdowns at q02 are comparably rare, based on the fitted results (Figure 15), we suspect the Weibull model seriously underestimates low-flow breakdowns. Figure 16 provides some evidence to support this as breakdowns at q02 are only predicted for very high flows. Under-estimation of capacity at q03 is more noticeable as predicted breakdowns are substantially rarer than observed
breakdowns (Figure 17). For both locations, predictions agree with our fitted results (Figure 15) as Weibull distributions, as compared to those of the GEV, are shifted toward larger capacity values. We conclude that our approach (GEV fitted to both censored and observed maxima), in addition to having a stronger theoretical justification, is a more reliable predictor of breakdowns than the Weibull approach.

Figure 16: Ten day sample of predicted breakdowns derived from a stochastic Weibull model fitted to breakdown flows (top) and observed breakdowns (bottom) for location q02. Red dots indicate breakdown.
Figure 17: Ten day sample of predicted breakdowns derived from a stochastic Weibull model fitted to breakdown flows (top) and observed breakdowns (bottom) for location q03. Red dots indicate breakdown.
CHAPTER VI: STOCHASTIC CAPACITY IN PRACTICE

As stated, the capacity distribution function, $F_c$, may be used to assess a roadway’s performance and efficiency. With this in mind, a procedure was developed to demonstrate the probabilistic approach under a variety of circumstances.

Based on observed flows, three distinct series were simulated to mimic average (observed), reduced, and elevated daily flow patterns. To produce such series, a smooth trend was first fitted to one weekday flow pattern from location q03 to serve as a reference. Next, residuals, or deviations of observed flow from the fitted smooth, were estimated by regions (time) of comparable variability. Flow variability is observed to be highly dependent on time of day as measurements during rush hours, for example, are always more volatile than those from off-peak hours. After identifying four such regions (morning commute, midday, evening commute, and off-peak hours), normal distributions were fit to each set of residuals separately. From these fitted distributions, normal noise was randomly generated and applied to the corresponding portion of the reference curve to yield a stochastically-produced representation of average daily flow. To produce the reduced and elevated flow series, the reference curve was shifted down and up, respectively, before random noise was added. In these cases, a 10% increase/decrease in flow was applied, a change in about 500 vph for peak hours. Finally, the procedure was replicated to create several
successive days of average, increased, and decreased flow. Figure 18 below presents generated flows for one day.

Figure 18: Increased, average, and decreased (blue, black, and red lines, respectively) simulated flow derived from one daily observed flow pattern at location q03.

To assess the effect of increased/decreased flow on predicted congestion, the three series were compared to randomly generated capacities from previously fitted distributions (See CHAPTER V: BAYESIAN APPROACH, Procedure and Results) at location q03. Similar to our validation procedure (See CHAPTER V: BAYESIAN APPROACH, Validation), breakdowns are predicted when simulated flows exceed generated capacities. Figure 19 below represents ten days of simulated flow under the three scenarios described above with
predicted breakdowns marked in red. We observe that a moderate 10% increase/decrease in daily demand can have a considerable effect on predicted congestion. That is, as compared to the average flow profile (middle), increased flows result in marked increases in both morning (lower peak) and evening (higher peak) commuting hour breakdowns. When demand is reduced, however, evening breakdowns, compared to average flows, are noticeably reduced and morning breakdowns are rare. To further quantify the effect of increased/decreased demand, the procedure above was replicated 100 times, and for each replication the number of predicted breakdowns was recorded for the three flow scenarios. Compared to the average number of predicted breakdowns without increasing/decreasing demand, a 10% increase/decrease resulted in a 30% increase/decrease in the number of predicted breakdowns, respectively. These average counts of predicted breakdowns were all found to be significantly different.
Traditionally, expected traffic demand during a specific peak hour (30th highest hour, for example) is compared to a fixed estimate of capacity to assess the quality of traffic flow and performance of the facility (HCM, 2000). Clearly, the analysis of one peak hour cannot reflect a roadway's performance as numerous high demands are not considered. To overcome this, a 'whole year' analysis (WYA) was proposed by Briton (2000) and further developed by Zurlinden (2003) in which 365 days of simulated demand is compared to randomly generated capacity estimates. Our procedure, though not a full year, is akin to the WYA and results from such an approach are more representative of the life-cycle of the roadway. Furthermore, the procedure illustrates how capacity distribution functions can be used to achieve practical, tangible results. That is,
with a fitted capacity distribution, breakdown occurrence (the distribution of capacity) can be predicted for a variety of demand profiles, any of which can be easily simulated/generated to represent specific traffic patterns. If administrators believe flows will increase a specific amount (due to population increase, route closures, etc.) for a certain roadway, for example, this procedure can be implemented to assess the impact of these projected, increased flows on local traffic. Clearly, this type of analysis would offer valuable and measureable insight into the operating limits of a roadway, and could potentially assist in policy-/decision-making. In fact, a procedure similar to the WYA, where three ‘typical’ demand profiles are analyzed, has been used in Germany for years to evaluate freeway performance (EWS, 1997).
CHAPTER VII: HIERARCHICAL MODEL

After performing the above analysis, observing the correspondence between daily flow maxima and breakdown, we are left with very few data, a total of 206 observations, many of which are censored values where breakdown was not realized. As mentioned previously, Bayesian methods are ideally suited for such analyses with limited data, and the Bayesian framework offers a very natural treatment of missing or censored values. After performing a Bayesian procedure on the two locations, q02 and q03, individually, the obvious extension is then to develop a hierarchical model structure for the combined capacity data.

The term 'hierarchical' refers to a wide variety of model forms. According to Gelman et al. (2003), a model is hierarchical when 'multiple parameters [are] related by the structure of the problem.' This definition, of course, extends to cases where similar measurements are taken from different locations, as is the case with our traffic stream data at locations q02 and q03. The basic idea of a hierarchical model is that it may be advantageous to use priors that themselves depend on other parameters not mentioned in the likelihood function. These parameters will require priors, which themselves may (or may not) depend on new parameters. Eventually, the process (or hierarchy) terminates when no new parameters are introduced.

A distinct advantage of hierarchical modeling, and the basis for its application to our two data locations, is what is known as the 'pooling' effect, or
the sharing of statistical strength. Hierarchical models provide a way of pooling information from disparate groups without assuming that they belong to precisely the same population. Thus, inference for one set of data can help our inference for another. This is especially useful when one set of data is difficult to model, because of limited or missing data, for example. In our case, we have significantly more capacity data from q03 than we have from q02. In theory, and when done properly/responsibly, the pooling of data should help strengthen our estimates at q02, where very few data are observed.

Specifically, for our capacity data, we consider a hierarchical model in which data from both q02 and q03 are considered i.i.d. (independent and identically distributed) subsets of a larger population. Capacity data from both q02 and q03, as before, are assumed to follow a GEV form, and justification for such an assumption was previously established. Evidence from our previous work, model fitting using the GEV and other EV distributions, suggests capacity data are most accurately assessed using distributions with finite upper tails, thus negative GEV shape parameters. Furthermore, for both locations, such estimates were consistently and significantly negative, regardless of other model characteristics. For this reason, the shape parameters for both q02 and q03 are assumed to come from one common distribution. However, because data from q02 and q03 may be somewhat different in many other respects, location and scale parameters for the two subgroups (q02 and q03) are assumed to be somewhat disparate. Thus, under a hierarchical format, both location ($\mu$) and scale ($\sigma$) parameters for the two locations are allowed to differ as each indexed $\mu$ and $\sigma$ are respectively considered draws from a common distribution specified
by the hyper-priors (hyper-parameters). We consider this framework to be a realistic structure connecting the data from the two locations. With these structural assumptions and related parameters, a statistical power is achieved without sacrificing realism. Again, the hierarchical model format exploits such relationships by borrowing strength from data used to estimate related parameters.

Freeway capacities, $C_i$, at upstream and downstream locations $i = \{1, 2\} = \{q02, q03\}$ are assumed to be generalized extreme value distributions (GEV)

$$C_i \sim GEV(\mu_i, \sigma_i, \xi)$$

with location, scale, and shape parameters $\mu_i, \sigma_i$, and $\xi$, respectively. Collected data, $q_{ij,max}$, are maximum daily traffic flows at location $i$ for day $j = 1, ..., n_i$, each of which is classified as a censored or uncensored capacity ($c_{ij}$) value based on the previously established definition. For our data, $n_1 = n_2 = 206$. Necessary for our Bayesian model approach, the parameters $\mu_i, \sigma_i$, and $\xi$ are assumed to follow prior distributions as follows.

$$\xi \sim Uniform(a_\xi, b_\xi)$$

where $a_\xi$ and $b_\xi$ are defined to be -1 and 0, respectively, based on prior information, and

$$\mu_i \sim Normal(\theta, \tau)$$

where $\tau$ is a precision parameter set to arbitrarily small number (0.001), and

$$\theta \sim Uniform(a_\theta, b_\theta)$$
where $a_\theta$ and $b_\theta$ are defined as 0 and 10, respectively, based on prior information, and

$$\sigma_i \sim Uniform(a_\sigma, b_\sigma)$$

where $a_\sigma$ and $b_\sigma$ are defined as 0 and 5, respectively, based on prior information obtained from previous model-fitting.

To gain a better appreciation for our model structure, the hierarchical form, we have the following illustration.

**Figure 20:** Flow-chart of hierarchical structure of parameters.

---

**Procedure and Results**

Using the model form discussed above, the standard MCMC resampling scheme was implemented through the OpenBUGS statistical software. In this case, a relatively large number of iterations, 10,000, were produced with the
initial 5,000 discarded as a burn-in period. Convergence and independence from the starting values were checked by CODA (distributions, traces, etc.), the standard tools in such cases. In all cases, starting values for the sampling scheme were generated from the defined prior distributions. Model diagnostics are unremarkable as parameter traces are well-mixed and convergent. Generally speaking, the hierarchical structure is particularly well-behaved. Below we have fitted distributions, densities and CDFs (where CDF = \( F_c(q) \)), for several sets of parameter estimates and corresponding median fitted distributions (as well as upper and lower 5% fitted distributions). To illustrate the effect of the hierarchical format on parameter estimates, Bayesian results obtained from individual model fitting are presented as well.
Figure 21: Hierarchical (right) and non-hierarchical (left) densities (top) and cumulative distributions (bottom) for location q02 with median curves (thick black lines) and upper/lower 5% curves (thin black lines).

Location q02

Density

Hierarchical Density

CDF

Hierarchical CDF
As evident by Figure 21, there is much to be gained from the Bayesian format. Most obvious, results from location q02 (Figure 21) benefit significantly from the sharing of data. That is, hierarchical results, when compared to those obtained from a non-hierarchical format, yield a distribution with noticeable decrease in uncertainty. Without using a hierarchical model, parameter estimation for location q02 is dominated by the high number of censored values, and the resulting distribution of capacity is highly diffuse. The hierarchical model, in theory, addresses this by borrowing information from q03 to achieve increased
precision, although this model structure is not a perfect solution to the small sample size of location q02. Despite the fact that, under the hierarchical model format, the location parameters ($\mu_1$ and $\mu_2$) of the GEV distributions are allowed to differ by location (q02 or q03), this sharing of information does allow for some 'shrinkage' toward a common value. If we believe our data are obtained from freeway segments with dissimilar flow characteristics, and if we feel the associated distribution 'centers' should be different, this shrinkage may be a problem.

To further assess the effect of the hierarchical structure, our posterior results are compared to those obtained from our previous, non-hierarchical model (See below for posterior estimates and credibility intervals for both model forms). From the table, we make several observations. First, the q02 hierarchical location parameter estimate is shifted up from its non-hierarchical counterpart. This change is possibly due to the combined effects of data pooling and a high number of censored values at location q02, but may simply be a byproduct of dependent parameters, where shape and scale parameters may compensate for changes in the location parameter. In any event, as compared to the non-hierarchical procedure, the hierarchical procedure results in a capacity distribution that is slightly less sensitive to breakdown, an even more realistic distribution when considering the high flows and low occurrence of breakdown at q02. Next, when comparing model forms, the hierarchical model has very little effect on q03 parameters, estimates or credibility intervals, although some increased precision is achieved. This is somewhat expected as q03 has relatively few censored values and therefore gains very little from the pooling of data.
Lastly, the effect of the hierarchical model is most noticeable in the scale and shape parameters for q02, which, again, may simply be a result of high correlation. Clearly, from Figure 21 and credibility intervals reported in Table 1 (below), q02 benefits from the hierarchical model as variability in the posterior estimates and resulting distribution is significantly reduced. That is, for q02, compared to the non-hierarchical results, hierarchical location and scale parameters are more precise, and the shape parameter, which is shared, is significantly less variable. This increased precision results in a less ambiguous distribution for q02, especially towards the tail of the distribution. We acknowledge that this positive result is a byproduct of our model and useful if (and only if) the model structure and prior specifications are well-founded.

Table 1: Posterior estimates and 90% credibility intervals (parentheses) for both hierarchical and non-hierarchical models.

<table>
<thead>
<tr>
<th></th>
<th>μ hierarchy</th>
<th>μ non-hierarchy</th>
<th>σ hierarchy</th>
<th>σ non-hierarchy</th>
<th>ξ hierarchy</th>
<th>ξ non-hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>q02</td>
<td>5.36</td>
<td>5.31</td>
<td>1.50</td>
<td>1.68</td>
<td>-0.60</td>
<td>-0.37</td>
</tr>
<tr>
<td></td>
<td>(5.13, 5.65)</td>
<td>(5.03, 5.69)</td>
<td>(1.28, 1.75)</td>
<td>(1.32, 2.19)</td>
<td>(-0.67, -0.52)</td>
<td>(-0.64, -0.07)</td>
</tr>
<tr>
<td>q03</td>
<td>5.21</td>
<td>5.28</td>
<td>1.36</td>
<td>1.39</td>
<td>-0.60</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td>(5.02, 5.37)</td>
<td>(5.08, 5.48)</td>
<td>(1.23, 1.54)</td>
<td>(1.24, 1.59)</td>
<td>(-0.67, -0.52)</td>
<td>(-0.72, -0.53)</td>
</tr>
</tbody>
</table>

To further assess our results, we have a side-by-side comparison of q02 and q03 cumulative distributions obtained from the hierarchical model fitting procedure (See Figure 23). Additionally, for each plot we mark the $q_{50}$ value, the flow value corresponding to a 50% chance of breakdown or a substantial risk of failure. The concept of the $q_{50}$ value is akin to the median lethal dose, LD$_{50}$
(abbreviation for ‘lethal dose, 50%’), used in the field of toxicology to represent the dose of a toxin/pathogen required to kill half the members of a tested population. As LD<sub>50</sub> is commonly used to quantify toxicity, we use q<sub>50</sub> as a descriptive statistic of capacity. We observe that q<sub>50</sub> for q02 is larger than that of q03 by about 200 vehicles per hour (flows are reported in thousands), not an insignificant amount. Furthermore, upper quantiles (beyond the median, say) at q02 are larger than corresponding quantities at q03, and this difference increases with probability. So, despite some shrinkage of distributions from q02 and q03 towards each other, the hierarchical format does preserve some of the distinguishing features.

Hierarchical results indicate that for equivalent values above the median, q02 has slightly lower probabilities of congestion than q03, and the magnitude of these differences increase with flow. This suggests that q02 more effectively manages higher demands, or that q03 is more sensitive to increases in flow. This makes intuitive sense as q03 is a physical bottleneck that experiences exiting/entering and merging traffic, while, on the other hand, q02 is a seemingly featureless stretch of road with none of these triggers. That is, in the absence of congestion-related features, drivers at q02 are more likely to maintain freeflow speeds, avoid stop-and-go conditions, etc. Drivers at q03, however, because of a more complicated structure, are prone to erratic behavior associated with breakdown occurrence. These results agree with our understanding of both the collection sites and freeway/roadway dynamics, and provide more evidence that our model form is realistic.
Lastly, while it is not of primary interest to our study, we observe the prior distribution of the hyper-parameter $\theta$, the mean of the distribution for location parameters $\mu_i$. A property of hierarchical models is that very large posterior precision of the hyper-prior implies that the sub-populations are drawn from a common distribution with very small variance. In such cases, the model has identified identical means or shrinkage toward a 'grand' mean due to the fact that variation from the population mean is random noise. On the other hand, very small posterior precision of the hyper-prior implies that the sub-populations are drawn from distributions with different means, or that there is little shrinkage toward an overall mean. From our model-fitting results, the posterior mean of $\theta$ is found to be 5.07 with a corresponding standard deviation of 2.91, somewhat diffusely distributed. This is a desirable result as it suggests the two location parameters are different enough to force their distribution to be spread over a
fairly wide range of values, or that a common estimate will not suffice. This result confirms our prior belief of disparate datasets with different location parameters and provides some evidence that the imposed structure of our hierarchical model is appropriate.
CHAPTER VIII: CONCLUSION AND DISCUSSION

In order for capacity estimates to be used in modeling and decision-making, it is critical that capacity be clearly defined and accurate (Minderhoud et al., 1996). With this in mind, the concept of stochastic capacity was investigated, and a number of techniques were examined to supplement the current methods of estimation and to offer new perspectives. Primarily, in this work, we have offered several 'adjustments' to the current modeling of stochastic capacity that, we feel, are improvements. First, we have slightly altered the method in which capacity data are extracted from traffic stream data. Aided by functional data models (FDA), traffic stream traces allow us to simultaneously mitigate noise in the data and achieve a more precise view of breakdowns. When daily maxima correspond to breakdowns, maximal flows are considered capacity data; otherwise, these maxima are considered censored (right-censored) estimates. Such estimates of capacity, maximum flows, may then be suitably approximated by the generalized extreme value (GEV) distribution for block maxima. To introduce the censored values, and to address small sample sized, a Bayesian framework was implemented using semi-informed priors base on previous work and our understanding of traffic flow data. Additionally, a hierarchical Bayesian model was introduced to pool data and offer a transferable format amenable to diverse freeway sections. In general, we offer a very modern approach to capacity estimation as our procedure uses
new techniques such as extreme value theory and FDA as well as computer-driven Bayesian/MCMC sampling.

Model validation provides evidence that the GEV form fitted to both censored and observed values suitably approximates daily flow maxima. Further validation of the procedure in general provides evidence that the combined application of methodologies (extreme value analysis, censoring, FDA) can yield accurate distributions of capacity.

Extending our procedure to a hierarchical model format proves especially beneficial as the high number of censored values at location q02 is mitigated by the sharing of information between both locations. From the hierarchical format, capacity distributions are realistic estimates of efficiency at the two locations. Specifically, as compared to q03, the capacity distribution derived from the hierarchical model is less sensitive to demand. That is, at comparable flow values, congestion is less likely to occur at q02 than at q03. This is perfectly in keeping with observed traffic flow at q02 where breakdown is especially rare.

Going forward, there are numerous areas into which our work can be extended, evidence of the flexibility, usefulness, and applicability of our modeling scheme. Much of this work involves broadening our hierarchical model to include additional information that, up to this point, has been ignored or considered unchanging. First, this analysis considered the roadway as a homogenous unit and made no distinction between lanes (left, right, median, etc.). As our work focuses on high-level travel behavior and the performance of the roadway in general, this approach makes sense; but capacity can certainly
be calculated on a finer level, for each individual lane. Identifying capacity
distributions for individual lanes is a logical extension of this work as evidence
suggests that traffic behavior between lanes may be quite dissimilar (See, for
example, Cassidy and Bertini, 1999). The New Hampshire Department of
Transportation (NHDOT) now collects individual lane data along I-93 which will
allow for such refinement.

Second, following the procedure of Brilon et al. (2005), we may introduce
external conditions into our hierarchical model. Ponzlet (1996) demonstrated
that capacities vary according to external conditions such as dry/wet road
surfaces or daylight/darkness. Using the PLM/Weibull approach, Brilon et al.
(2005) found that on a wet road surface capacity was reduced by around 11%,
for all freeway sections in their analysis. On these same roads, however, it was
clearly found that darkness did not shift the capacity distributions, contrary to the
results of Ponzlet (1996). So long as detailed, local records are available,
identifying the effects of weather variables on capacity distributions is feasible.
Also, since a precise time stamp is available for every traffic observation, the
effect of daylight/darkness (and the effect of sun glare) is possible.

Third, the effect of driver behavior should be investigated. Using the
current data sets and models for q02 and q03, covariates for time-of-day, day-
of-week, and month could be added to our model. During our initial data
inspection, most breakdowns were observed during either morning or evening
commuting hours (rush hours), but the time-of-day effect should be examined
more thoroughly. Certainly, obtaining a model that can identify the probability
of congestion by both time and location would be very beneficial. In their analysis of traffic flows, Tebaldi et al. (2002) introduced a day-of-the-week component in their hierarchical models. The authors, however, concede that a more formal and thorough modeling of this effect is required. In our analysis, days are considered exchangeable and indistinguishable from one another. In reality, some differences likely exist between days. Lastly, a monthly or seasonal (true seasonal) effect may be pursued. In New Hampshire, a vacation destination during multiple seasons, we suspect that a large influx of tourist travel adversely affects the traffic along I-93, and thus the distribution of capacity. Unfortunately, long series of data are required to accurately assess this claim, and our traffic data collected over nine months is insufficient to do so.

Fourth, capacity distributions could be established for more diverse road sections. In our analysis, two somewhat diverse freeway sections, q02 and q03, were included to offer some proof of transferability. While the analysis provided anecdotal evidence that different structures yield different capacity distributions, a more thorough and extensive analysis is required. Ponzlet (1996), in addition to studying other external conditions, found that capacity may vary based on the prevailing purpose of the freeway section. For example, long-distance and commuter traffic sections are expected to have distinct capacity profiles. Brilon et al. (2007) established that the capacity of an intersection may also be treated as a random variable, and the capacity distribution may be examined for two-lane undivided highways, structures that are very common in rural states such as New Hampshire.
Fifth, as stated by Minderhoud et al. (1997), a primary advantage of the capacity distribution function is that it allows for the choice of a capacity value based on certain quality considerations. That is, given a distribution of capacity, and thus the probability of congestion for a range of flow values, a capacity estimate (single value) can be chosen by planners/administrators based on an 'acceptable' risk of congestion/breakdown. Capacity distribution models may then be used in conjunction with models to predict traffic delay times, thus associating flow values with both probability of breakdown and expected breakdown duration. Extending capacity analyses to risk management in such way seems very natural and initial work in this area is already being undertaken.

Sixth, we may wish to extend our hierarchical platform with the inclusion of an effect for spillback (spillage or tailback), an issue that was not completely resolved. Up to this point, this queuing effect has been disregarded as only very few cases were identified as potential spillback. That is, the presence of spillback cannot be supported without more specific measurements, such as traffic stream data measured more incrementally between locations. However, such information may be available, and a flexible model form could account for the spillback effect. Typically, measurements caused by spillback are omitted from capacity estimates, so a model could be devised in which the sample size is treated as a stochastic element. Knowing which measurements are potentially spillback events, we could select a sample of these cases with some predetermined probability. Although this would not be particularly advantageous in our case, for q02 and q03, a hierarchical model with spillback
adjustment would be one step closer to an adaptable network model for freeway capacity.

Seventh, the transferability of our procedure remains unresolved, and it is not clear that process can be suitably applied to locations dissimilar to those used in this analysis. That is, although locations q02 and q03 are somewhat diverse, they naturally share some characteristics as they are adjacent segments of the same freeway. It is possible that data from a two-lane segment, for example, behaves much differently and is not amenable to our procedure. Ideally, the acquisition of traffic stream data from dissimilar roadway segments will allow for further testing and evaluation of our approach.

Many researchers are urging traffic operations practitioners and the profession to adopt probabilistic methods of capacity estimation (See, most notably, Lorenz and Elefteriadou, 2001; Brilon et al., 2005). While no one methodology is agreed upon, we have shown in the preceding chapters that our GEV approach is an easy-to-use analysis tool for evaluating performance of a highly variable and complex freeway breakdown process. The approach removes the traffic signal noise and thus makes it possible for traffic managers, operators, and planners to understand and come to grips to what is happening in the field. By no means is our approach perfect as numerous issues remain unresolved and require further research. However, we have offered a number of new perspectives on the current view of stochastic capacity that ideally will inspire new research.
APPENDICES
APPENDIX A: STOCHASTIC CAPACITY

In their study of freeway capacity and breakdown, Lorenz and Elefteriadou (2001) analyzed speed and volume data collected at two freeway bottleneck sites in Toronto, Ontario (Canada). The authors develop preliminary models for the probability of breakdown as a function of flow rates. For both sites, increasing trends are observed where probability of breakdown increases with flow. Moreover, breakdown was observed to occur over a range of flow rates, some higher and some lower than traditional capacity estimates. These results illustrate the fundamental concept of stochastic capacity and refute the deterministic notion that defines a predictable relationship between breakdown and a fixed threshold. The authors recommend that the HCM incorporate probability of breakdown component in the definition of capacity.

Persaud et al. (2001) pursue stochastic capacity by identifying the probability of breakdown as an increasing function of volume at the critical location. To define this function, the authors use a simple logistic regression model calibrated to flow values measured immediately prior to breakdown. This approach yields a distribution of capacity from which probability of breakdown may be estimated for any flow. This probability-of-breakdown method was developed as a basis for ramp-metering in which volumes are managed to maintain performance of the roadway.

Of course, when investigating stochastic capacity, we would be remiss not to mention the innovative work of Brilon et al. (2005). This work introduces the capacity distribution function, the derivation and theoretical basis for which is described throughout the remainder of this section. To estimate this function (the capacity distribution function), the authors use the Product Limit Method (PLM) (Kaplan and Meier, 1958) based on the general approach for statistical analysis of lifetime data (Lawless, 1981), and originally intended to estimate survival functions. Later, Minderhoud et al. (1996) would formulate the PLM for the capacity analysis of freeways, and Zurlinden (2003) would adopt the PLM for congestion analysis in Germany.

The Product Limit Method approach considers breakdown as a ‘failure’, and the distribution of capacity (c) is treated as analogous to that of lifetime (T) (van Toorenburg, 1986). First, the non-parametric PLM used to describe survival functions is given by the expression:

\[
S(t) = \prod_{j: t_j < t} \frac{n_j - d_j}{n_j}
\]

where \(S(t)\) is the estimated survival function, \(n_j\) is the number of individuals with a lifetime \(T \geq t_j\), and \(d_j\) is the number of deaths at time \(t_j\). We may substitute traffic volume (\(q\)) with time (\(t\)), the concept of breakdown with death, and capacity
variable \( C \) with lifetime variable \( T \) to create an expression for capacity. By further observing that the lifetime distribution function is the compliment of the survival function, or that \( F(t) = 1 - S(t) \), we have an expression for the distribution of capacity data as follows:

\[
F_c(q) = 1 - \prod_{i: q_i < q} \frac{k_i - d_i}{k_i}, i \in \{B\}
\]

where \( F_c(q) \) is the capacity distribution function, \( q \) is traffic flow, \( q_i \) is traffic flow in interval \( i \), \( k_i \) is the number of intervals with volume \( q \geq q_i \), \( d_i \) is the number of breakdowns at volume \( q_i \), and \( \{B\} \) is the set of breakdown intervals. Such an approach is required as capacity data cannot easily be estimated because it cannot be directly observed. The PLM can then be used to estimate this capacity distribution function based on samples that include both uncensored (breakdown is observed) and censored (breakdown not observed, capacity is greater than demand) data. That is, using the PLM, uncongested flows with higher flow rates than the lowest observed capacity rate contribute to the capacity estimate since this observation gives additional information about the capacity value.

In order to estimate the distribution of the capacity, the capacity observations are assumed to be identically and independently distributed with probability density function \( f_c(q) \), probability distribution function \( F_c(q) \), and probability survival function \( S_c(q) = 1 - F_c(q) \). Then, as per Minderhoud et al. (1996), the likelihood is given by:

\[
L = \prod_{i=1}^{n} f_c(q_i)^{1-\delta_i} \cdot S_c(q_i)^{\delta_i}
\]

where \( n \) is the number of observation periods, and \( \delta_i \) is 0 or 1 for uncensored or censored flow values, respectively.

In application, this estimation is hindered by the fact that a 'complete' distribution can rarely be realized. To overcome this, the estimation is performed parametrically where a known distribution function is assumed. For highways in Germany, Brilon and Zurlinden (2003) investigated the utility of various 'plausible' distributions such as the Weibull, Normal, and Gamma. For capacity estimation, it was found that the Weibull distribution was a 'very good approximation' to the nonparametric estimation (Zurlinden, 2003). Thus, for capacity data we have:

\[
F_c(q) = 1 - \exp\left(-\left(\frac{q}{b}\right)^a\right)
\]

and the resulting likelihood function is given by:

\[
L = \prod_{i=1}^{n} \left(a \cdot b^{-a} \cdot q_i^{a-1} \cdot e^{-\left(\frac{q_i}{b}\right)^a}\right)^{1-\delta_i} \cdot \left(1 - e^{-\left(\frac{q_i}{b}\right)^a}\right)^{\delta_i}
\]
where \( a \) and \( b \) are the Weibull shape and scale parameters, respectively, which may be estimated via maximum likelihood techniques.

Under the survival analysis framework, the capacity distribution function, \( F_C(q) = \text{Prob.}(C \leq q) \), corresponds to the probability of breakdown at the flow value \( q \), and where flow values resulting in breakdown are designated as capacity data, \( c_i \). The approach is based on the belief that each roadway has an instantaneous or momentary capacity at any given time. Since every flow rate greater than the capacity must, by definition, cause a breakdown, this cumulative distribution represents the associated probability. The PLM/Weibull approximation is generally accepted as the most realistic approach to estimating capacity, and the work of Brilon et al. (2005) is ubiquitous in traffic studies.

In a related work, Geistefeldt and Brilon (2009) compared the performance of the capacity distribution function to the method of direct estimate of breakdown probability by flow grouping (the most pragmatic/logical approach to stochastic capacity). Empirical comparisons were based on data from German freeways, and the consistency of the models was assessed using a macroscopic simulation model. Empirical analyses confirm that the two methods give significantly different estimates, particularly at high volumes. For these high quantiles, the PLM/stochastic approach using censored data performs much better and achieves significantly more precise estimates. The direct probability method, on the other hand, tended to underestimate breakdown probability at high traffic flows. The authors enthusiastically recommend the PLM approach for estimating capacity distribution functions. Similar conclusions are stated in Geistefeldt (2010), another related work in which the same models are compared.
APPENDIX B: SPEED THRESHOLD SELECTION

To identify breakdown, we have adopted a traditional approach that relies on a fixed speed threshold to identify the transition from freely flowing traffic to congested conditions, or a breakdown. Numerous works have used such a fixed speed value as a threshold between congested and freely flowing traffic (for example, Lorenz and Elefteriadou, 2001; Yeon et al., 2009; Geistefeldt and Brilon, 2009; Habib-Mattar et al., 2009; Brilon et al., 2005). Typically, deciding on the exact threshold is based on empirical evidence, or based on an understanding of traffic behavior in general. The threshold used in this analysis is simply derived from inspecting flows versus speeds and identifying distinct freeflow and congested regimes (See Figure 24 below). That is, we can visually assign realistic lower and upper speed boundaries to the freely flowing and congested regimes, respectively, and then simply find the midpoint between them. This midpoint will serve as a threshold to distinguish traffic regimes and identify congested states/breakdown occurrence. In the absence of a more standard approach, this is an empirically based approach that avoids arbitrarily selecting a speed boundary. After identifying upper and lower speed boundaries respectively at 60 and 35 mph, a midpoint is identified at 48 mph, a reasonable threshold to distinguish breakdowns. In fact, this speed threshold is similar to those of Geistefeldt and Brilon (2009), Yeon et al. (2009), and Brilon et al. (2005) who used fixed values of 47 mph (75 kph), 50 mph, and 43 mph (70 kph), respectively. We note that this analysis was performed at q03 only as it is expected that q02 will yield similar results.

Figure 24: Plot of observed flows versus speeds at location q03. Red lines separate distinct regimes of congested and freely flowing traffic while blue line represents midpoint between these boundaries used as a breakdown threshold.
APPENDIX C: COLLECTION SITES

Based on the 15-minute aggregation, and the aforementioned breakdown criteria, data from several locations (devices) along I-93 were considered in this analysis. But, many traffic stream data (time series of flow and speeds) were deemed unreliable, and breakdowns identified at many locations were likely attributable to spillback and thus unusable. In the end, just three consecutive locations, q02, q03, and q04, from the I-93 collection sites will be regarded as promising sources of capacity data. Descriptions of these sites, details of their roadway characteristics, and evidence of their appropriateness (or inappropriateness) for a capacity analysis will now be presented.

Locations 3 and 4 (devices q03 and q04)

First, we observe device location 3 (q03), a site that observes both daily breakdowns and heavy demand. Device q03 is located just north of an exit ramp, exit 1, and just south of an entrance ramp to I-93 (See Figure 25). Leading up to exit 1, and travelling northbound, I-93 is three lanes of ‘ideal roadway’ (12-foot wide travel and breakdown lanes, etc.). Exit 1 is a major artery to the Salem, NH area and serves a large mall, a busy downtown/consumer area, and a dense population. From I-93, exit 1 immediately forms a two-lane off ramp to service this area. Immediately after the exit, I-93 remains a three-lane highway up to measurement device q03, less than 1,500ft after the exit ramp. Just north of the device (downstream), I-93 is physically constricted to two-lanes. I-93 continues as two-lanes through the on-ramp just north of q03 (less than 1,500ft) and for several more miles northbound.

Figure 25: Layout of I-93 at locations q02-q04 illustrating number of lanes, exit/entrance ramps, and approximate distances between structures.

Device location 4 (q04) is located about 1 mile north (downstream) of q03. Device q04 is located immediate north of exit 2 and immediately south of
an on-ramp to I-93. Exit 2 is a busy off-ramp servicing the same general Salem, NH area as exit 1 (upstream). Distances between q04 and both exit and entrance ramps are less than 500ft (See Figure 25). The positioning of the device in relation to on-and off-ramps is similar to that of q03, but the physical narrowing of the roadway observed at q03 does not exists at q04. That is, the physical constriction of I-93 occurs about a mile upstream of q04 and I-93 remains two lanes through device q04 and beyond. By HCM standards, and because of their location to off-ramps, q03 and q04 are considered merge influence areas and weaving is deemed ‘discretionary.’ That being said, merging, especially in the presence of heavy demand, is likely a contributor to bottleneck formation, or congestion/breakdown (Cassidy and Bertini, 1999; Elefteriadou et al., 2005).

Cassidy and Bertini (1999) found that bottlenecks always formed at the same location and that they were always activated by a sustained surge in the flow from upstream. Thus, traffic transitioned from free flow to queued conditions in a predictable way; the queues formed at an inhomogeneity, the bottleneck, due to increased flows. According to Horowitz and Bertini (2007), a bottleneck is a location on a freeway that separates downstream freely flowing traffic from queued upstream congestion. Daganzo (1997) considers a bottleneck activation to occur when these same conditions are met and deactivation to occur when downstream traffic becomes queued and spills back into the active bottleneck or when demand decreases at the bottleneck site. Furthermore, in their analysis of freeway bottlenecks and speed drop sequences, Ogun and Banks (2005) surmised that there is rarely a single bottleneck location within critical freeway sections, which suggests that many bottlenecks should be thought of as extended sections rather than points or isolated segments. This extends the definition of a bottleneck quite literally. Most agree, however, that a bottleneck exists when congestion occurs at a specific location and repeatedly at the same time.

Based on these definitions of bottlenecks, the area immediately downstream of q03 conforms to the traditional concept of bottleneck where a physical narrowing disrupts freely flowing traffic. That is, after exit 1, I-93 transitions from three lanes to two lanes. In addition to this constriction of the roadway, an on-ramp to I-93 is located just downstream of device q03 and the reduction of lanes. Not only does this increase demand on the segment by supplying the roadway with an influx of vehicles, but weaving and merging on entering traffic likely adds to congestion. We are confident that a bottleneck is formed by these combined effects, both the physical narrowing of lanes exacerbated by effects of the on-ramp, and is initiated at the point where the on-ramp and the 2-lanes of I-93 intersect, just downstream of q03 (See Figure 25). Lastly, we note that the positioning of device q03 is an ideal location as data collected here fit perfectly to the theoretical framework of flow-based capacity analyses (Minderhoud et al., 1997). As stated by Brilon et al. (2005), the capacity of the freeway segment is analyzed most precisely at or slightly upstream of a bottleneck.
Like q03, demand at q04 is often heavy and breakdowns are experienced frequently. In fact, from field research, queues typically formed along the entire distance (about 6,000ft) between q03 and q04 on a daily basis. This long queue between q03 and q04 is ultimately relieved (and terminated) by the off-ramp at exit 2 (located immediately before device q04) where downstream traffic usually returns to freely flowing conditions. We suspect that inadequate exit ramp capacity at exit 2 results in the formation of an additional bottleneck at the exit prior to q04. That is, because it cannot sufficiently process the traffic demand, the exit ramp is an 'operationally influenced' deficiency (USDOT/FHWA, 2012), the cause of routine 'recurring' congestion, a recurrent bottleneck. Because of the positioning of device q04 after the exit-ramp (exit 2), it is difficult to provide evidence to support this claim, aside from the anecdotal evidence observed first-hand. Ideally, an additional measurement device would be located between q03 and exit 2. However, at q04, we do observe a high number of congestion events, or speeds below our threshold. In these cases, q04 is likely detecting vehicles immediately downstream of a bottleneck as they accelerate to freeflow velocity, but have yet to attain it. These are essentially the after-effects of the terminated bottleneck. For these reasons, we will not further regard data from q04 in our capacity estimation.

**Location 2 (device q02)**

Next, we examine device location q02, a much simpler structure that experiences far fewer breakdowns. Device q02 is located about 1 mile south of q03 and is not in the vicinity of any exit- or on-ramps to I-93 (See Figure 25). Prior to and beyond q02, I-93 is three lanes of 'ideal roadway' with 12-foot wide travel and breakdown lanes and no physical narrowing. Device q02 is classified by the HCM as a basic freeway segment. This means q02 is a limited access facility with high design standards, is outside the influence area of on- and off-ramps, is one direction only, experiences optional lane changing, etc. Clearly, no physical/structural bottleneck exists here, yet a fair number of breakdowns are observed (about 100).

Although 'spontaneous breakdown' (either microscopic of macroscopic) away from bottlenecks or incidents is conceivable/possible (Polus and Pollatschek, 2002; Kerner and Rebhorn, 1997), it is not commonly observed and prediction of such events is complicated (Habbib-Mattar et al., 2009; Banks, 1990, 1991a, 1991b). Location 2 observed about 100 breakdowns, corresponding to about one every other day. These breakdowns are too numerous and are too frequent to occur spontaneously away from a bottleneck. Rather, assuming these breakdowns are not caused by upstream spillback, it is likely that q02 qualifies as a bottleneck under an 'expanded' definition in which a narrowing or obstruction is not required. The Federal Highway Administration presents this expanded definition of [recurrent] bottleneck as 'a localized section of highway that experiences reduced speeds and inherent delays due to a recurring operational influence or a nonrecurring impacting event' (USDOT/FHWA, 2012). Thus, recurring bottleneck conditions may result from weaving, sun glare, a vertical climb, etc. Unfortunately, the exact cause of the
bottleneck at q02 has yet to be identified, but breakdowns here are likely attributable to secondary roadway characteristics in the presence of exceptionally high demand.

As discussed by Brilon et al. (2005), the capacity function estimation would be very restrictive to specific physical criteria if it were only applicable to data extracted from physical bottlenecks. Regler (2004) extended the approach and applied the same PLM/Weibull estimation to locations where no physical, distinct bottleneck is apparent. Similarly, for these cases, again, spillback (tailback) must be identified and thusly removed from consideration. This extension is an important result as the approach is then suitable for a variety of freeway sections, so long as breakdown is observed. However, we suspect that these 'extended' locations are simply recurrent bottlenecks caused by factors other than lane reductions/narrowing, i.e. the expanded definition of 'bottleneck' put forth by the FHWA. In any event, whether we expand our definition of bottleneck to include q02, or whether we rely on the generalization of Regler (2004), data acquired from q02 is seemingly suitable for a capacity study.
APPENDIX D: FUNCTIONAL DATA MODELS

Suppose that our functional data, \( Y(t) \), is observed through the model
\[
Y(t_i) = X(t_i) + \varepsilon(t_i)
\]
where the residuals, \( \varepsilon(t) \), are independent of \( X(t) \). We can then express the original signal \( X(t) \) in terms of a linear smoother:

\[
\hat{X} = \sum_{i=1}^{n} s_{ij} y_i
\]

where \( s_{ij} \) is the weight that the point \( t_j \) gives to the point \( t_i \), and \( y_i \) is the observed value of the variable \( y \) at point \( t_i \). Typically, basis methods are used for smoothing. A basis is a set of known functions such that a linear combination of some number of these functions can sufficiently approximate another function, a set of functional building blocks (Ramsay et al., 2009). So, for a set of basis functions, \( \phi_k \) where \( k = 1, ..., K \), a function \( x(t) \) may be expressed (in basis function expansion) as

\[
x(t) = \sum_{k=1}^{K} c_k \phi_k(t),
\]

where \( c_1, ..., c_k \) are simply the coefficients of the expansion. We note that the above expression refers to the basis function expansion of the value of function \( x \) at argument value \( t \), but the expansion of \( x \) is better expressed as

\[
x = \sum_{k=1}^{K} c_k \phi_k.
\]

We have opted to smooth our 15-minute flow aggregates via a Fourier series basis, a recommended approach for periodic data, such as speed and flow traffic measurements. The Fourier series is

\[
\phi_1 = 1
\]
\[
\phi_2 = \sin(\omega t)
\]
\[
\phi_3 = \cos(\omega t)
\]
\[
\phi_4 = \sin(2\omega t)
\]
\[
\phi_5 = \cos(2\omega t)
\]
\[
\vdots
\]

which is often used for functions that repeat over some period \( T \). In such cases, the constant \( \omega \) is related to \( T \) by the relation \( \omega = 2\pi/T \). In most cases, the value
of $T$ can simply default to the range of $t$ values spanned by the data (Ramsay et al., 2009), which leaves only the number of basis functions, $K$, to be determined.

Figure 26: Cyclic trend observed in traffic flows at location q03.

Literature on FDA rarely agrees on the optimal basis and the way parameters are estimated. It is generally accepted, however, that basis functions should be chosen to reflect the characteristics of the data. In our case, our flow data follows a distinct, daily pattern (See above), so, as recommend for periodic data (Ramsay and Silverman, 2005), Fourier basis functions were used.

For our functional data models, 15-minute aggregates were used as they provide realistic estimates of capacity and are generally better data. If we were to use very short intervals, we would likely end up modeling individual driver behavior instead of general traffic stream characteristics. So, can we simply smooth the 2-minute data more to achieve more realistic results? Unfortunately, such an approach does not overcome the detrimental effect of the volatility in the shorter intervals.

Lastly, for our 15-minute data, FDA models primarily serve to reduce noise in the traffic stream data. That is, we believe the 15-minute data provide realistic trends, but some aggregates based on few observations yield unrealistic traffic measurements. So, to maintain the observed trend while mitigating especially noisy data, minimally smooth traces were fitted. To obtain smooths that maintain much of the daily pattern, FDA models using 21 Fourier basis functions were fitted to each day's traffic stream data individually (using more basis functions yielded identical results). For the majority of cases, however, FDA smooths for flow only slightly underestimate the original 15-minute aggregated flow maxima.
APPENDIX E: MCMC SAMPLING

For our Bayesian models, OpenBUGS statistical software is used to generate samples from the joint distribution of the parameters. OpenBUGS operates under a Gibbs sampling framework where conditional distributions of each parameter given all the others (these are known as full conditional distributions) are successively sampled. Gibbs sampling is a special case of the Metropolis-Hastings (Hastings, 1970) algorithm and is considered to be widely applicable to a broad class of Bayesian problems. The work of Gelfand and Smith (1990) sparked a renewed interest in the Gibbs sampler and brought about a major increase in its application. Furthermore, modern advances in computing have allowed for computer-intensive MCMC algorithms to be easily implemented. A simple overview of the Gibbs framework is offered below.

Very generally, for a parameter vector, \( \theta = (\theta_1, \theta_2)' \), the posterior density \( \pi(\theta|data) \) of the parameter vector is estimated using the following Gibbs sampling framework.

1.) Start with initial values: \( \theta^{t-1} = (\theta_1^{t-1}, \theta_2^{t-1})' \).
2.) Sample \( \theta_1^t \) from \( Pr(\theta_1|\theta_2^{t-1}, data) \).
3.) Sample \( \theta_2^t \) from \( Pr(\theta_2|\theta_1^t, data) \). This yields \( \theta^t = (\theta_1^t, \theta_2^t) \).

More generally:

1.) Initialize the sampler with starting values \( \theta^{(0)} \).
2.) Let the sampler run, generating \( \theta^{(1)}, \theta^{(2)}, ... \).
3.) Under a wide set of conditions, as \( t \to \infty \), each Gibbs sample, \( \theta^{(t)} \), can be viewed as samples from the posterior density \( \pi(\theta|data) \).

As the sampler updates, it moves away from the initial values, providing a "random tour" of the parameter space (possibly high-dimensional parameter space), visiting locations in the space with frequencies proportional to the posterior density. Formally, the output of the sampler forms a ergodic Markov chain on the parameter space for \( \theta \), with transition probabilities such that the 'limiting' distribution of the sampler is the posterior density \( \pi(\theta|data) \) of direct interest. Estimates using this approach get 'better' with increased samples (more samples), so modern computing power makes this approach to estimation and inference possible. In our case, for the GEV, \( \theta \) is a three-dimensional parameter comprised of \( \mu, \sigma, \) and \( \xi \), the respective location, scale, and shape parameters.

For censored values, the Gibbs sampler and MCMC approach is particularly well-suited. Information from censored observations is incorporated into the sampling scheme as follows.
1.) For a current value of the parameters, $\theta^{(m)}$, a vector $Y_c^{(m+1)}$ for the
censored data is sampled from $Pr(Y_c | Y_0, \theta^{(m)})$ where $Y_c$ represents the
censored values and $Y_0$ represents the observed (un-censored) values.

2.) Based on $Y_c^{(m+1)}$, $\theta^{(m+1)}$ is sampled from $Pr(\theta | Y_0, Y_c^{(m+1)})$, the complete
data posterior for $\theta$.

Thus, for each iteration of the chain, we are 'augmenting' the data with imputed
values for the censored observations.

We note that, depending on prior specification, a full conditional distribution
may not have a closed-form expression and thus direct sampling may be
difficult. In these cases, for these steps of the Gibbs framework, OpenBUGS uses
the Metropolis-Hastings or slice samplers as alternatives. The Metropolis-Hastings
sampling algorithm is the most general sampling scheme and, in principle, works
in every situation. The Metropolis-Hastings algorithm is as follows:

1.) Given a current value of the parameter $\theta^0$, select a candidate $\theta^*$ from a
    'jumping/transition' distribution $f(\theta^* | \theta^0)$, very often a normal distribution.
2.) Compute and 'acceptance ratio' given by:

   \[ r = \frac{Pr(\theta^* | data) / f(\theta^* | \theta^0)}{Pr(\theta^0 | data) / f(\theta^0 | \theta^*)} \]

3.) Accept $\theta^*$ as $\theta^1$ with probability $\min(r, 1)$. If $\theta^*$ is not accepted, then
    $\theta^1 = \theta^0$.
4.) Repeat as necessary.

As we can see, unlike a pure Gibbs sampler, the Metropolis-Hastings
algorithm does not necessarily generate a new value at each iteration.
However, based on this above acceptance/rejection scheme, samples tend to
stay in high-density regions of the desired distribution and only occasionally visit
low-density regions. For this reason, the M-H algorithm returns samples that follow
the desired joint distribution. Slice sampling (Neal, 2003) is a general purpose
algorithm for single site updating that always produces a new value at each
iteration.
LIST OF REFERENCES


Part II: STATISTICAL DOWNSCALING OF PRECIPITATION EXTREMES
There is great societal interest in assessing the impacts of projected climate change, and more specifically, there is an intense interest in the impact of change in variability and extreme events that could accompany global climate change predictions (Tebaldi et al., 2006). Increases in these extremes have already been observed as precipitation events, heat waves, and drought are occurring with greater intensity and frequency over the past few decades (USCCSP 2008). Over the same time period, Karl et al. (1996) concluded that the climate in the United States is generally more extreme, based on aggregated set of climate change predictors. Among extreme events, precipitation extremes are a primary concern as these events are typically more impactful [sic] than precipitation events alone and are responsible for a disproportionately large part of climate-related damages (Kunkel et al. (1999), Easterling et al. (2000), Meehl et al. (2000)). Natural systems may also be affected by changes in precipitation extremes, as these events have been shown to cause shifts in ecosystem distributions, to trigger extinctions, and to alter species morphology and behavior (Parmesan et al., 2000). Furthermore, extreme rainfall often translates into extreme flooding and consequently great material and economic losses, erosion and damage to crops, collapse of lifeline infrastructure, the breakdown of public health services (Douglas and Fairbank, 2011), fatalities (Kunkel et al., 1999), and structural damage to dams, bridges, and coastal roads.
Precipitation in the United States has increased over the past century, and this increase is primarily reflected in heavy and extreme daily precipitation events (Karl and Knight, 1998). Further studies have provided additional evidence that precipitation extremes are becoming more and more extreme and will continue to do so in the future (e.g. Zwiers and Kharin (1998), Groisman et al. (1999), Meehl et al. (2000), Tank and Konnen (2003), Kharin and Zwiers (2005)). Tebaldi et al. (2006) aimed to survey the most recent projections of climate extremes provided by the latest state-of-the-art global circulation models (GCMs) for climate ‘indices’ defined by the Intergovernmental Panel on Climate Change’s Fourth Assessment Report (IPCC-AR4). They concluded that models agree with observations over the historical period that there is a trend towards a world characterized by intensified precipitation, with a greater frequency of heavy-precipitation and high-quantile events, although with substantial geographical variability.

Locally, in New England, precipitation extremes, extreme rain or snow events, seem to be more prevalent. In fact, from October 2005 through April 2007, coastal New England experienced three ‘historically’ extreme and statistically anomalous precipitation events. In a related study, Douglas and Fairbank (2011) investigated the presence of trends in extreme precipitation (annual maximums) in northern New England between 1954 and 2008. Studying daily maxima of rain depths for the region, they found a strong increase in the magnitude of extreme precipitation events over the last three decades, especially in eastern Massachusetts and southern New Hampshire. Based on calculated 100-year return levels, they found that Technical Paper no. 40 (TP-40)
underrepresents coastal storm depths in northern New England and strongly suggested that storm estimation methods be updated. Furthermore, their research suggests a strong increase in the magnitude of extreme precipitation events over the last three decades, especially in the eastern half of MA and the southern half of NH where it appears that precipitation is becoming more extreme. And, similar to the findings of Tebaldi et al. (2006), they found significant geographic variability among precipitation extremes. The recent record-breaking storm events in New England anecdotally support these conclusions.

Because of their serious implications on the environment, this analysis focuses on assessing future, extreme precipitation events by applying a method of statistical (probabilistic) downscaling to large-scale model output. This downscaling process will be applied to several locations throughout all six New England states. This analysis outlines both the downscaling procedure and the parametric, extreme value analysis required to generate an accurate, high-resolution distribution of precipitation extremes at the local-scale. From these distributions, extreme impacts will be assessed through the calculation of return level estimates.
Atmosphere-ocean general circulation models, or GCMs, are combinations of atmosphere and ocean models that simulation weather at a global scale. Global circulation models are the primary tool used to quantify and assess climate change impacts and they underpin most climate change impacts studies (Wilby and Harris, 2006). However, because global weather simulation is so computationally expensive, these models provide predictions/projected scenarios at an extremely coarse scale (250KM by 250KM, in most cases). The issue is that environmental impact models are sensitive to local climate characteristics, and the drivers of local climate variation are not captured at the coarse scales of GCMs (Maurer and Hidalgo, 2008). That is, GCMs do not provide an accurate description of local climate. To overcome this discrepancy, methods of ‘downscaling’ are applied to produce local-scale climate predictions based on corresponding GCM scenarios. Downscaling itself is a new science as it relies heavily on GCM outputs which are products of recent advances in the climate science community (Benestad et al., 2007).

Downscaling appears in two forms: Dynamical and statistical downscaling (or empirical statistical downscaling). Dynamical downscaling is a computationally-intensive technique in which a fine-scale climate model is nested within a coarse scale model. More specifically, dynamical downscaling makes use of the lateral boundary conditions combined with regional-scale forcings such as land-sea contrast, vegetation cover, etc., to produce regional
climate models (RCMs) from a GCM. RCMs are normally gridded regions with scales in the tens of kilometers. Simulations from RCMs are not intended to mimic the observed weather conditions, but are intended to accurately portray weather characteristics over a defined time period. Some works have based climate predictions solely on dynamically downscaled (RCM) projections (for example, Dominguez et al., 2012), while others have questioned the value of using RCMs instead of GCMs themselves (for example, Castro et al., 2005).

Statistical downscaling (SD), on the other hand, is a more efficient (less computationally expensive) alternative that may be applied to achieve a variety of results. Essentially, statistical downscaling is a two-step process consisting of 1) the development of statistical relationships between local climate variables and large-scale predictors, and 2) the application of such relationships to the output of large-scale output to simulate local climate characteristics in the future (Hoar and Nychka, 2008). Statistical downscaling is a realistic approach to develop a specific, local-level climate prediction.

Typically, SD methods are applied to GCM projections, but may also be applied to RCM output as these results may not be representative for the local climate (Skaugen et al., 2002; Engen-Skaugen, 2006). Furthermore, RCM output may simply have inadequate spatial resolution for some impact studies, and some kind of statistical downscaling must be applied to the dynamical model results (Benestad et al., 2007).
CHAPTER II: METHODOLOGICAL MOTIVATION

This analysis focuses on a method of 'probabilistic downscaling' to project a single variable, extreme precipitation, into the future. While traditional ESD models the link between large- and local-scale variables, probabilistic downscaling is a branch of statistical downscaling that models the relationship between large- and local-scale statistical entities. In this case, the statistical entities are the corresponding cumulative distribution functions (CDFs) of the large- and local-scale precipitation extremes. In this way, probabilistic downscaling techniques do not retain the chronology, or exact ordering, of events. Techniques have been proposed to recreate time series from a distribution (Benestad, 2010), but accurate descriptions of future climate distributions are themselves sufficient predictions. This is especially true in our case as we do not aim to predict weather, but rather the distribution of a weather variable (precipitation extremes). A further discussion of single variable statistical downscaling techniques can be found in Benestad et al. (2007) and Hayhoe et al. (2004), for example.

When dealing exclusively with CDFs, the simplest form of downscaling is what is referred to as 'quantile mapping' or 'quantile matching'. This non-parametric technique downscales a large-scale value $x_G$ by selecting a local-scale value $x_s$ based on the following:

$$F_s(x_s) = F_G(x_G) \text{ with } x_s = F_s^{-1}(F_G(x_G))$$
where $F$ is a CDF of a climate random variable. Once a mapping has been defined, it is then applied to large-scale dataset to create a local-scale prediction. The method does not take into account the information of the distribution of the future modeled dataset (Michelangeli et al., 2009). Furthermore, the method of quantile mapping cannot provide local-scale quantiles outside the range of the historical observations (Michelangeli et al., 2009). Proposed by Wood et al (2004), the technique was applied to downscale monthly precipitation and temperature output from a GCM, and became known as bias-correction and spatial downscaling (BCSD). BCSD was originally developed to streamline the translation of GCM output and has proven reasonable for hydrological applications (Maurer and Hidalgo, 2008).

To overcome the clear shortcomings of the quantile matching methodology, Michelangeli et al. (2009) proposed an extension to this simple mapping for downscaling CDFs. Their technique, called the CDF-t, is similar to quantile mapping as it compares local- and large-scale distributions, but it accounts for changes in the large-scale CDF between historic and future periods. Let $X$ denote a variable from climate model output, such as precipitation, and let $X_c$ denote the series of the variable over the current, or calibration, period. Then, $X_P$ denotes the variable projected into the future, the time series from runs of the climate model in the future. Similarly, let $Y_c$ and $Y_P$ denote the current and future series for the local-level station. We note that while $Y_c$ is observed, $Y_P$ will need to be predicted or downscaled. Finally, a transformation, $T(\cdot)$, is assumed to exist between the large- and local-scale variable such that $T(\cdot) : [0,1] \to [0,1]T(\cdot)$. We then have the relationship:
\[ F_{yp}(x) = T \left( F_{xp}(x) \right) = F_{yc} \left( F_{xc}^{-1}(F_{yp}(x)) \right) \]

where \( F_{yp} \) and \( F_{xp} \) are the respective empirical CDFs for the local- and large-scale prediction, and \( F_{yc} \) and \( F_{xc} \) are the respective CDFs of observed (historic) local-level data and observed large-scale, or regional data. For further details see Michelangeli et al. (2009). The clear improvement over quantile mapping is that the future, local-scale distribution is a function of both historic observations and large-scale information that may be distributed differently between calibration and projection periods. The prediction takes advantage of all available data series.

Most climate and hydrological studies have been based on the downscaling of mean values (Fowler et al., 2007) or the center of the distribution of a climate variable. For these cases, the CDF-t method is perfectly applicable. However, for precipitation data, the center of the distribution associated with small and moderate precipitation amounts is uninteresting as we are concerned with the extreme events. In these cases, where the tails, which correspond to the extremes or high quantiles, are of primary interest, the non-parametric CDF-t is not suitable. Because most precipitation measurements are zero as it is a rare event, the corresponding empirical CDFs are heavy-tailed. With so few data at the extreme ends of the distribution, quantiles at the tails, as a result, have large variance. That is, a non-parametric CDF estimate requires a good amount of data to achieve a level of precision, and in the tails where data is sparse, the estimates are based on few observations and tend to be poor. Furthermore, these tails may also be strongly influenced by a single extreme event. Lastly,
observations of historical changes, as well as future projections, confirm that changes in the distributional tails of precipitation (extremes) may not occur in proportion to changes in the mean and may not be symmetric in nature (Kharin and Zwiers, 2005; Robeson, 2004; Tank and Konnen, 2003; Easterling et al., 2000). We can conclude, then, that the CDF-t will not allow for proper inference on extreme cases, as desired.

In light of these shortcomings, Kallache et al. (2011) proposed the XCDF-t technique to downscale the distribution of extremes exclusively. The technique is analogous to the CDF-t technique of Michelangeli et al. (2009) in that it makes use of the same transformation function form to link large- and local-scale distributions of climate variables. Unlike the CDF-t method, however, the XCDF-t links the distributions of large- and local-scale extremes only. To do this, elements of extreme value theory (EVT) are used to fit appropriate distributions to subsets of the data deemed 'extreme'. The advantage is that the framework of EVT (eg., Coles, 2001) allows for more precise estimation of the extreme portions of distributions. Also, while the CDF-t method was a non-parametric procedure, the XCDF-t method estimates parametric extreme value distributions based on limiting properties of max-stable data. Specifically, in Kallache et al. (2011), the generalized Pareto distribution (GPD) is used exclusively. Like the CDF-t, the XCDF-t uses the following transformation:

\[ F_Y(x) = T \left( F_X(x) \right) = F_Y \left( F_X^{-1}(F_X(x)) \right). \]
We note that now, $F_{X_p}$, $F_{Y_c}$, and $F_{X_c}$ are not empirical CDFs, but rather GPDs for the extremes of the large-scale predicted, local-level observed (historic), and large-scale observed series, respectively. For further details, see Kallache et al. (2011).
CHAPTER III: EXTREME VALUE THEORY

The XCDF-t downscaling scheme of Kallache et al. (2011) relies on the GPD to model exceedances at the regional- and local-scale. Individually, these subsets of data, the large-scale historic, large-scale future, and local-scale historic data, may be analyzed as univariate extreme value variables. An efficient way of doing this, especially with precipitation measurements, is by using a threshold model (Coles, 2001). Considering only interval maxima is thought to be 'wasteful' as many data below the maxima are discarded (Coles, 2001). By observing all threshold exceedances, we retain many more of these data. In threshold models, data above a set (but not necessarily constant) threshold are considered 'extreme.'

Consider a sequence of independent random variables, $X_1, X_2, ..., $ with common distribution function $F$, where $F(x) = \Pr\{X_i \leq x\}$. Let $M_n = \max(X_1, ..., X_n)$, the maximum of the process over $n$ time units of observation. Suppose that for large $n$,

$$\Pr\{M_n \leq z\} \approx G(z),$$

where

$$G(z) = \exp \left\{ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}$$
for some $\mu, \sigma > 0$, and $\xi$. As this threshold, $u$, becomes sufficiently large (approaches the upper endpoint of the distribution), the limiting distribution of exceedances are given by

$$F_u(z) \approx 1 - \left(1 + \frac{\xi(z-u)}{\sigma}\right)^{-\frac{1}{\xi}},$$

for $z > u$ which is known as the generalized Pareto distribution and where $\xi$ and $\sigma$ are shape and scale model parameters, respectively (Pickands, 1975; Balkema and de Haan, 1974). That is, if block maxima have approximating distribution $G$, then the threshold exceedances have a corresponding approximate distribution within the generalized Pareto family.

This result is significant for application as exceedances above a sufficiently high threshold may be approximately estimated by the generalized Pareto distribution, regardless of the distribution of the original data (Smith, 1989; Davison and Smith, 1990). Recently, the emphasis in EVT has shifted toward methods based on exceedances over thresholds rather than maxima (Smith, 2001), and the GPD has become a standard model for extremes. Furthermore, the GPD is 'generalized' and is a rich class of distributions that is very flexible and includes a variety of tail behaviors. This form, the qualitative behavior of the GPD, is dominated by $\xi$, the shape parameter. For $\xi > 0$, the distribution of excesses is unbounded and has the traditional 'Pareto', heavy tail; for $\xi < 0$, the distribution has a finite upper bound and resembles a Weibull-type distribution; and $\xi = 0$ corresponds to an unbounded, exponential-type distribution.
There are several issues concerning the fitting of a GPD to precipitation exceedances. First, the GPD only applies to excesses above a sufficiently large threshold, the selection of which can be unclear. When only extremes are analyzed, the number of observations, of course, tends to be quite few. But, since the theory assumes a high threshold, excessive lowering to increase observations would induce a strong bias as the GPD will fit the exceedances poorly. On the other hand, too high of a threshold will reduce the number of exceedances and thus yield erratic data, instability, and generally increase the estimation variance. To reach a compromise between bias and variance, Davison and Smith (1990) introduced the 'mean excess' or 'mean residual life' plot to identify appropriate levels of threshold. If the GPD is the correct model for the exceedances above a threshold \( u \), then mean excess, i.e., the mean value of \((Z - u)\) is given by:

\[
E(Z - u|Z > u) = \frac{\sigma + \xi u}{1 - \xi}
\]

which is linear in \( u \). Thus, when the mean excess is plotted against \( u \), the plot should appear linear if the GPD is an appropriate approximation. In theory, this plot then identifies suitable values for \( u \). But, precise estimation of \( u \) is often subjective. A second approach to threshold selection is to simply estimate the model at a range of thresholds. Above a certain level at which the asymptotic properties of the GPD is valid, estimates of the shape parameter should be approximately constant, while estimates of the scale parameter should be linear in \( u \). In practice, however, a combination of both procedures is usually recommended.
CHAPTER IV: DATA

Local-level (or point-level) precipitation data, \( Y_c \) in our downscaling notation, were obtained from the National Climatic Data Center (NCDC), a climate data archiving and retrieval system operated by the National Oceanographic and Atmospheric Administration (NOAA) Satellite and Information Service. Hourly precipitation accumulations reported in hundredths of inches were originally obtained for 69 meteorological stations from all six New England states covering a period from 1948 to 2010. Not all New England stations, however, had complete precipitation records for the entire period (1948-2010), and few had data for only short intervals therein. Extreme events are expected to be quite rare, so time series of observations collected over decades are typically required to accurately model these cases. Thus, the few stations with shorter precipitation series were omitted from that analysis and only stations with continuous measurements between 1970 and 2000 were retained. Additionally, the data from the Mt. Washington station in New Hampshire were omitted from the analysis as this location experiences weather conditions atypical of all other locations. That is, Mt. Washington is known to experience some of the world’s most extreme weather, and data from this station are truly outliers. In the end, of the original 69 stations, hourly precipitation data for 58 monitoring stations in New England were considered. For each of the remaining stations, hourly accumulations were aggregated into daily precipitation totals in
inches. Furthermore, for each of these stations, the geographic features of latitude, longitude, and elevation were utilized from the NCDC database.

Despite being physically consistent representations of smaller regions (typically, 50KM by 50KM) of the atmosphere (Benestad et al., 2007), regional climate model (RCM) scales are often still too coarse for local-level prediction. Thus, RCM output is used for large-scale precipitation data, both historic and predicted, or \( X_c \) and \( X_p \) from the downscaling transformation function, respectively. The four regional model outputs used in downscaling were acquired from the North American Climate Change Assessment Program (NARCCAP), an international program which acts as a custodian for regional climate simulations to be used in impact assessment and research. Each of the RCM outputs is driven by larger-scale atmospheric and oceanic boundary conditions provided by global circulation models, or GCMs. Thus, RCMs are highly dependent on their GCM ‘drivers.’ A total of eight different RCM/GCM combinations were used in conjunction with the NCDC station-level data. Information regarding the specific RCMs used is listed below (Table 2). We also note that all GCMs have been forced with the SRES A2 greenhouse gas emissions scenario for the 21st century (See, Nakicenovic et al. (2000) for Special Report on Emissions Scenarios (SRES) commissioned by the IPCC).
Table 2: RCM/GCM combinations used for our large-scale model output.

<table>
<thead>
<tr>
<th>Name</th>
<th>RCM Name</th>
<th>Modeling Group</th>
<th>GCM Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRCM-CGCM3</td>
<td>Canadian Regional Climate Model</td>
<td>OURANOS / UQAM</td>
<td>3rd Generation Coupled Global Climate Model</td>
</tr>
<tr>
<td>CRCM-CCSM</td>
<td>Canadian Regional Climate Model</td>
<td>OURANOS / UQAM</td>
<td>Community Climate System Model</td>
</tr>
<tr>
<td>HRM3-GFDL</td>
<td>Hadley Regional Model 3 / Providing Regional Climates for Impact Studies</td>
<td>Hadley Centre</td>
<td>Geophysical Fluid Dynamics Laboratory GCM</td>
</tr>
<tr>
<td>HRM3-HADCM3</td>
<td>Hadley Regional Model 3 / Providing Regional Climates for Impact Studies</td>
<td>Hadley Centre</td>
<td>Hadley Centre Coupled Model, version 3</td>
</tr>
<tr>
<td>RCM3-GFDL</td>
<td>Regional Climate Model, version 3</td>
<td>University of California, Santa Cruz</td>
<td>Geophysical Fluid Dynamics Laboratory GCM</td>
</tr>
<tr>
<td>RCM3-CGCM3</td>
<td>Regional Climate Model, version 3</td>
<td>University of California, Santa Cruz</td>
<td>3rd Generation Coupled Global Climate Model</td>
</tr>
<tr>
<td>WRFP-CGCM3</td>
<td>Weather Research and Forecasting Model</td>
<td>Pacific Northwest National Laboratory</td>
<td>3rd Generation Coupled Global Climate Model</td>
</tr>
<tr>
<td>WRFP-CCSM</td>
<td>Weather Research and Forecasting Model</td>
<td>Pacific Northwest National Laboratory</td>
<td>Community Climate System Model</td>
</tr>
</tbody>
</table>

For each of the eight cases, historic RCM data, or RCM "current" data, are produced by nesting the RCM within corresponding GCM. Current data of 42 climate variables are produced for the period 1970-1999. Large-scale precipitation output are available as three-hourly average instantaneous flux in units of kg/m²s which was then converted to accumulation in inches. That is, precipitation data are produced as a total amount (accumulation) for each
three-hour segment spanning the entire historic period. Similarly, the RCM
projected model output are produced for the period 2040-2069 and also appear
as three-hourly precipitation accumulations in inches. In each case, the
datasets are precise and of high-quality with no missing entries for the entire
measurement period. For both model output, historic and future (projected),
and the station data, respective three-hourly and hourly accumulations were
aggregated into daily totals of precipitation measured in inches. For our
purposes, this resolution is sufficient for interpretable predictive results.

RCM output is not meant to predict weather, but rather to accurately
represent large-scale climate features across a given region. In this way, RCM
output represents average weather over a region and/or time scale. This
downscaling method does not rely on an hour-to-hour correspondence
between station and large-scale precipitation. Instead, like most downscaling
techniques, only the translation or relationship between the statistical distribution
of observations and model output is of direct interest.
CHAPTER V: PROCEDURE

Much of this analysis was motivated by CDF-t procedure of Michelangeli et al. (2009) and the extension to the extreme case, the XCDF-t procedure prosed by Kallache et al. (2011). Following these approaches, we intend to achieve local-level forecasts of extremes based on the maximum amount of available data. That is, we aim to train a model on the historic relationship between large-scale and local-scale precipitation extremes, and then translate this relationship onto future, large-scale output to achieve a local-level forecast. In this way we are projecting climate ‘on the ground’ and into the future. Taking advantage of the model output in developing local-level predictions makes this a downscaling procedure instead of simply a predictive model.

Assumptions

First, like all statistical downscaling techniques, the XCDF-t method assumes that the relationship between large- and local-scale extreme remains constant between calibration and prediction period. That is, the models are based on historical data, and there is no guarantee that the past statistical relationships between different data fields will hold in the future. This so-called ‘stationarity’ assumption is made with any ESD procedure and is virtually impossible to verify. This may be considered an inherent fault with ESD that cannot be overcome. However, if long enough current time series are available, this assumption may be tested by partitioning the series into calibration and validation sets.
Next, it is generally agreed upon that New England does not have a particularly wet season as precipitation throughout the year is fairly constant. From our own data investigation, run charts and plots of precipitation totals over the entire period shows little or no evidence of true seasonality, so no seasonal component was included in our analysis. If true seasonality or trend was detected, however, accounting for such non-stationarity is, theoretically, a relatively simple task. We may introduce time as a covariate by allowing the threshold or parameters of the GPD to 'evolve' temporally. The use of such non-stationary EV distributions for climate change studies is becoming quite a popular approach (Smith, 2003). For example, Maraun et al. (2010) use a time covariate to integrate seasonality into their models of extreme daily precipitation in the UK.

Lastly, the extreme value distributions assume that data are independent of one another. Like precipitation events themselves, extreme events tend to exhibit temporal dependence and occur in clusters, especially at high temporal resolutions. In our case, for both station and model output data, respective hourly and three-hourly accumulations are first aggregated into daily precipitation totals. Then, after thresholds are established, a series of exceedances are extracted from the data that are assumed to be approximately independent of one another, and thus suitably modeled by an extreme value distribution.
Data grouping

We aim to apply our downscaling procedure to each of the 58 New England locations, which means that each of these stations must be grouped/matched with a corresponding region from the eight RCMs to establish the required $X_C, X_P,$ and $Y_C$ series. Although all RCMs are comprised of equally sized 50 kilometer by 50 kilometer grids, these layouts are not identical. In the majority of cases, stations are matched to model output based simply on gridded boundaries. For some coastal areas, however, stations are matched to output based on latitude, longitude, and elevation information. This ensures that these locations are not matched with predominately ocean-related grids. This process, in theory, results in each of the precipitation data from the 58 stations being matched with eight corresponding historic and future RCM output that most closely represents 'on the ground' characteristics.

Generalized Pareto distribution

Following the method of Kallache et al. (2011), a GPD is first fitted to the precipitation exceedances associated with the three available series, $X_C, X_P,$ and $Y_C$. The distribution of future precipitation exceedances at the local level, $F_{Y_P}$, is a combination of these three GPDs and defined through the downscaling transformation function. This resulting distribution, however, is not necessarily a GPD. For ease of calculation and interpretation, one may assume a simplified form of the distribution of local, predicted exceedances. If desired, one may first assume $\xi_{X_P} = \xi_{X_C}$, that the model output shape parameters are equal and that the resulting $F_{Y_P}$ is of GPD form. With these assumptions, the parameters of the
resulting \( F_{yp} \) can be expressed in terms of the component GPD parameters, with \( \xi_{yp} = \xi_{yc} \) and \( \sigma_{yp} = \sigma_{yc} \cdot (\sigma_{xp}/\sigma_{xc}) \), and the resulting parametric distribution is very simply calculated. The first assumption, the assumption of equal model shape parameters, is typically justified as changes in this parameter for the same variable are expected to be small (Kallache et al., 2011), and if small deviations should occur, they may be compensated for by associated scale parameter estimates.

Initially, the full downscaling transformation function of the XCDF-t was applied to each series grouping **without** assuming simplified version of the technique outlined above. That is, the transformation function was used to estimate the predicted distributions of station-level extremes and their corresponding return levels without assuming the resulting distribution is of GPD form. For this procedure, we do, however, constrain the analysis to series whose distributions, \( F_{xc}, F_{xp}, \) and \( R_{yc} \), have positive or identically zero shape parameters corresponding to heavy- and exponentially-tails, respectively. The vast majority of GPD model-fitting results yielded positive shape parameters, but those that resulted in negative estimates were subsequently set to zero. This assumption is supported by the widely held belief that precipitation data, particularly maxima and extremes, consistently appear as heavy-tailed (e.g. Katz et al., 2002; Smith, 2001). Furthermore, it is known that precipitation extremes typically follow a heavy-tailed or exponential-tailed distribution, so it is commonplace to assume a constant zero or slightly positive shape parameter for multiple series.
Aside from the obvious advantages, the simplified procedure avoids 'getting stuck' in the quantile/probability matching procedure. That is, $F_{Y_p}$ is defined only on values between the minimum and maximum of the future simulation dataset. For values outside this range, the result is an incomplete distribution taking on minimum and maximum probabilities different from 0 and 1, respectively. See Michelangeli et al. (2009) for a thorough discussion related to this topic and proposed remedial measures. Fortunately, in our case, our data was 'well-behaved' and $X_p$ input series typically yielded complete probability distributions. Later, when a Bayesian procedure is implemented, a non-parametric approach is developed to address such rare instances.

The simplified XCDF-t makes a fairly strong assumption that the $Y_p$ exceedances follow a GPD form. To test this claim, the 'full' XCDF-t was performed (without making the GPD assumption) and the resulting distribution was compared to that achieved from the simplified version. This comparison was performed at all New England stations and for every model combination, and results were extremely positive. That is, the XCDF-t without the GPD assumption yields distributions of exceedances that matched GPD forms almost identically. To further quantify these results, Kolmogorov-Smirnov (K-S) tests of equal distributions were performed for each station/model combination, and corresponding K-S test p-values were plotted (See Figure 27). From the figure, red markers represent small p-values (< 0.10), evidence to reject the claim of equal distributions, while green and blue markers represent little and no evidence to reject such a claim, respectively. Clearly, the vast majority of markers are blue, and the GPD model assumption is sufficiently justified.
Therefore, going forward the simplified procedure was used for return level estimation and the validation of local-level, predicted distributions to be discussed later.

Figure 27: K-S test for assumption of GPD model form for each New England location (Red dots indicate significant differences).

The XCDF-t procedure requires fitting a GPD and thus selecting a threshold for the three series, \(X_C, X_P,\) and \(Y_C\). As discussed, the issue becomes balancing bias with variance. Producing and analyzing MRL plots and parameter stability plots the three series at each station location would be unmanageable, so the method of Karl et al. (1996) was employed where a fixed, high percentile was chosen as a threshold. The 98.5\textsuperscript{th} percentile was chosen as a threshold for all series, at all stations, which equates to 5 observations per year, on average, or
about 200 observations for the entire collection period. We consider this a fair compromise between a high threshold level and allowing ample data for analysis. Lastly, related to threshold selection, a threshold for the predicted series, $Y_p$, must also be defined, otherwise exceedances are meaningless. In the spirit of statistical downscaling, we define the difference between historic and predicted local-level series to be proportional to the difference between model series, or $u_{Y_p} - u_{Y_c} = \lambda(u_{X_p} - u_{X_c})$, with $\lambda = \frac{\sigma_{Y_c}}{\sigma_{X_c}}$. In this expression, $\lambda$ is called the 'inflation factor' and ensures a complete distribution is downscaled. Furthermore, we consider $\lambda$ defined in this way to be advantageous as this relation is robust to large exceedances in the series.

**Return levels**

Kallache et al. (2011) present their results as summarized parameters of the downscaled distribution (shape, scale). These results, however, are more suited to statisticians and are less meaningful to scientists or those working in climate change assessment. Rather, we extended the XCDF-t method to include downscaled return levels. A return level is a quantile (typically high) of an extreme value distribution that will be exceeded with some known probability. This approach has improved tangibility as return levels are universally understood and relate directly to location-specific climate impact assessments such as flooding, potential erosion, etc. As stated in Coles (2001), it is usually more convenient to interpret extreme value models in terms of quantiles or return levels, rather than individual parameter values.
Suppose that a GPD with parameters $\sigma$ and $\xi$ is a suitable model for exceedances of a threshold $u$ by a variable $X$. Then, for $x > u$, we have:

$$\Pr(X > x | X > u) = \left[1 + \xi \left(\frac{x - u}{\sigma}\right)\right]^{-1/\xi}.$$ 

It follows that

$$\Pr(X > x) = \varsigma_u \left[1 + \xi \left(\frac{x - u}{\sigma}\right)\right]^{-1/\xi},$$

where $\varsigma_u = \Pr(X > u)$. Return level estimation requires the above expression, referred to as the 'unconditional' distribution, as it is no longer conditional on $X > u$. That is, since the GPD is defined for only those values above a defined threshold, the inclusion of the $\varsigma_u$ term allows us to estimate the proportion of exceedances from all observations and thus a proper return level.

So, the level $x_m$ that is exceeded, on average, once every $m$ observations is the solution of

$$\varsigma_u \left[1 + \xi \left(\frac{x_m - u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}.$$

Finally, we may rearrange the expression to find:

$$x_m = u + \frac{\sigma}{\xi} \left((m\varsigma_u)^{1/\xi} - 1\right).$$

This quantile is the $m$-observation return level, or the quantity that is expected to be exceeded once every $m$ observations. In practice, it is often more convenient and interpretable to give return levels in terms of years on an annual scale, or an $N$-year return level expected to be exceeded once every $N$ years, in
which case \( m = N \cdot n_y \) where \( n_y \) is the number of data points per year. Return level estimation requires maximum likelihood estimates of both \( \sigma \) and \( \xi \), and also an estimate of \( \zeta_u \), the probability of an observation exceeding the threshold \( u \). This value is approximated by \( \zeta_u = \frac{n_u}{n} \), where \( n_u \) is the number of exceedances and \( n \) is the number of observations. These estimates are easily attained from the data.

To produce return level estimates with associated measures of uncertainty, a parametric bootstrapping procedure was devised and implemented at each of the 58 locations (local-level stations) and for all 8 RCM/GCM combinations. As per Rice (1995), let \( \theta \) be the parameter of interest, in our case the return level of the predicted distribution of precipitation extremes, let \( \hat{\theta} \) be an estimator of that parameter, and let \( \theta_0 \) be the true, unknown value of the parameter. If we knew the distribution of \( \hat{\theta} - \theta_0 \), then we would have

\[
\Pr(\hat{\theta} - \theta_0 \leq \delta) = \frac{\alpha}{2}
\]

\[
\Pr(\hat{\theta} - \theta_0 \leq \bar{\delta}) = 1 - \frac{\alpha}{2}
\]

where \( \delta \) and \( \bar{\delta} \) are the \( \frac{\alpha}{2} \) and \( 1 - \frac{\alpha}{2} \) quantiles of the distribution, respectively. After manipulating, we have the expression

\[
\Pr(\hat{\theta} - \bar{\delta} \leq \theta_0 \leq \hat{\theta} - \delta) = 1 - \alpha
\]

and the corresponding 100(1 - \( \alpha \))% confidence level as \((\hat{\theta} - \bar{\delta}, \hat{\theta} - \delta)\). However, since \( \theta_0 \) is not known, \( \hat{\theta} \) is used in its place. That is, from a distribution with value \( \hat{\theta} \), we generate many samples and construct an estimate of \( \theta \) from each. So, for
our case, $\hat{\theta}$ (predicted return level estimate) is first calculated from GPD model parameters fitted to the $X_c$, $X_P$, and $Y_c$ series. Next, series of random exceedances are generated from these fitted parameters, GPD parameters are refit to these exceedances, and local-scale, predicted parameters are 'built' as specified by the simplified XCDF-t procedure. From these predicted parameters, the 25-year return level is estimated. This process is repeated many times and the collection of estimates, the bootstrap estimates, are denoted as $\theta_j^*, j = 1, 2, ..., B$, where $B$ is the number of samples (500 in our case). The distribution of $\hat{\theta} - \theta_0$ is suitably approximated by $\theta^* - \hat{\theta}$ (Davison and Hinkley, 1997), the bootstrap estimates minus the original estimate. Finally, from this distribution, the 0.05 and 0.95 quantiles are used to form the 90% 'mean corrected' (or 'basic') bootstrap confidence interval for each station/model combination.

Longer return periods, 50- and 100-year periods, were investigated as they are routinely used by engineers and infrastructure is often designed to have a 50-year lifespan. However, results based on these periods were highly variable as expected. That is, for our procedure, extrapolation beyond the simulation run length yields unreliable results. The 25-year return period, however, seems a much more realistic choice and is long enough to be useful and meaningful for climate impact assessment. In other words, these estimates represent a fair compromise between accuracy and interpretability. Lastly, we note that 90% confidence intervals were chosen simply as conservative choices.
Point Estimates

Figure 28 below presents the 25-year return levels estimates associated with predicted distributions for all 58 New England locations. From this figure, we observe how our choice of RCM/GCM combination affects our projections and resulting return level estimates. For example, for six of the eight RCM/GCM combinations, Pinkham Notch, NH (circled below) had the highest predicted 25-year return level estimate of all locations. Upon closer inspection, however, the magnitude of such estimates is variable and highly dependent on model combination choice. That is, the RCM/GCM combinations HRM3_hadc and RCM3_gfdl yield 25-year return levels of 6.82 and 10.85 inches per day, respectively. Thus, for this location, daily 25-year return level estimates can differ by more than 4 inches, depending on the model selected. Similar variability exists at other locations.
Next, we observe differences between the 25-year return levels associated with model predictions, $X_p$, and those associated with historic model output, $X_c$. This represents the projected effect of climate change on extreme precipitation as provided by the RCM model output. In Figure 29 below, red and blue markers indicate positive and negative changes, respectively, and the size of the marker indicates the magnitude of change. Results reveal the inconsistency and disagreement that exists due to both RCM and GCM driver choice. To illustrate the effect of the GCM driver on differences between historic and future periods, we observe cases where the RCM is held constant. Combinations CRCM_ccsm and CRCM_cgcm both exhibit moderate differences across locations with some subtle disagreement in terms of magnitude and sign of changes. On the other hand, both RCM3_gfdl and
RCM3_hadc show large positive differences in coastal Massachusetts, Rhode Island, and Maine, but generally disagree at other locations. In fact, RCM3_gfdl shows large negative differences throughout northern New Hampshire, Vermont, western Massachusetts, and Connecticut, while RCM3_hadc shows no such negative differences at these locations, and occasionally shows significant positive changes. Thus, differences in model output return levels are inconsistent, and results exhibit varying degrees of dependence on GCM driver.

To examine the effect of the regional model on return level differences, we observe cases where the GCM driver is held constant. Results from RCM3_cgcm, CRCM_cgcm, and WRFG_cgcm combinations are considerably different and often contrary. RCM3_cgcm shows large negative changes throughout Vermont, New Hampshire, eastern Massachusetts and Rhode Island, while WRFG_cgcm shows large positive changes at the same locations, generally. Results from CRCM_cgcm are near those of the WRFG_cgcm in terms of sign difference, but disagree significantly in terms of order of magnitude.

Next, we observe HRM3_gfdl and RCM3_gfdl combinations. HRM3_gfdl shows a coastal/inland effect with negative differences throughout VT, northern and western NH, western MA, and western CT, shows considerable positive differences throughout ME, central and eastern NH, eastern MA, and RI. Results from RCM3_gfdl show some agreement in RI, eastern MA, and western and northern NH in terms of sign change only. RCM3_gfdl, however, shows no such coastal effect and locations throughout ME and central/eastern NH show large negative changes. Thus, model output differences seem to be highly dependent on choice of RCM as varying this factor always yields different results.
We next compare the 25-year return levels associated with the local-level (station) predictions, $Y_P$, to those associated with the observations, $Y_C$. This represents the projected effect of climate change on extreme precipitation at the local station level via downscaling. Figure 30 below presents these differences, where red and blue markers indicate positive and negative changes, respectively, and the size of the marker indicates the magnitude of such change. We first observe that, unlike the model output (Figure 29), results seem to be more sensitive to the choice of GCM driver. That is, among common RCMs, different GCM drivers yield considerably different results. Second, like the model output, differences in station-level return levels are highly dependent on
our choice of RCM, and this choice yields considerably different and occasionally contradictory results. Finally, unlike the differences in return levels for the model output (above), the station results generally exhibit fewer decreases and more moderate increases. Such a comparison to the previous figure illustrates the effect that the downscaling procedure has on our estimates, how the procedure adds uncertainty to our estimates and fewer significant differences are observed. Subsequent figures will further quantify this effect.

Figure 30: Difference in 25-year return level estimates for $Y_P$ and $Y_C$.

**Climate Change Uncertainty Quantification**

Figure 31 below presents the projected 25-year return level estimates and associated 90% bootstrap confidence intervals at the local-level (New
Hampshire locations only). This figure illustrates the differences in return level estimates derived from RCM/GCM combinations (as above), but also includes the variability corresponding to each estimate. First, we observe that many return level estimates are consistent and comparably variable across RCM/GCM combinations. Such locations include Pittsburgh, Hanover, Errol, North Strafford, and Concord. On the other hand, Pinkham Notch, for example, exhibits both drastic differences in mean estimates and inconsistent variability based on RCM/GCM choice. Lincoln and other locations across New England exhibit similar inconsistent variability, to different degrees. In addition to influencing downscaled projections and return level estimates, characteristics of RCM/GCM combinations may be propagated through the downscaling process to affect variability in return levels as well. For bootstrapping, sample exceedances are generated from GPDs fitted to model output and observations. When these data are very extreme, fitted models and corresponding bootstrap samples will reflect this, and resulting return levels will be erratic. This is likely the case for locations such as Pinkham Notch with inconsistency across model combinations.
Figure 31: 90% bootstrap confidence interval for downscaled \((Y_P)\) 25-year return level estimates for New Hampshire locations only.

Figures 32 and 33 below introduce the variability attributable to downscaling and compares the differences in lower bounds of the 90% confidence intervals for 25-year return levels. Top figure compares differences in return levels for \(X_P\) and \(X_C\) series, while bottom figure compares differences between \(Y_P\) and \(Y_C\). Locations marked in red identify positive differences between return level lower bounds, which correspond to significant increases in 25-year return levels from historic to future periods. Like those results based on differences in return level estimates, we find differences in model output return level lower bounds to be highly dependent on our choice of RCM. However,
unlike previous differences in estimates, results seem to be less sensitive to GCM choice. That is, results based on common RCMs with different GCM drivers seem to change little in terms of both sign and order of magnitude.

Referring to the local-level differences (Figure 33), GCM seems to have more influence on resulting differences. For example, WRFG_ccsm and WRFG_cgcm show significant differences and contrary results at numerous locations. In this case, evidence suggests that our choice of GCM does contribute to the variability observed in downscaling results. Next, like the model output, station differences in return level lower bounds show significant dependence on our choice of RCM. This is illustrated in the significant disagreement between results obtained from RCM3_cgcm, CRCM_cgcm, and WRFG_cgcm combinations, for example. We conclude that there is little agreement between the 8 RCM/GCM combinations in terms of return level projections, and both the regional model and GCM driver choice contribute to such variability.
Figure 32: Difference in 90% confidence interval lower bound for $X_p$ and $X_c$ 25-year return level estimates.

RT25 CI 90% lower : future - present
Overall, observing differences in return level lower bounds, we find significant increases at relatively few stations. This is a marked difference from model output results that exhibit numerous significant increases in return levels from historic to future periods. Furthermore, the station results based on return level lower bounds are significantly different from station results based on return level estimates only, where numerous increases were observed. These comparisons illustrate the variability attributable to the procedure, how downscaling adds uncertainty to our results. We observe that of the few significant increases, most are found in southern New England, throughout Massachusetts and Connecticut, and associated with HRM3_hadc and
WRFG_cgcm. Not coincidentally, these same combinations yield the largest
difference in model output at these locations.

Nonparametric Bootstrap and Bias Corrected and Accelerated
Method

The simplest approach to calculating bootstrap confidence intervals is the
percentile method, or simply taking percentiles of the bootstrap samples. This
method is based on the assumption that sampling distributions are symmetric, so
the approach is not appropriate in cases where significant skew or asymmetry is
present, such as distributions of 25-year return levels which are known to be
skewed/asymmetric. In such cases, where very large values are occasionally
produced from the procedure, simple percentile confidence intervals are not
expected to have the correct coverage (Davison and Hinkley, 1997). Our
approach (described previously) addresses the skew of return level distributions
by calculating mean corrected confidence intervals, calculations believed to
be more accurate in the presence of asymmetry. However, the method, like the
percentile method, does not allow for the possibility that the bootstrapped
sampling distribution may be a biased estimate of the true sampling distribution.

As a remedial measure for such cases (as well as other cases), the BCa
(bias-corrected and accelerated) procedure was proposed by Efron and
Tibshirani (1993) as a method of achieving 'better confidence intervals.' In this
procedure, BCa interval endpoints are also given by percentiles of the bootstrap
distribution, but adjusted to account for the skew and bias of the data. The
actual percentiles used depend on two numbers referred to as \( a \) and \( z_0 \), the
acceleration and bias correction, which, generally speaking, measure the rate of change of the standard error and median bias of our estimator, respectively. In practice, these values are estimated by repeatedly sampling the data, and resulting BCa intervals are simply calculated. In addition to correctly considering the shape of the bootstrap distribution, the BCa has the advantage of being 'second-order accurate,' which, as stated by Efron and Tibshirani (1993), leads to much better approximations of exact endpoints.

As before, let \( \theta \) be our parameter of interest, in this case the return level of the predicted distribution of precipitation extremes, let \( \hat{\theta} \) be an estimator of that parameter, and let \( \hat{\theta}^* \) be the estimate of \( \theta \) based on the resampled (bootstrapped) data. Also, let \( \hat{G} \) be the cumulative distribution function of \( B \) (again, 500, in this case) bootstrap replications. The central \( 1 - \alpha \) BCa interval is given by \( (\hat{\theta}_{BCa}\left[\frac{1}{2}\right], \hat{\theta}_{BCa}\left[1 - \frac{\alpha}{2}\right]) \) where

\[
\hat{\theta}_{BCa}[\alpha] = \hat{G}^{-1}\Phi\left( z_0 + \frac{z_0 + z^{(\alpha)}}{1 - \alpha(z_0 + z^{(\alpha)})} \right),
\]

\[
z^{(\alpha)} = \Phi^{-1}(\alpha),
\]

with \( z_0 \) the bias-correction constant and \( \alpha \) the acceleration. The BCa algorithm estimates \( z_0 \) by

\[
z_0 = \Phi^{-1}\left\{ \frac{\#{(\hat{\theta}^*<\hat{\theta})}}{B} \right\},
\]

\( \Phi^{-1} \) of the proportion of bootstrap replications less than \( \hat{\theta} \). Furthermore, the acceleration, \( \alpha \), is accurately approximated by
\[ \hat{a} = \text{SKEW}_{\hat{\theta}} \left( \ell_\theta (\hat{\theta}) \right) / 6, \]

or one-sixth the skewness of the score function evaluated at \( \theta = \hat{\theta} \) (Efron, 1987).

For further details regarding the derivation of the BCa interval, the reader is directed to Efron (1987).

The acceleration and bias correction, and thus the adjusted BCa interval, are estimated non-parametrically, unlike the previous parametric bootstrap. First, the \( X_c, X_p, \) and \( Y_c \) series are collectively sampled with replacement. Next, for each of the three resulting series, GPD parameters are fit to the threshold exceedances and the simplified version of the XCDF-t procedure is applied. That is, the series are downscaled and the distribution of projected, local-level exceedances is 'built' from the model and local-level historic fitted GPD parameters. Finally, from the downscaled distribution, a 25-year return level estimate is calculated, and the entire process is repeated 500 times to produce a distribution of return levels from which the adjusted BCa interval can be estimated. Using these results, we now calculate differences between return levels and compare our results to those achieved previously.

As discussed, the BCa intervals are expected to capture the positive (right-) skew of the distribution of return levels and produced confidence intervals (90%) that are generally wider and shifted up as compared to previously calculated intervals. Below we have a comparison of BCa return level intervals for the \( Y_p \) series with corresponding mean corrected (basic) intervals for Connecticut stations only. From these results, compared to the mean corrected intervals, we immediately notice that virtually all BCa intervals are shifted up. This
suggests that distributions for return levels are somewhat right-skewed and the BCa adjustment captures these larger values. For station/model combinations that result in small return level intervals, the shift introduced by the BCa adjustment is typically very slight, while, for station/model combinations that result in larger estimates, the shift is very sizeable.

Figure 34: Comparison of BCa and mean corrected 25-year return level estimates.

Figure 35 below introduces the variability component by comparing the differences in BCa lower bounds of the 90% confidence intervals for $Y_p$ and $Y_c$ 25-year return levels. Locations marked in red identify positive differences between return level lower bounds, which correspond to significant increases in 25-year return levels from historic to future periods. Generally speaking, we find that differences in lower bounds based on BCa intervals to be similar to those obtained without the adjustment. That is, regardless of method (BCa or mean correction), increases and decreases follow the same broad pattern across stations and model combinations. We do note, however, that the BCa adjustment tends to result in more moderate (less extreme) negative values...
(fewer large negatives) and inflates the positive differences. This suggests a larger shift in the $Y_c$ interval than the $Y_p$ series which results in more positive differences than those observed without applying the BCa adjustment. This is somewhat expected, is consistent with our understanding of the BCa, and makes some intuitive sense as the BCa is expected to capture the extreme behavior of the projected series. That said, it is satisfying to observe marked similarities between results achieved with and without the adjustment. The BCa adjustment clearly makes some difference in our confidence interval estimation, but the resemblance between our current results to those obtained previously suggests that the mean correction is sufficient.

Figure 35: Difference in 90% confidence interval lower bound for $Y_p$ and $Y_c$ BCa 25-year return level estimates.

RT25 CI 90% lower : future - present

CRCM_ccsm

HRM3_gfdl

RCM3_cgcm

WRFG_ccsm

CRCM_cgcm

HRM3_hadc

RCM3_gfdl

WRFG_cgcm
CHAPTER VII: VALIDATION

Based on associated return level estimates, downscaling has shown to be sensitive to RCM/GCM combination. While this is a useful result unto itself, such differences tell us nothing of the accuracy (probabilistic predictive capabilities) of downscaling in an absolute sense. To address this, to evaluate the downscaling procedure, and to identify the most useful region climate models, we rely on validation. The procedure is a cross validation (CV) procedure where data is partitioned into non-overlapping training and validation sets. In our case, the first 75% of $X_C, X_P,$ and $Y_C$ were used as training sets and the XCDF-t was applied to these distributions to create a distribution of projected exceedances at the local-level. The last 25% of $Y_C$ was used as a validation set, and the downscaled (predicted) distribution was compared to this 'future' distribution of extremes. By training on earlier periods and validating on later periods, our framework mimics prediction into the future and may identify previously unidentified trends in extremes. Visual assessment of the procedure, comparing downscaled CDFs to validation CDFs, indicates overall success (See Figure 36).
Figure 36: Comparison (for all 8 RCM/GCM combinations and 8 New Hampshire locations) of downscaled (predicted) distribution and distribution of exceedances of validation period. Black lines represent distribution of validation set exceedances and red lines represent predicted distributions.

Next, Kolmogorov-Smirnov tests, tests of equality of distributions against one- and two-sided alternatives, were performed. K-S tests measure physical differences between distributions and, in our case, can be used to provide objective comparisons between training-based downscaling and validation data. After performing two-sided and one-sided tests for all 58 NE stations and 8 RCM/GCM pairings, K-S test p-values were produced and examined. Generally speaking, test results are unremarkable and downscaling appears to be successful at most locations/for most model combinations. However, the one-sided, 'less than' K-S test, the test with the alternative hypothesis that the downscaled GPD is stochastically greater than the validation exceedances (the
< alternative), reveals some downscaling failures. Below we have the associated p-value plot (See Figure 37) for this test where significant p-values, those <0.05 and indicated in red, provide evidence to reject the null hypothesis and indicate unsuccessful downscaling, or over-prediction of extremes. Locations marked in blue or green correspond to p-values >0.05 or between 0.05 and 0.10, and indicate successful and moderately successful downscaling, respectively.

From Figure 37 we observe strong agreement among model combinations as evident by the pattern of significant and weakly significant p-values (red and green dots, respectively) throughout northern, central, and southern NH, CT, and some locations in VT. In these cases, the null hypothesis of equal distribution of extremes is rejected in favor of the alternative that the downscaled distribution is significantly greater than the GPD of validation set. (Here, 'greater' refers to 'stochastically greater' or more extreme). For these cases, results suggest that downscaling over-estimates the distribution of extremes, or that observed precipitation will be less extreme than future predictions. We note, however, that these conclusions are made tentatively as it is not clear that K-S tests are optimal tests for validating our procedure. That is, because K-S test statistics are calculated from vertical differences between distributions, results from these tests may be misleading.
Figure 37: p-values for one-sided Kolmogorov-Smirnov tests by location. Red dots provide evidence to reject the claim that predicted distribution fit to training data is equal to distribution of validation data.

KS-test, Training > Validation

CRCM_ccsm  HRM3_gfdl  RCM3_cgcm  WRFG_ccsm

CRCM_cgcm  HRM3_hadc  RCM3_gfdl  WRFG_cgcm
CHAPTER VIII: BAYESIAN APPROACH

Next, we aim to develop a Bayesian model approach to supplement the downscaling of RCM precipitation data and station observations. As before, precipitation exceedances are assumed to follow a GPD which is asymptotically justified by extreme value theory. The procedure will thus result in distributions for these extreme value distribution parameters that will serve as a basis for downscaling and ultimately return level estimation.

As described earlier, our downscaling procedure was used to generate realizations from the predicted, station-level (local-level) extremes (exceedances), or the tail of the precipitation series. From this predicted tail, a bootstrap sample for a return level was generated from substituting the bootstrap sample parameter values into the appropriate return level expression. However, it is well known that bootstrap procedures are not consistent for extreme value problems as there is a tendency for these samples to generate shorter tails than the true sample distribution (Coles and Simiu, 2003). That is, when applying a bootstrapping for extremes, the tail of the bootstrap series is generally shorter than the tail of the original series. It is likely that some correction in the bootstrap procedure must be pursued, and a number of such remedies are available, such as the previously applied bias correction bootstrap. In our case, we have decided to supplement our previous results with a Bayesian approach. It is believed that this procedure will yield a more accurate predicted series and consequently more accurate and precise return level estimates.
The basic framework of a Bayesian analysis requires both a prior
distribution on the parameters of interest and a likelihood function of the data
given those parameters. Unlike traditional statistical methods, the parameters in
a Bayesian setting are considered random variables. Furthermore, the defined
prior distribution of these parameters contains information separate from the
actual data. The information provided by this prior distribution is alternately
considered the greatest strength and weakness of Bayesian inference (Coles,
2001). That is, proponents argue that prior distributions can supplement limited
data, while opponents of the methodology argue that prior specification can
subjectively alter results. While results can be sensitive to the prior distinction, it
has been reported that Bayesian inference, on extremes in particular, is robust
across a range of non-informative prior distributions (Coles and Tawn, 2005).

There are, however, indisputable and numerous other benefits to the approach.
Bayesian statistics is based, of course, on Bayes' Theorem which results in a
complete distribution of parameter given the data, or posterior distribution. Thus,
inference on the parameter can be summarized by characteristics of the
posterior distribution and asymptotic theory, as required by maximum likelihood
estimation, can be altogether avoided. Another advantage is that Bayesian
statistics gives a more realistic version of prediction. Predicted densities
necessarily contain terms for model uncertainty as well as uncertainty due to the
variability in future observations. In statistics, these measures of uncertainty are
considered as important as the estimates themselves. And, this becomes
especially true in the context of extreme value analysis where extreme quantiles,
for example, are known to have high levels of variability (Coles and Tawn, 2005).
Additionally, Bayesian techniques have a practical advantage over traditional approaches as they can effectively handle models with very large numbers of parameters and/or complex hierarchical structures. Bayesian techniques allow for inference on complex models that would otherwise be impossible. Finally, the Bayes approach is useful in predictive inference where the ultimate objective is not so much to learn the values of unknown parameters, but rather to establish a meaningful probability distribution for future unobserved random quantities.

**MCMC Sampling**

A major limitation of the Bayesian approach is that obtaining the posterior distribution often requires the integration of high-dimensional functions. Direct integration is often impossible and often computationally very difficult. Among other alternatives to direct integration, Markov Chain Monte Carlo (MCMC) methods attempt to simulate draws from some complex distribution of interest. In the MCMC approach, previous sample values are used to randomly generate subsequent sample values, thereby generating a Markov chain (as the transition probabilities between sampled values are only a function of the most recent sample value). MCMC methods can be traced to the Metropolis algorithm (Metropolis and Ulam 1949, Metropolis et al. 1953), a method to assist physicists compute complex integrals. By expressing these integrals as expectations of some distribution, they could be estimated by drawing samples from that distribution.

For our Bayesian models, OpenBUGS statistical software is used to generate samples from the joint distribution of the parameters. OpenBUGS operates under
a Gibbs sampling framework where conditional distributions of each parameter
given all the others (these are known as full conditional distributions) are
successively sampled. Gibbs sampling is a special case of the Metropolis-
Hastings (Hastings, 1970) algorithm and is considered to be widely applicable to
a broad class of Bayesian problems. The work of Gelfand and Smith (1990)
sparked a renewed interest in the Gibbs sampler and brought about a major
increase in its application. Furthermore, modern advances in computing have
allowed for computer-intensive MCMC algorithms to be easily implemented. A
simple overview of the approach is offered below, but more detailed treatment
of MCMC methods can be found in Gelman et al. (1995), for example.

Very generally, for a parameter vector, $\theta = (\theta_1, \theta_2)'$, the posterior density
($\pi(\theta|data)$) of the parameter vector is estimated using the following Gibbs
sampling framework.

1.) Start with initial values: $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})'$.
2.) Sample $\theta_1^t$ from $Pr(\theta_1|\theta_2^{t-1}, data)$.
3.) Sample $\theta_2^t$ from $Pr(\theta_2|\theta_1^t, data)$. This yields $\theta^t = (\theta_1^t, \theta_2^t)$.

More generally:

1.) Initialize the sampler with starting values $\theta^{(0)}$.
2.) Let the sampler run, generating $\theta^{(1)}, \theta^{(2)}, ...$
3.) Under a wide set of conditions, as $t \to \infty$, each Gibbs sample, $\theta^{(t)}$, can be
viewed as samples from the posterior density $\pi(\theta|data)$. 

138
As the sampler updates, it moves away from the initial values, providing a 'random tour' of the parameter space (possibly high-dimensional parameter space), visiting locations in the space with frequencies proportional to the posterior density. Formally, the output of the sampler forms an ergodic Markov chain on the parameter space for \( \theta \), with transition probabilities such that the 'limiting' distribution of the sampler is the posterior density \( \pi(\theta | \text{data}) \) of direct interest. Estimates using this approach get 'better' with increased samples (more samples), so modern computing power makes this approach to estimation and inference possible.

We note that, depending on prior specification, a full conditional distribution may not have a closed-form expression and thus direct sampling may be difficult. In these cases, for these steps of the Gibbs framework, OpenBUGS uses the Metropolis-Hastings or slice samplers as alternatives. The Metropolis-Hastings sampling algorithm is the most general sampling scheme and, in principle, works in every situation. The Metropolis-Hastings algorithm is as follows:

1.) Given a current value of the parameter \( \theta^0 \), select a candidate \( \theta^* \) from a 'jumping/transition' distribution \( f(\theta^*|\theta^0) \), very often a normal distribution.

2.) Compute and 'acceptance ratio' given by:

\[
    r = \frac{Pr(\theta^*|\text{data})f(\theta^0|\theta^*)}{Pr(\theta^0|\text{data})f(\theta^*|\theta^0)}
\]

3.) Accept \( \theta^* \) as \( \theta^1 \) with probability \( \min(r, 1) \). If \( \theta^* \) is not accepted, then \( \theta^1 = \theta^0 \).

4.) Repeat as necessary.
As we can see, unlike a pure Gibbs sampler, the Metropolis-Hastings algorithm does not necessarily generate a new value at each iteration. However, based on this above acceptance/rejection scheme, samples tend to stay in high-density regions of the desired distribution and only occasionally visit low-density regions. For this reason, the M-H algorithm returns samples that follow the desired joint distribution. Slice sampling (Neal, 2003) is a general purpose algorithm for single site updating that always produces a new value at each iteration.

**Procedure**

For our case specifically, R statistical software was used in conjunction with OpenBUGs through the ‘R2OpenBUGs’ package to produce posterior distributions for the GPD parameters, $\xi$ and $\sigma$ (xi and sigma, respectively). This, of course, was done for each of the three known series, $X_C, X_P, \text{and } Y_C$, to produce corresponding distributions for the respective parameters $\xi_{X_C}, \sigma_{X_C}, \xi_{X_P}, \sigma_{X_P}, \xi_{Y_C}, \text{and } \sigma_{Y_P}$. For each series, a beta prior distribution was assumed for shape parameters. That is,

$$\xi_{X_C} \sim Beta (\alpha_{\xi_{X_C}}, \beta_{\xi_{X_C}}),$$

$$\xi_{X_P} \sim Beta (\alpha_{\xi_{X_P}}, \beta_{\xi_{X_P}}),$$

and

$$\xi_{Y_C} \sim Beta (\alpha_{\xi_{Y_C}}, \beta_{\xi_{Y_C}}),$$

where the beta parameters, alpha and beta ($\alpha, \beta$), are set to be 1.5 and 2.5, respectively, for each series. These distributions are semi-informative at they
restrict estimates to values between 0 and 1 (thus, fairly non-informative), and similar to the approach of Martins and Stedinger (2000). As stated earlier, precipitation extremes typically follow a heavy-tailed or exponential-tailed distribution, so it is commonplace to assume a constant zero or slightly positive shape parameter and thus restrict our model to only such cases, or value deemed ‘sensible.’ Literature generally supports this claim as evidence suggests that the distributions of hydrologic variables are heavy tailed.

For scale parameters, prior distributions were assumed to be uniform. That is,

\[ \sigma_{x_c} \sim \text{Unif.}(a_{\sigma_{x_c}}, b_{\sigma_{x_c}}) \],

\[ \sigma_{x_p} \sim \text{Unif.}(a_{\sigma_{x_p}}, b_{\sigma_{x_p}}), \text{ and} \]

\[ \sigma_{y_c} \sim \text{Unif.}(a_{\sigma_{y_c}}, b_{\sigma_{y_c}}). \]

For these parameters, the uniform distribution is defined from 0 to 50 (a and b, respectively). Thus, the prior specification is non-informative and extreme diffuse as parameter estimates from traditional model-fitting were typically found to be between 1 and 3. Thus, for the scale parameters, we are truly allowing the data to guide the posterior analysis.

The Gibbs sampler was initialized with starting values, \( \xi^{(0)} = 0.48 \) and \( \sigma^{(0)} = 3 \) for each series, reasonable values based on previous model-fitting. Individual samplers were run for each series using 20,000 iterations with a burn-in period of 5,000 iterations to produce three distinct sets of shape and scale posterior
distributions. Convergence and independence from the starting values were checked by CODA (distributions, traces, etc.), the standard tools in such cases.

**Non-parametric extension/semi-parametric approach and results**

Up to this point, the approach, the parameter estimation by means of OpenBUGS, has been purely parametric. That is, the distribution of precipitation exceedances (of a threshold) has been assumed to be a generalized Pareto distribution (GPD). As mentioned in describing our previous parametric bootstrap, applying the transformation function for downscaling is not defined for values outside the range of predicted model exceedances. Furthermore, for this reason, the ultimate local-level predicted distribution may not be a 'complete' distribution. Kallache et al. (2011) address this issue by inflating and shifting the large-scale series so that $Y_c$ and $X_c$ have approximately the same range before applying XCDF-t. We, however, adopt a semi-parametric approach where the parametric analysis is supplemented with an empirical estimation of values below the defined threshold. This empirical extension is pragmatic solution that avoids the need for shifting or inflating the historic series. Our approach can best be explained by a step-by-step description.

First, a sample of simulated exceedances is generated based on both model and local-level observations. Next, as described previously, the GPD parameters are estimated for each series under the Bayesian framework. For each set of (many) estimates, a GPD or quantile function is calculated, depending on the series. Next, a non-parametric, empirical CDF is estimated for values between some low 'buffer' limit and our extreme value threshold (98.5th
percentile). For all series, this lower buffer is set at the 90th percentile as values below this level are not expected to affect the extreme value predictions. This non-parametric ‘extension’ is calculated for all three series, $X_C$, $X_P$, and $Y_C$, and combined with the corresponding parametric distribution. Finally, the XCDF-t transformation function is applied to the three ‘extended’ distributions. The result is a ‘complete’ distribution of precipitation extremes obtained semi-parametrically. We note that, on average, very few values necessitate the non-parametric extension, but the simple appendage makes for a more rigorous approach. Lastly, we note that the XCDF-t transformation function used here is not the simplified version, the version that assumes the resulting, predicted distribution of extremes is a GPD, but the ‘full’ transformation that assumes no model form of the downscaled distribution.

Figure 38: Illustration of non-parametric extension for one location.
Since our goal is to quantify climate change impacts, our analysis is not complete until return-levels are properly estimated. With the downscaling procedure clearly defined and automated, estimating return-levels is a simple application. For each location/model combination (456), and for a predetermined return period, a return-level point estimate is calculated from the local-level prediction, $F_{yp}$, by selecting the appropriately defined quantile. For reasons previously explained, 25-year return periods were used in all cases. So, for each station, 15,000 sets of GPD parameters are estimated via our Bayesian model, but, for ease of calculation, this number is further thinned to around 1,000. For each of these, the semi-parametric downscaling procedure as above is applied and a single 25-year return-level is calculated. The result is a set of 1,000 25-year return-level estimates for each station/model combination. From this set, a 90% credibility (confidence) interval is then simply extracted.

In Figure 39 below, return-level estimates and 90% credibility intervals are compared to estimates and confidence intervals obtained from our parametric bootstrap procedure for Connecticut stations only. From these results, which are typical of other New England stations, we first notice that, compared to the parametric results, Bayesian return level estimates (red circles) are somewhat similar in the vast majority of cases. Next, we observe that Bayesian credibility intervals (red lines), compared to bootstrap intervals (black lines), are consistently wider indicating substantial increases in variability of return level estimates. While the Bayesian approach was pursued in an attempt to reduce this variability, this result, as illustrated below, was not achieved. It is worth mentioning, however, that our Bayesian intervals are not perfectly comparable to the intervals...
achieved previously. That is, 25-year return level confidence intervals obtained via parametric bootstrapping assumed the simplified XCDF-t form in which $Y_p$ is assumed to be GPD, while the Bayesian (non-parametric extension) procedure made no such assumption regarding the downscaled distribution. The assumptions associated with the simplified XCDF-t are conceivably reducing the variability of the estimates in an artificial way. Lastly, we observe that the Bayesian intervals, compared to the bootstrapped intervals, have longer right tails. This is consistent with our understanding of the distribution of return levels, and suggests that our credibility intervals, while not very precise, are accurate.

Figure 39: Comparison of Bayesian and bootstrap 25-year return level estimates.

Next, we compare Bayesian-derived point estimates for $Y_p$ and $Y_c$ series by observing differences in the 25-year return levels. These plots are directly comparable to those produced previously from parametric bootstrap (and BCa adjustment) methods. In Figure 40 below, red and blue markers indicate positive and negative changes, respectively, and the size of the marker indicates the
magnitude of change. From these results we notice small increases in the \( Y_p \) estimates in many stations, and some larger increases throughout northern New England/Maine. Also, for all model combinations, we observe decreases (negative differences) throughout Connecticut, a result dissimilar to that obtained previously via parametric bootstrapping. Despite some general trends across model combinations, results do appear to be dependent on choice of both GCM and RCM.

Figure 40: Difference in \( Y_P \) and \( Y_C \) Bayesian 25-year return level estimates.

Lastly, below we compare the differences in lower bounds of the 90% confidence intervals for the \( Y_p \) and \( Y_c \) 25-year return levels. Locations marked in red identify positive differences between return level lower bounds, which
correspond to significant increases in 25-year return levels from historic to future periods. From the plots below, we observe blue markers at virtually all stations, and thus little evidence of significant increases between $Y_p$ and $Y_c$ return-levels. These results strongly agree with parametric bootstrap results which also resulted in very few significant increases (red markers). Unlike those results (those obtained from the parametric bootstrapping), however, Bayesian results (below) do not indicate any significant increases (red markers) in Connecticut/southern New England for any model combinations. This is not surprising as the high variability in the Bayesian return level intervals were not expected to reveal many significant changes between historic and projected periods. Certainly, future work in the area will involve identifying methods to reduce such variability and thus more precisely assess climate change.
Figure 41: Difference in 90% confidence interval lower bound for $Y_P$ and $Y_C$ Bayesian 25-year return level estimates.

**RT25 90% CI lower: future - present**

CRCM_ccsm

HRM3_gfdl

RCM3_cgcm

WRFG_ccsm

CRCM_cgcm

HRM3_hadc

RCM3_gfdl

WRFG_cgcm

148
CHAPTER IX: CONCLUSION AND DISCUSSION

Theoretically, this work is heavily influenced by the work of both Michelangeli et al. (2009) and Kallache et al. (2011) who respectively introduced the CDF-t and its parametric extension to extremes, the XCDF-t. While our work and results are made possible by these procedures, we make several refinements. First, we verify that the resulting, downscaled distribution of precipitation exceedances is justifiably a GPD. Within the downscaling procedure, we introduce an updated inflation factor that is suspected to be more robust to high-leverage points. Next, we extend the procedure to produce return level estimates from the downscaled distribution, a widely interpretable estimate critical to climate change impact assessment. A bootstrapping procedure was then employed, and associate measures of uncertainty were produced using both mean correction and BCa adjustments to account for skewness/bias. Furthermore, uncertainty estimates of future return levels are additionally obtained via downscaling under a Bayesian estimation framework. We feel our work is a useful adjunct to the XCDF-t methodology and our results are a valuable contribution to the climate change community.

The application of the downscaling was motivated by the work of Douglas and Fairbank (2011) who investigated trends in extreme precipitation for stations in northern New England. Like their procedure, our predictions are in the form of return levels, but unlike their work, our estimates consider information of the future, the predicted model output, as well as current data. In their work,
the authors conclude that coastal areas of northern New England have significant increases in maximum precipitation records. Based on differences between local-level observations and downscaled future estimates, parametric bootstrapping and BCa techniques yield few significant increases in 25-year return level estimates, mostly throughout southern New England. These results, however, are highly dependent on model choice. In fact, only combinations with the RCM3 regional model yield increases corresponding to Douglas and Fairbanks' results. When we extend our procedure to a Bayesian/MCMC framework, we find larger increases in return level estimates throughout northern New England and decreases throughout southern New England. Bayesian results are also highly dependent on RCM/GCM combination.

Results from our downscaling procedure are further assessed by observing differences in 25-year return level lower bounds between station-level historic and projected periods. From these comparisons, for both parametric bootstrapping and Bayesian estimation of model parameters, local-level return levels are not found to increase significantly from historic to future periods. Differences in return levels from model output, on the other hand, exhibit numerous, significant increases over the same time periods. The disagreement is a concern as our procedure is expected to yield local-level differences that have some correspondence to model output differences. This failure to identify significant local-level increases is likely due to the downscaling procedure itself. That is, downscaling generally yields wide confidence intervals for local-level projections that mask differences from historic periods. As a remedial measure, a Bayesian methodology was pursued as it was believed that this approach would
yield more precise estimates (narrower intervals) and thus identify local-level increases in return levels. However, these Bayesian results were not particularly precise and widespread local-level increases in return levels were never realized. Future work will entail developing suitable techniques to decrease sampling variability, to calculate more precise estimates, and to ultimately achieve more meaningful results.

The choice of RCM and GCM driver has been shown to impact the downscaling procedure substantially. While it was not the primary purpose of this analysis, we have indirectly illustrated the effect this choice can have on downscaling and climate predictions. In their presentation of guidelines for downscaling climate variables, Wilby et al. (2004) note that it is increasingly recognized that any comprehensive impact study should be founded on multiple GCM (or large-scale) model outputs. We strongly support this position as our analysis illustrates the effect that model output can have on downscaling results. Similar findings were reported in Schliep et al. (2009) who analyzed historic output from six RCMs via a spatial Bayesian hierarchical model. Comparing extreme precipitation generated by these models, the authors found that while similar spatial patterns for 100-year return level estimates were produced, the RCMs yielded substantially different estimates. Similar to our results based on model output, differences between RCMs were also identified when accounting for the uncertainty associated with the return level estimates.

Going forward there are additional opportunities to extend and refine our work. Most obviously, we can quantify the effect of RCM and GCM on
projections by means of a functional analysis of variance, or functional ANOVA. Related and recent work in this area has been undertaken by Kaufmann and Sain (2010) and Sain et al. (2010), for example. Such an analysis will hopefully quantify the variability in model output attributable to both RCM and GCM effects.

Next, it is generally believed that winter precipitation is caused by large-scale atmospheric conditions, while summer precipitation is due to local-scale, or regional, processes. Thus, winter and summer predictions may be dominated by global and regional model output, respectively. A seasonal analysis may produce a more nuanced picture of results and, depending on the season, common RCM/GMCs may yield more consistent results. Also, not all models have resolved the snow-precipitation conversion, and some models are expected to provide more accurate winter precipitation projections. Along these lines, Cooley and Sain (2010) analyzed seasonal extreme precipitation in the western United States using a spatial Bayesian hierarchical model. Based on historic and future output of one RCM, the authors found a general increase in 100-year precipitation return levels for the winter season and a significant decrease for the summer season. Their results disagree with our local-level predictions which identified no significant trend in return levels. However, a direct comparison is difficult as Cooley and Sain (2010) based their results solely on RCM output and did not address the effect of uncertainty on differences in return levels between historic and projected periods.
Lastly, the results lend themselves to a regional, spatial analysis. By establishing a spatial correlation between location parameters, data can be suitable pooled to 'share' information and ultimately reduce variability in return level estimates. Also, a smooth spatial process can be applied to the parameters of the projected, local-level extreme value distributions to allow for interpolation at unobserved locations. Cooley et al. (2007) used a spatial hierarchical model to create maps of precipitation return levels and uncertainty measures in a region of Colorado. The authors identify significant differences between the plains and mountains, with mountain areas having extreme distributions with lighter tails and significantly lower return levels. While these results are interesting, they are based solely on historic, station observations and do not consider large-scale model output. Cooley and Sain (2010), on the other hand, analyzed extreme precipitation using a spatial Bayesian hierarchical model for RCM output only (mentioned previously). By combining these approaches within a downscaling framework, by considering model output, historic observations, and the relationship between them, the pooling of data through a spatial model format will improve our uncertainty quantification.
LIST OF REFERENCES


