ESTIMATING PRECIPITATION VOLUME DISTRIBUTIONS USING DATA FROM THE SPATIALLY DENSE COCORAHS NETWORK

Abstract

(Under the Direction of Lynne Seymour)

A novel approach to modeling the distribution of precipitation volume is developed using a combination of traditional and new techniques in spatial statistics. Data are taken from the Community Collaborative Rain, Hail and Snow (CoCoRaHS) network; this network of trained volunteers provides daily precipitation depth measurements across the country. Data for three regions in Colorado were selected due to its spatial density. Combined variogram clouds were calculated for each region, and variograms were fitted to this data using weighted least squares. Precipitation depths were estimated using ordinary Kriging, and bilinear interpolation was used to approximate daily precipitation volumes. Distributions were fitted to the seasonal volume estimates using maximum likelihood, and fit comparisons were done using negative log-likelihood and the Anderson-Darling test.

Key Words: Bilinear Interpolation; CoCoRaHS; Distribution fitting; Kriging; Precipitation; Variogram; Volume.

ESTIMATING PRECIPITATION VOLUME DISTRIBUTIONS USING DATA FROM THE SPATIALLY DENSE COCORAHS NETWORK

by

PATRICK JAMES KRIEBEL

BBA, University of Georgia, 2013

BS, University of Georgia, 2013

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2016

ESTIMATING PRECIPITATION VOLUME DISTRIBUTIONS USING DATA FROM THE SPATIALLY DENSE COCORAHS NETWORK

by

PATRICK JAMES KRIEBEL

Major Professor:	Lynne Seymour
Committee:	Jaxk Reeves
	Thomas Mote

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

Contents

1. Introduction

Accurately modeling precipitation intensity and return periods for extreme storms is an active area of research in climatology and hydrology, often employing mathematical and statistical techniques to assist in the process. Accurate estimates for the likelihood of these events are crucial for designing infrastructure in different communities. As seen by the flooding in Colorado in 2013, Houston in 2015, and West Virginia in 2016, overwhelmed infrastructure to handle run-off precipitation can have catastrophic effects. By definition, these storms are extreme events that lie far out in the tails of the distributions used to model them. Historic methods of modeling these distributions rely on averaging data from fixed stations over long time horizons in order to develop a distribution for the precipitation depth. It is these distributions that are used to estimate the return periods. Averaging in this way smooths out and diminishes the impact of large storms. In addition to using the arithmetic or weighted averages to estimate precipitation depth, intensity is often measured via the rate of accumulation [11] [10]. There is a dearth of resources available for estimating precipitation volume. Rather, depth estimates are used to estimate runoff volume [10]. However, the networks reporting precipitation depth are typically spatially sparse. Colorado, the state containing the three regions explored in this paper, has 24 stations from the Global Historical Climatology Network (GHCN) providing daily reports. By comparison, there are an average of 32 Community Collaborative Rain, Hail and Snow Network (CoCoRaHS) stations in Fort Collins alone providing daily precipitation records. Spatial density is important because it increases the likelihood of measuring the most intense parts of the storm. As a storm develops and moves over land, a spatially dense network

can better measure this storm than a sparse network. Mattingly et al. (2016) [6] show that precipitation depth distributions based on the GHCN network dramatically understate the probability of experiencing high amounts of precipitation. This leads to return periods which are longer than in reality. Furthermore, meteorological records show an increase in average precipitation levels (Karl et al. 1996) [4], and the Intergovernmental Panel on Climate Change (IPCC) projects an increase in the intensity and frequency of extreme weather events. The need for a better means to estimate the likelihood of extreme storms is paramount.

A collection of daily precipitation measurements from the Community Collaborative Rain, Hail and Snow (CoCoRaHS) network for Boulder, Fort Collins, and Lakewood, Colorado, from January 1, 2005 through December 31, 2014 is used to generate seasonal 24-hour precipitation volume distributions. Method for combining variogram clouds for spatial data collected in the same region over a period of time are used to calculate the empirical variogram for each region (Walter et al. 2007)[13]. Spherical variogram models are fit to the empirical data using weighted least squares, and an interactive plotting script developed for this research is used to tune the parameters by hand. Ordinary Kriging is used to generate estimates of precipitation depth over a grid of points encompassing each region. The daily precipitation volumes are calculated using bilinear interpolation. The precipitation volumes displayed clear seasonal dependence, and were subsequently divided seasonally. Seasonal distributions were fit to the data and used to estimate the likelihood of extreme storms as measured by large 24-hour precipitation volumes. This approach found distributions that fit the estimated volumes well. Through the use of spatially dense data and the modeling techniques described

above, more accurate estimates of the likelihood of extreme storm events can be made than with current methods.

2. Methods

2.1 Data and Data Aggregation

CoCoRaHS is a network of volunteers who measure and map precipitation, which is defined as rain, hail, and snow. The volunteers go through training to learn accurate measurement and reporting techniques so that the network can provide accurate data. CoCoRaHS increases the density of high quality precipitation data available to researchers over that of GHCN data. While the GHCN often has only 1 station for a given region, CoCoRaHS can have any number of volunteers.

The data are a collection of daily precipitation records for CoCoRaHS stations in three regions of Colorado: Boulder, Fort Collins, and Lakewood. The three regions from which data were collected have anywhere from 7 to 32 stations reporting on a given day. In particular, our data comes from stations within a circular area with radius 6 km for each region. The reports include precipitation depth in inches, which was converted to tenths of a millimeter, latitude and longitude of the station, and elevation in feet, which was converted to meters. Measurements for CoCoRaHS member stations are taken daily at 7:00 am. We use daily reports from the time period starting January 1, 2005 and ending December 31, 2014. Unfortunately, GHCN stations measure and report the precipitation depth daily at 9:00 am. This removes any spatial correlation from the GHCN and CoCoRaHS

daily precipitation measurements for each region, and makes grouping the daily GHCN and CoCoRaHS precipitation data inappropriate.

For each of the three regions, a grid of roughly 10,000 uniformly spaced points was created within circles centered on the mean latitude and mean longitude of the CoCoRaHS stations. The **R** package `sp` [1][8] was used to sample within a polygon object and select points separated by the same horizontal and vertical distance. Distance in this case is merely the Euclidean distance between pairs of points of the form $(longitude_i, latitude_i)$. As all of the points are within 12 km of each other, it can be safely assumed that the curvature of the surface of the grid is essentially zero, i.e., the grid is flat.

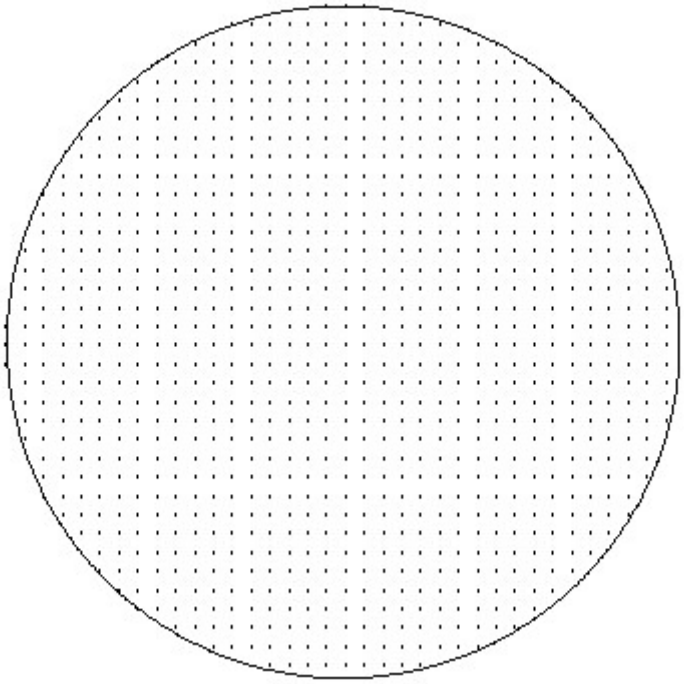


Figure 2.1.1: Example Kriging Grid

2.2 Variogram Estimation

Geostatistics is, "a hybrid discipline of mining, engineering, geology, mathematics, and statistics" [3] used to study and model both spatial trend and spatial correlation of a random process. The first step in geostatistical analysis is finding a suitable variogram model for the spatial process being investigated. The variogram is used to model the spatial dependence of a random process. In this case, we wish to model the spatial correlation of precipitation depth for each of the three CoCoRaHS zones. The precipitation depth at a particular point in a given region is treated as a random variable, and the values reported as measurements from CoCoRaHS stations are realizations of this process. We need a variogram in order to model the spatial dependence of this process so that we may perform Ordinary Kriging over the spatial grids and obtain estimates of precipitation depth at each point. More formally, the variogram is defined as the variance of the difference between field values at two locations (x and y) across realizations of field[3], denoted $Z(x)$, where

$$x = (x_{lon}, x_{lat}), \quad y = (y_{lon}, y_{lat}).$$

In this case, the locations are the coordinates of CoCoRaHS stations, and the realizations are observed precipitation depths. Assuming that the process is stationary implies that it has constant mean, $E(Z(x)) = \mu$, and the covariance depends only on the distance, h, between points

$$Cov(Z(y), Z(x)) = Cov(Z(x+h), Z(x)) = C(h).$$

Then the variogram, $2\gamma(x, y)$, can be written as

$$2\gamma(x, y) = Var(Z(x) - Z(y)) = Var(Z(x + h) - Z(x)) = 2\gamma(h),$$

and the variogram depends only on the distance and direction between locations. If it is assumed that the process is isotropic, then the variogram depends only on $h = |x - y|$, the distance between points, and direction no longer matters. Under the assumptions of stationarity and isotropy, the variogram can be expressed as

$$2\gamma(h) = var(Z(x + h) - Z(x)) = E[(Z(x + h) - Z(x))^2] = 2\sigma^2 - 2C(h),$$

where

- $Z(x)$ is the realization of the point process at point x.
- $Z(x + h)$ is the realization of the point process at point $x+$ distance h, and
- $\gamma(x)$ is called the semivariance

A variogram has three main components:

- **Nugget:** $2\gamma(h), \quad h \to 0$. In a perfect world, the nugget is zero. We would expect that, as samples are taken closer together, the observed values will be more similar. So at a separation distance of zero, we expect no measurement variability. However, measurement errors occur because devices are not perfect. The nugget captures this effect. Variation that occurs on a scale smaller than the sampling distances will also show up in the nugget.

- **Sill:** $2\sigma^2$, is the value reached when the variogram levels off as $h \to \infty$ and $C(h) \to 0$. It represents the variability of the data.

- **Range:** the distance at which the variogram reaches the sill. Points separated by a distance beyond the range are considered uncorrelated.

Matheron (1963)[5] proposed the following unbiased and minimum variance estimator for the theoretical variogram

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (z(x_i) - z(x_j))^2$$

where

- $2\hat{\gamma}(h)$ = estimated variogram at distance h

- $z(x_i)$ = sample value at point x_i

- $z(x_j)$ = sample value at point x_j

- $N(h) = \{(i,j) \colon \text{dist}(x_i, x_j) = h\}$

- $|N(h)|$ = the cardinality of $N(h)$, i.e., the number of pairs points separated by by a distance of h.

In reality we rarely find multiple pairs of points that are *exactly* a distance of h apart. In order to calculate the empirical variogram, one typically selects distance classes and bin the data accordingly, e.g., all the pairs of points between 950 meters and 1050 meters could be binned together for distance $h = 1000$ meters. After distance classes are selected, the pairs of points are binned, and the average squared difference between pairs of observations

is calculated for each bin. The resulting pairs $(h, \hat{\gamma}(h))$ make up the empirical variogram. It is vital that appropriate distance classes are selected because the empirical variogram is used to estimate the true variogram, and the accuracy of Kriging depends on accurately modeling the spatial dependence with the variogram. Of particular importance for selecting distance classes is the number of pairs of points in each class. Too few pairss will inflate the variogram value for a specific distance. Cressie (1985) [2] suggests 30-50 pairs, while more recent research, Webster and Oliver (2001) [14], recommends 100 or more pairs. In order to obtain the threshold of 100 or more pairs, we need 15 or more points at which observations are recorded.

Often it is the case that data are sparse spatially, but dense temporally. This occurs when a small number of locations are sampled frequently over time. CoCoRaHS data has this trait: daily precipitation measurements are taken at a limited number of stations within a larger geographic region. As selecting an accurate variogram model is of the utmost importance, methods to increase the information available to infer the variogram are quite useful. Walter et al. (2007)[13] develop a method to combine variogram clouds from multiple time points for a particular region into an empirical variogram. Their method makes the following assumptions about the data:

- The measurements are taken over time from a random process in the same region that shares a common covariance structure, i.e., the spatial dependence of the random process is not changing over time.
- The process exhibits second-order stationarity and isotropy within the time points.

Their method works as follows

1. Standardize the realizations for each time point. In our case, the precipitation depths recorded at each station for a given day are standardized to a common $N(0, 1)$ scale.

$$s(x_i) = \frac{Z(x_i) - \bar{z}}{SD(z)}$$

where $\bar{z}$ is the mean of the realizations for that time point and $SD(z)$ is their standard deviation.

2. The variogram cloud is made for each time point. The variogram cloud is simply a plot where the vertical axis is the squared difference between realizations at points x and y, $((Z(x_i) - Z(x_j))^2)$, plotted against the horizontal axis $h = \text{dist}(x, y)$ for all pairs of points with $\text{dist}(x, y)$ less than a predefined cutoff. Typically the cutoff is less than half the maximum distance between any two points (Cressise 1993) `[citation]`.

3. Because the realizations have been standardized, the variogram clouds for each time point are on the same scale and can be combined into one variogram cloud. These data are typically analyzed via a plot of squared differences versus distance, and appropriate distance classes can be selected from inspecting the plot.

4. Bin the data according to the selected distance classes and calculate the average squared difference between realizations for each bin. This produces the empirical variogram for the random spatial process of interest.

Finally, parameters for a theoretical variogram are estiamted from the empirical variogram using weighted least squares, where weights are given by

$$\frac{|N(h)|}{h^2}.$$

This is performed using the `fit.variogram()` function in the `gstat` package produced by Edzer Pebesma [7]. The estimates returned by the fitting function minimize the sum of squared errror (SSE) of the model *for the available data points*. However, we know the data points near and beyond the range of the empirical variogram can behave oddly and bias the results of the fitting algorithm. Therefore, the parameter estimates are used as initial values for the interactive plotting tool. The plotting tool allows a user to change the values of the parameters while the plot updates in real-time, displaying the fit variogram, the tuned variogram, and the empirical variogram.

The spherical model was selected because it tends to be more flexible than others. The variogram function increases smoothly at a decreasing rate over the interval $h \in [0, \text{range}]$. It levels off for $h \geq 0$, indicating that covariance for points separated by a distance greater than the range is 0. Finally, the functional form of the model is quite simple. Thee spherical variogram model is given by

$$2\gamma(h) = \begin{cases} c + b\{\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\} & \text{if } 0 < h \leq a \\ c + b & \text{if } h \geq a \end{cases}$$

where the nugget is given by c, the range is given by a, and the sill is $c + b$.

2.3 Kriging

Kriging is a procedure used to interpolate or predict unobserved values of a random spatial process. In particular, if x_0 is a point which was not sampled, and $x_1, \ldots, x_n$ are the sampled locations with observed values $z_1, \ldots, z_n$, then Kriging uses a weighted average

$$\hat{Z}(x_0) = [w_1, \ldots, w_n] \cdot [z_1, \ldots, z_n]' = \sum_{i=1}^{n} w_i(x_0) z_i$$

to interpolate the value at x_0. The weights, $w_1, \ldots, w_n$, are selected such that $\hat{Z}(x_0)$ is unbiased and possesses minimum variance. The Kriging procedure produces these estimates for each point in the grid for every day in the data set. The interpolated values are then backtransformed using the observed mean and standard deviation to the original scale, mm of precipitation.

2.4 Volume Estimation

The ultimate goal of this process is to accurately model the volume of precipitation dropped over the region during a storm event. We use the interpolated daily precipitation values to approximate the volume. Recall Figure 2.1.1, the example Kriging grid, and note that the points are laid out uniformly in the region. Fixing the longitude, the difference in latitudes for consecutive points is the same. Fixing the latitude, the difference in longitude for consecutive points is the same. If one treats the precipitation values as a height associated with each point, then the precipitation volume for a given day is the volume of the object whose surface is given by the points $(x, Z(x))$. We do not have a perfectly smooth surface, rather, we have

a mesh grid. Unfortunately, computing the volume of a mesh grid is no easy task unless the grid adheres to strict properties. Our grid does not possess properties like convexity. However, we can

1. Divide the Kriging grid into squares of 4 neighboring points (Figure 2.4.1, left).
2. Use these points to form the base of a 3-dimensional object whose height is given by the precipitation depth at each point.
3. Find the volume of this 3-dimensional object (Figure 2.4.1, right).
4. Sum the volume for all of these objects to approximate the volume of the entire mesh grid.

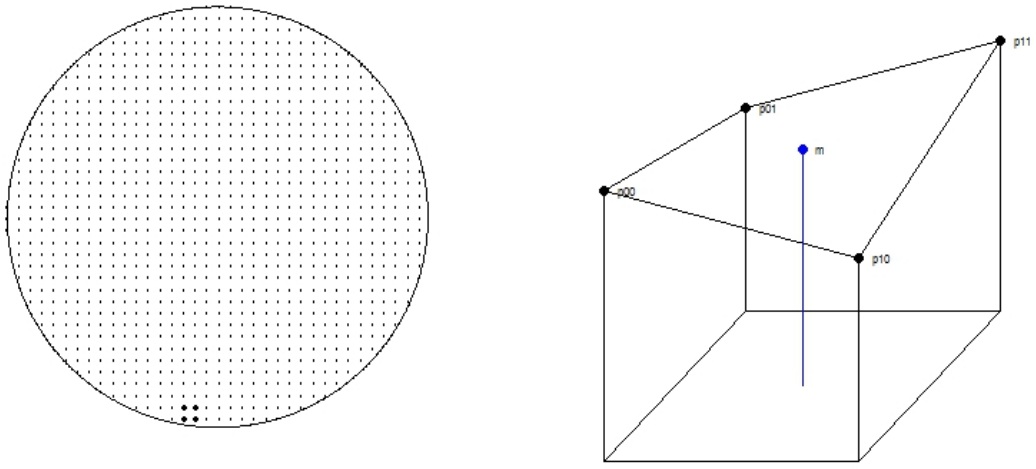


Figure 2.4.1: Kriging Grid Square and 3-D volume object

Suppose the four points of the square are $(0, 0, p_{00}), (0, 1, p_{01}), (1, 0, p_{10}), (1, 1, p_{11})$, they form a unit square with the origin as a corner. We can use bilinear interpolation to approximate the quadric surface passing through those four points. Bilinear interpolation gives the following form for the surface $S(m)$, where $m = (m_1, m_2)$ is an interior point of the square[9]:

$$S(m) = (1 - m_1)(1 - m_2)p_{00} + m_1(1 - m_2)p_{10} + (1 - m_1)m_2p_{01} + m_1m_2p_{11}.$$

Then the volume under this surface is simply found through integration

$$\int_{m_1=0}^{1} \int_{m_2=0}^{1} P(m)dm_1dm_2 = \frac{p_{00} + p_{01} + p_{10} + p_{11}}{4}.$$

To account for the fact that each square in the Kriging grid does not have a unit base, one simply multiplies the volume by the area of the base. This process can be repeated for each square in the grid, and all of the volumes summed to approximate the total volume. Each point in the grid is included in either 1, 2, 3, or 4 bases depending on if it is an outer corner, and outer edge, an inner corner, or an interior point. By classifying the points in the grid, one can compute the volume as the inner product of the vector of precipitation depth at each grid point and the vector of 1's, 2's, 3's, and 4's, indicating how many times each precipitation depth is summed.

2.5 Volume Distribution Fitting

After volumes were estimated for each day, we had a 3,652 estimated volumes. The goal of this research is to accurately model the distribution of the precipitation volume dropped over

a particular region during a storm event. We have particular interest in the likelihood a large volume of precipitation will be released during a storm, as this is when catastrophic damage to infrastructure is likely to occur. Various approaches were taken to model the distribution of precipitation volumes. Daily data were categorized into two sets according to the number of CoCoRaHS stations reporting positive precipitation depths. The first set contains the volumes for every day with at least one station reporting positive precipitation depth. The second contains the volumes for every day with all of the stations reporting positive precipitation depth. For each data set, the daily precipitation volumes were classified by season, and distributions were fit. A custom Matlab script fit all available parametric distributions using maximum likelihood estimation to estimate parameter values, and candidate models were selected based on fit statistics. The Anderson-Darling test was carried out for all candidate models. The final model for each season was selected and used to perform inference on the likelihood of storms dropping certain volumes of precipitation.

3 Analysis and Results

3.1 Procedure Algorithm

The algorithm below provides a general outline of the steps taken to perform this analysis.

1. Variogram Estimation

    1.1. Normalize daily precipitation depths.

    1.2. Calculate daily variogram clouds and merge to one data set.

1.3. Select distance bins and calculate empirical variogram.

1.4. Fit theoretical variogram to empirical variogram.

2. Interpolate Precipitation Depths and Estimate Volume

2.1. Perform Ordinary Kriging for each point in Kriging grid for every day of data.

2.2. Back transform precipitation depth estimates to original scale (millimeters).

2.3. Use bilinear interpolation to estimate daily precipitation volume (in billions of liters).

3. Fit distributions to Volume Estimates

3.1. Subset daily volume data by season.

3.2. Fit distributions to seasonal volume data.

3.3. Assess fit and calculate seasonal volume percentiles.

3.2 Variogram Estimation

Prior to performing the variogram estimation method proposed by Walter et al[13], the data for each region was divided into two sets:

- The daily observations for all days in which at least 1 of the reporting stations records a positive precipitation value.
- The daily observations for all days in which all of the reporting stations record a positive precipitation value.

The first set for each region has between 1100 and 1200 days, while the second has between 400 and 650 days. This subsetting of the full data set was done for a few reasons. First, any day where all stations report zero precipitation will clearly have no precipitation volume, and witnessed no storm events. These days are not of interest for modeling the distribution of precipitation volume during storm events. We must condition on the fact that some precipitation fell. While the first set does satisfy this condition, the data are noisy because there are many days with only a handful of stations reporting a small amount of precipitation. By focusing on days where all stations report positive precipitation, the second set provides a clearer picture of the spatial dependence structure for precipitation in each region.

Boulder and Fort Collins were model data sets, Fort Collins more so than Boulder. The Fort Collins data have an average of 27 stations reporting daily, and Boulder averages 11 stations reporting daily. Recall the assumptions made when combining variogram clouds in order to calculate the empirical variogram.

- The measurements are taken over time from a random process in the same region that shares a common covariance structure, i.e., the spatial dependence of the random process is not changing over time.
- The process exhibits second-order stationarity and isotropy within the time points.

The first assumption is easily satisfied as the precipitation measurements are reported daily, and taken from stations within the same 6 km circle in each region. Because the data come from a 10-year period, it is reasonable to assume the covariance structure for precipitation over each region stays constant. Over a much longer time period, geographic changes could

invalidate this assumption. The assumption of second order stationarity is typically safe to make with precipitation data over a short time horizon. Prior to computing the variogram cloud, a cutoff distance, D, must be selected. Any pair of observed values with a distance between them that is greater than D will not be included in the variogram cloud. The literature recommends selecting D to be a value less than half the maximum distance between any two observed values. For Fort Collins, the cutoff distance was 5.4 km. For Boulder, the cutoff was 4.85 km. The combined variogram clouds for each of these regions had tens of thousands of data points that we used to infer the true variogram model for each region. Figure 3.2.1 displays the combined variogram cloud for Fort Collins. The vertical gaps indicate distances that do not exist for any pairs of stations in the data. By selecting distance classes appriately we can smooth out these gaps and produce an empirical variogram. As distance increases the density of points with higher squared deviations increases, which is typical of spatially correlated data.

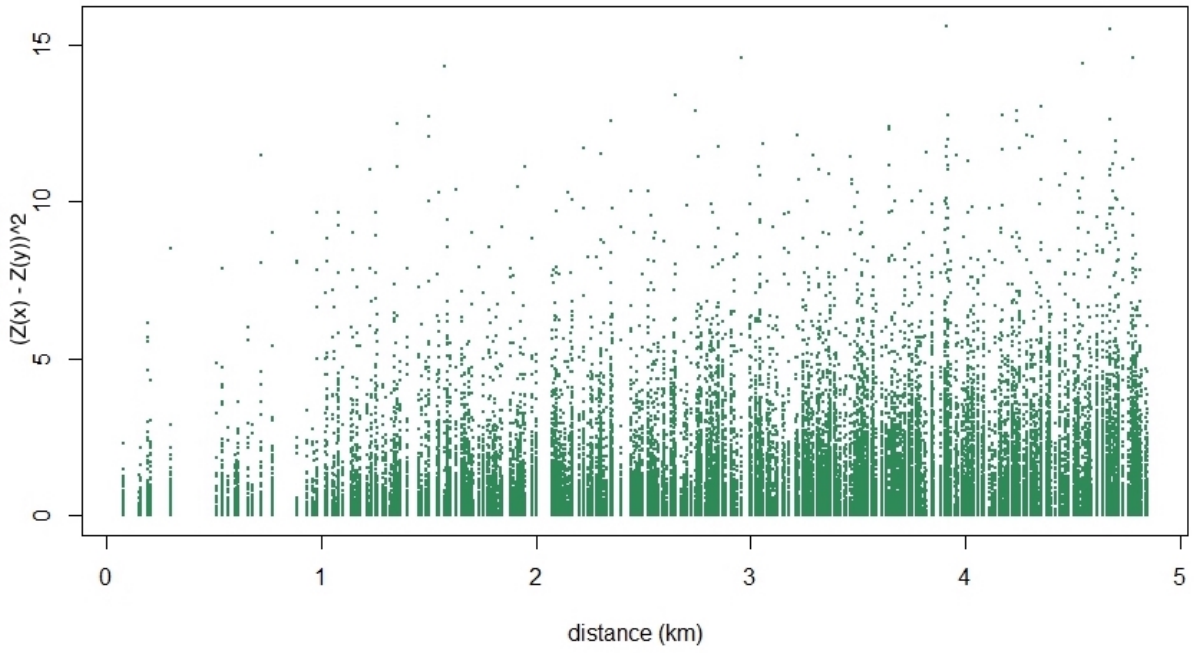


Figure 3.2.1: Fort Collins Combined Variogram Cloud

Typically when one selects distance classes, one must balance the number of distance bins with the number of points in each bin. As the number of bins increases, each bin has fewer points, and the average squared difference of observed values in each bin is more variable. On the other hand, fewer bins means fewer points for fitting a variogram model. By combining daily variogram clouds, we get a massive increase in the number of data points available to infer the true variogram model. We were able to bin each variogram cloud into 25 equally sized bins from a distance of 0 to the cutoff value. Each bin had hundreds or thousands of data points which far surpasses the thresholds of 50 to 100+ in the literature. Having 25 distance classes for computing the empirical variogram leads to 25 pairs of $(h, \hat{\gamma}(h))$ to be used in estimating the variogram parameters. The large number of points provided a clear picture of the spatial relationship for precipitation depth in both Boulder and Fort Collins.

After calculating the empirical variograms for Fort Collins and Boulder, theoretical models were fit using the `fit.variogram()` function from the `gstat` package in **R**[7] and an interactive plotting tool developed for this research.

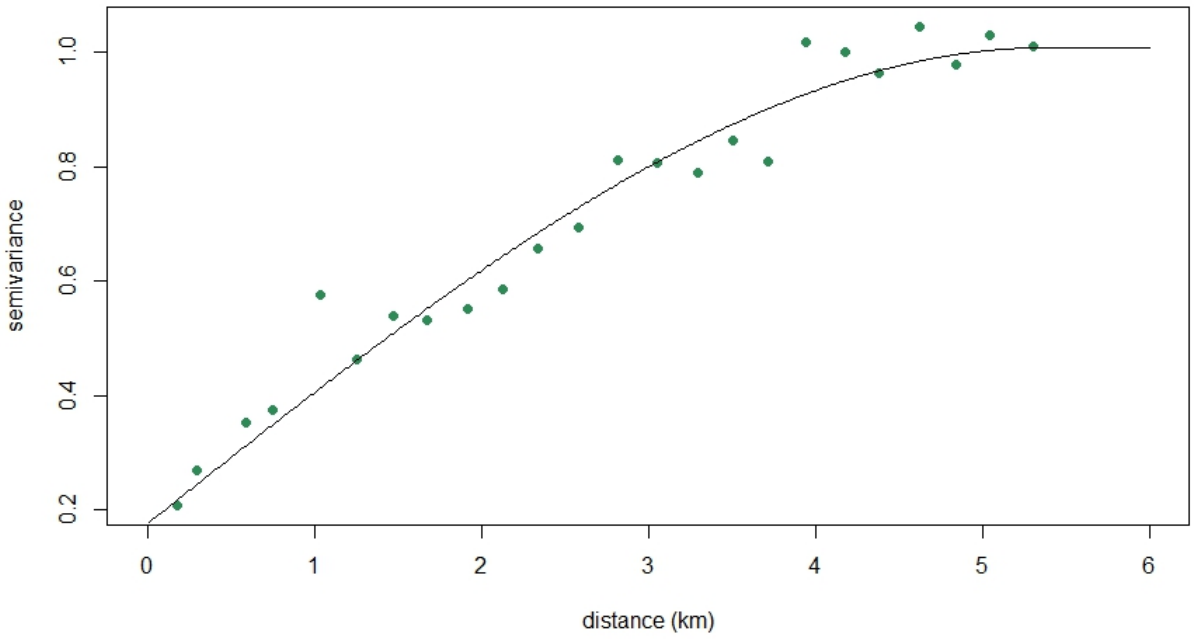


Figure 3.2.2: Fort Collins Empirical and Theoretical Variogram

After calculating the empirical variograms for Boulder and Fort Collins, the theoretical variograms were calculated using a fitting algorithm and graphical parameter tuning. Figure 3.2.2 displays a plot of the empirical and fitted variogram for Fort Collins. The spherical model selected describes the spatial correlation of the data well. The non-zero nugget is the result of tuning by hand. In theory this represents microscale variations or measurement error. Around a distance of 4.5 km the variogram begins leveling off as the correlation between locations decreases rapidly. At 5.38 km, the range, the curve flattens completely, and locations separated by this distance or more are uncorrelated.

The theoretical model for Boulder is given by (range in km)

$$2\gamma(h) = \begin{cases} 0.155 + 1.096\{\frac{3}{2}\left(\frac{h}{4.096}\right) - \frac{1}{2}\left(\frac{h}{4.096}\right)^3\} & \text{if } 0 < h \leq 4.096 \\ 0.155 + 1.096 & \text{if } h \geq 4.096 \end{cases}.$$

The theoretical model for Fort Collins is given by (range in km)

$$2\gamma(h) = \begin{cases} 0.175 + 1.009\{\frac{3}{2}\left(\frac{h}{5.381}\right) - \frac{1}{2}\left(\frac{h}{5.381}\right)^3\} & \text{if } 0 < h \leq 5.381 \\ 0.175 + 1.009 & \text{if } h \geq 5.381 \end{cases}.$$

That both models have similar parameter estimates is a good sign. Boulder and Fort Collins are in the same state and only 74 km apart. We would expect similar spatial dependence structures for precipitation in the two regions. However, the Flatiron Foothills border the entire western side of Boulder, and have a large impact on the way storm systems move in and across Boulder. The absence of this feature in Fort Collins likely contributes to the differences in the two theoretical variograms.

Lakewood has 17 stations in total with around half of them reporting most days. Still, there was not enough data at different distances to accurately infer the theoretical variogram.

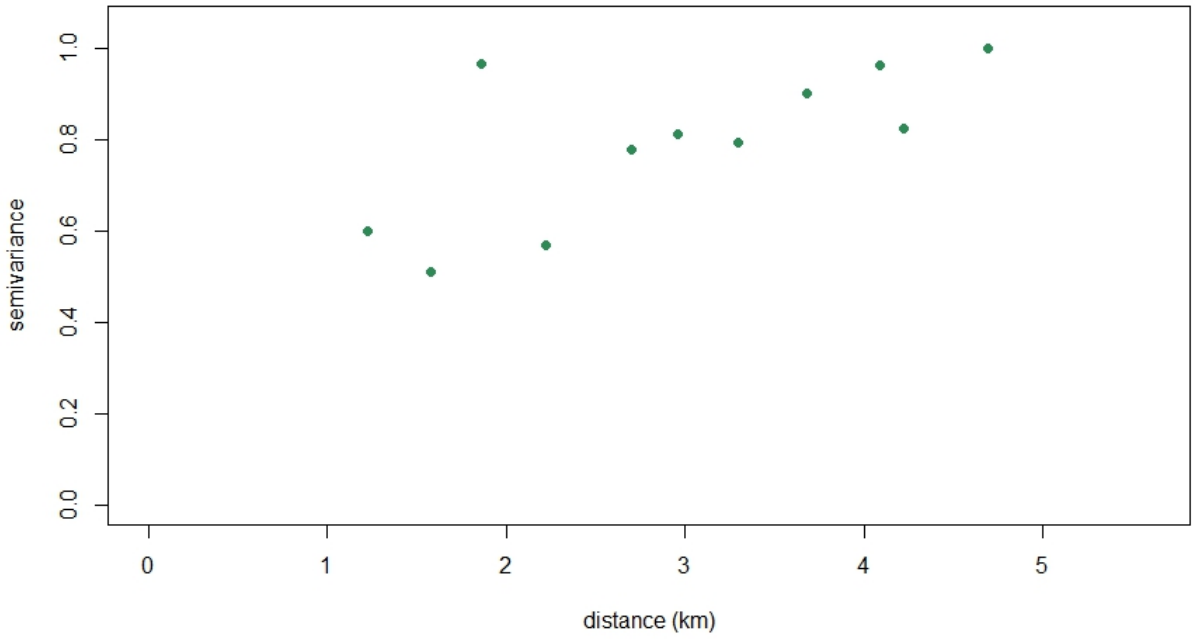


Figure 3.2.3: Lakewood Empirical Variogram

Figure 3.2.3 displays the empirical variogram for Lakewood. The minimum distance between stations was about 1 km. This highlights a limitation of the variogram cloud overlay method. While it can provide more information than is available spatially sparse data, if the repeated measurements do not have distances between pairs throughout the entire range, then the resulting empirical variogram cloud will not be useful for estimating the theoretical variogram. Lakewood is 55 km from Boulder, and 110 km from Fort Collins. Due to its proximity to Boulder, it was decided to use Boulder's fitted variogram for Lakewood.

3.3 Kriging Daily Precipitation Depths

Once variogram models were selected, ordinary Kriging was performed over each region's Kriging grid for every day from January 1, 2005 through December 31, 2014. Ordinary Kriging provides estimates for the precipitation depth at each point in the Kriging grid as a weighted average of the observed precipitation depth at each of the stations reporting that day. As stated in the methodology section, the weights for the observed values are selected such that the estimator is unbiased and has minimum variance. It should be noted that the Kriging estimates were back-transformed to the original scale. Recall that daily precipitation values were standardized, i.e., the mean was substracted from each observation and the result was divided by the standard deviation. So the kriged estimates for each day were multiplied by the standard deviation of the observed precipitation values for that day, and the mean of the obseved values was added to this result. Ordinary Kriging assumes the random process exhibits second-order stationarity, isotropy, and that the realizations of the process are normally distributed. The isotropy and second-order stationarity assumptions are met as discussed in section 3.2. However, the normality assumption is not. Fortunately, the daily precipitation depths produced by Ordinary Kriging were only used for smoothing purposes, not for inference. So the predication variance, which will be impacted by violating the normality assumption, is not a concern.

Once the daily precipitation depths were calculated for every day in each region, daily volume estimates were computed from the back-transformed daily precipitation values using bilinear interpolation.

Unfortunately, there is no resource available where one can designate a region on the globe

and a date, and receive the precipitation volume that fell over the specified region on the given date. This made benchmarking the volume estimates difficult. The only means of benchmarking available was through an application on the U.S. Geological Survey's website [12]. This tool allows one to specify:

1. The precipitation depth, p.

2. The area of the region, A.

and it outputs a precipitation volume estimate for the region using $p * A$. For a random sample of 20 days from each region, the volume estimates produced by bilinear interpolation were within a few million liters of the USGS estimate (using the mean precipitation depth for that day and the area of the Kriging grid). When many of the volumes are in the billions of liters range, plus or minus a few million liters of the USGS calculator was considered adequate, especially since the USGS calculator is a less refined method than ours.

3.4 Modeling The Distribution of Precipitation Volumes

This was the ultimate goal of the research: to develop a means of modeling the distribution of precipitation volume during storm events for each of the regions. The first step in this process was exploring the histograms of the daily volume estimates for each region. The histograms for the volumes corresponding to the days when at least 1 station reported positive precipitation are skewed more heavily than than the distribution of volumes for days when 100% of the stations report precipitation. See Figures 3.4.1 and 3.4.2 for the regional volume histograms for each of the two reporting conditions. As expected, for each region, there

are many more days with an estimated volume near 0 liters for the set of days when at least 1 station reports precipitation. Furthermore, we see large outliers far to the right of each histogram. These huge volume estimates come from September 2013, when Colorado experienced historic levels of rainfall and widesrpead flooding.

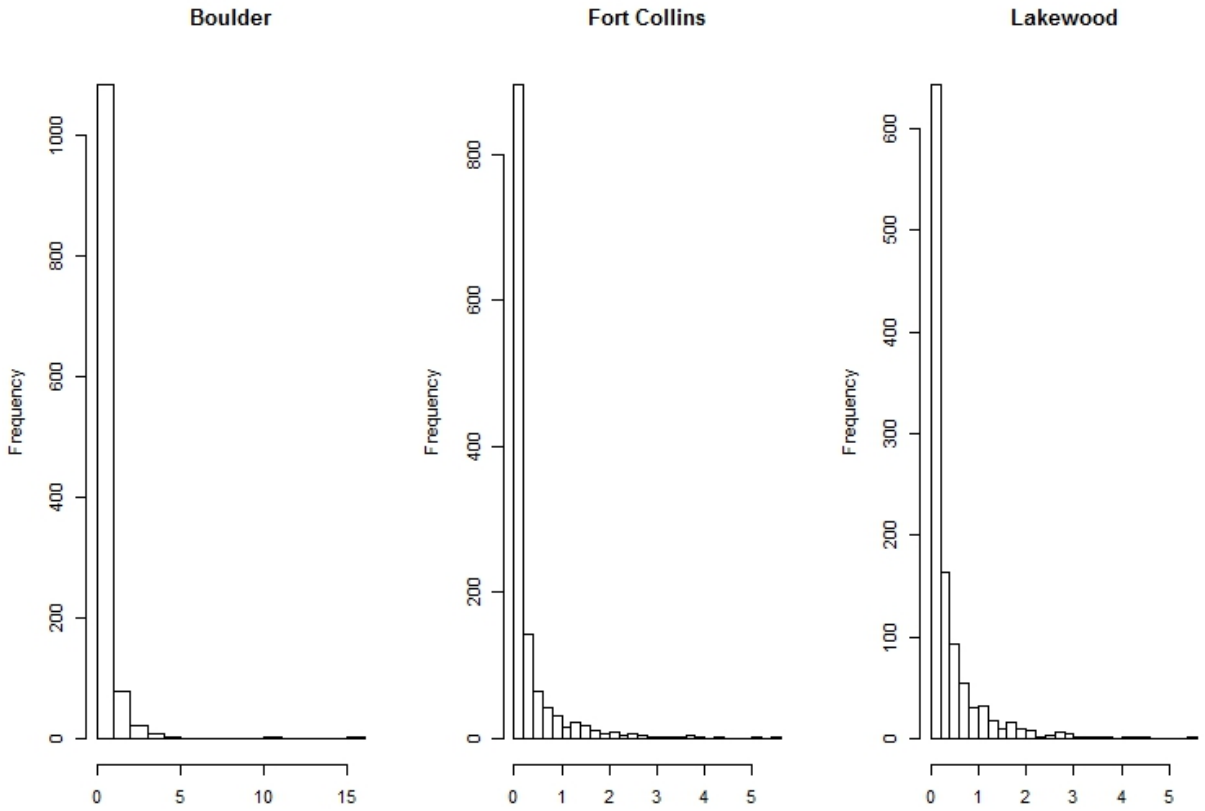


Figure 3.4.1: Precipitation Volumes (billion liters): at least one station reporting precipitation

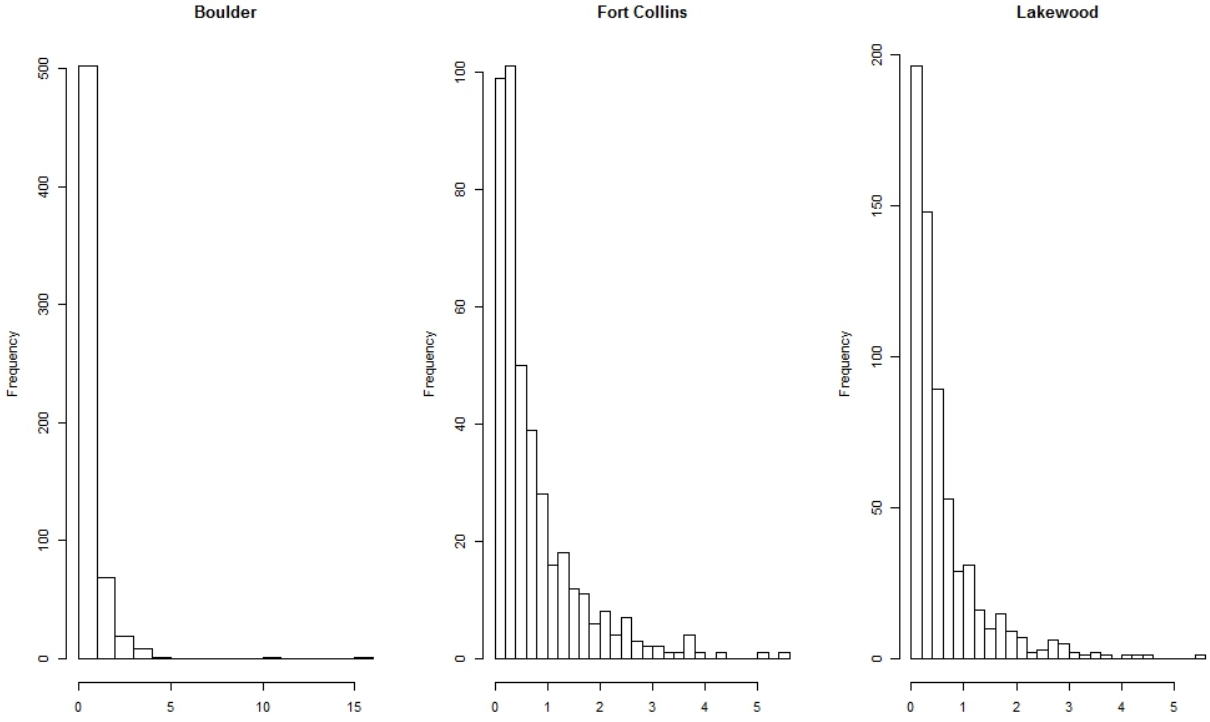


Figure 3.4.2: Precipitation Volumes (billionliters): all stations reporting precipitation

Using Matlab, an automated procedure was used to fit all parametric distributions available, using maximum likelihood estimates for the parameters, to each of the two sets of volumes for each region. Disastrous results occurred when distributions were fitted to the volumes for each region. This happened regardless of whether the volumes were subset to days when at least one station reports precipitation or days when 100% of stations reported precipitation. At this point, the volume data was subset by season in order to get a clearer picture of the distributions. Because the volume estimates are conditioned on the fact that some stations record precipitation for that day, it was necessary to get an idea of the probability that precipitation will occur.

The percentage of days where at least one station reports precipitation were calculated for each month from January 2005 through December 2014. These monthly percentages were then grouped according to season. Figure 3.4.3 displays the seasonal probability of precipitation for each region. From the boxplot it is clear that the probability of precipitation is highly dependent on the season. The summer months have the highest probability of precipitation, while winter and fall have the lowest. Winter is the most consistent season in terms of probability of precipitation, while summer is the most variable. Boulder appears to be the most variable of the three regions, which could be due in part to The Flat Iron Foothills resting along the entire western edge of Boulder. This mountain range has a significant impact on the weather experienced in Boulder. For a given season, the probability of precipitation is similar across the 3 regions.

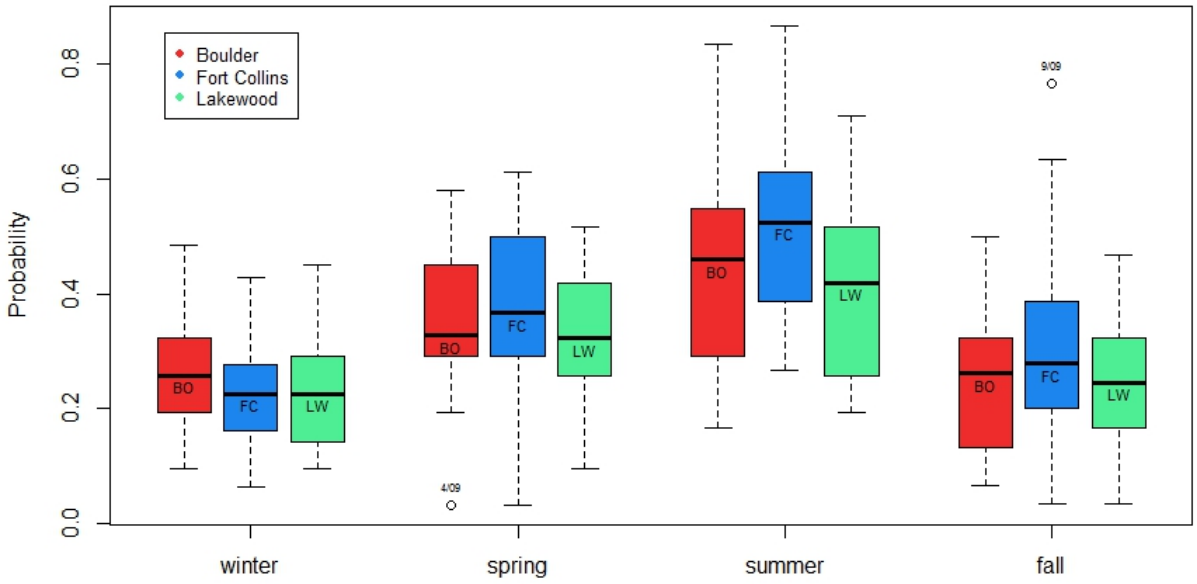


Figure 3.4.3: Seasonal Probability of Precipitation

An attempt was made to incorporate data on the El Niño and La Niña effects, as it is believed these phenomena have a large impact on weather patterns. However, due to the limited time horizon of 10 years and weak El Niño and La Niña effects during this stretch, this did not produce useful results. Figure 3.4.4 displays the seasonal precipitation volume for each region, and Figure 3.4.5 elucidates the seasonal distributions by excluding the extreme outlier present due to the historic levels of rainfall experienced during September 2013 in Colorado. Aside from December 2006 when blizzards swept through Colorado and dropped historic snowfall, winter is the least variable and has the lowest average precipitation volume. Fall and spring experience the widest range of precipitation volumes, and summer is only slightly more consistent.

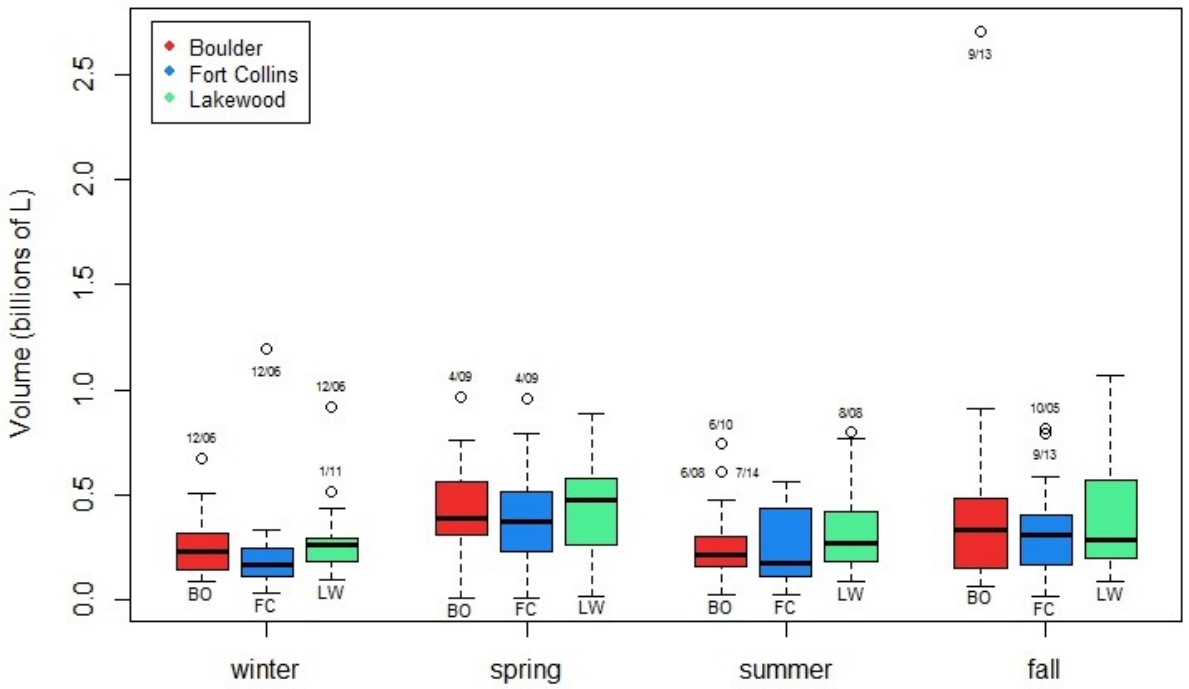


Figure 3.4.4: Seasonal Average Precipitation Volume (L)

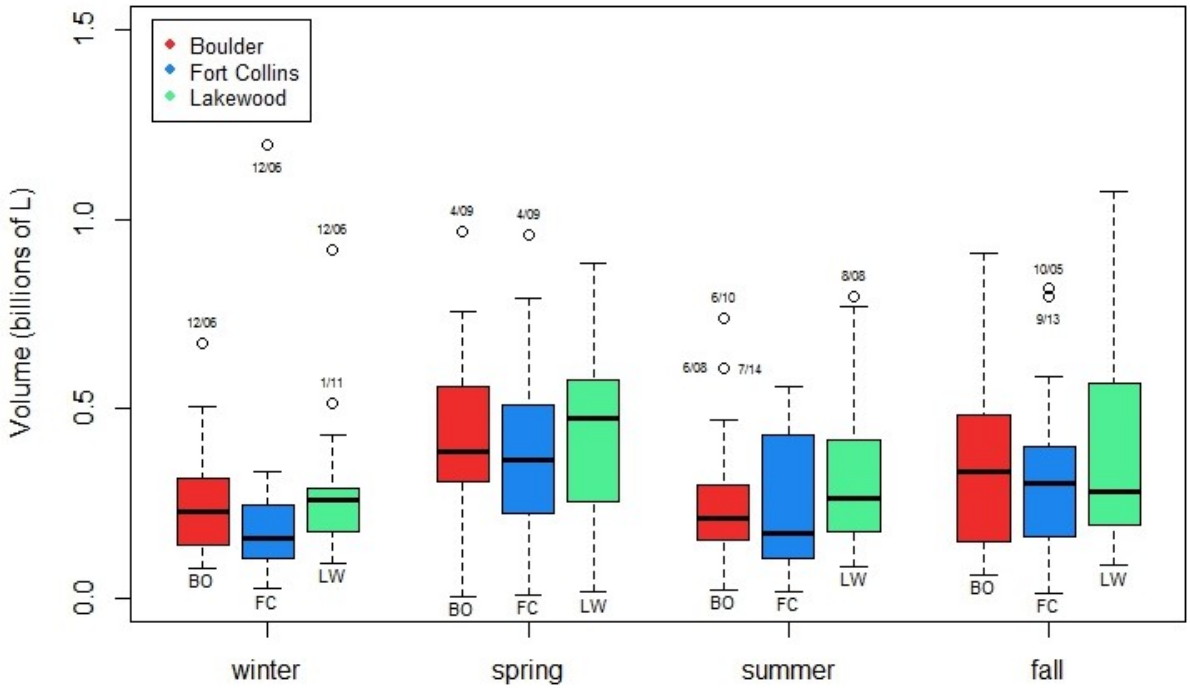


Figure 3.4.5: Seasonal Average Precipitation Volume (L), excluding 2013 floods

Distributions were fit to the seasonal precipitation volumes for each region. Dividing the

volumes seasonally produced much better fits. From the distributions fit by the automated script, the top four candidates for each volume data set were selected according to negative log-likelihood. The $10^{th}, 25^{th}, 50^{th}, 75^{th},$ and 90^{th} percentiles were calculated for each distribution, and compared to the same precentiles from the data. The candidate distributions fit the seasonal data remarkably well. Table 3.4.1 contains the results from the fit analysis for Fort Collins spring precipitation volume distribution, and Figure 3.4.6 displays the histogram of Fort Collins's spring precipitation volumes with the candidate distributions plotted. The observed and estimated percentiles are nearly identical. The corresponding tables and figures for the other region and season combinations can be found in the appendix.

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			120	201	570	1250	2100	**3870**
Birnbaum Saunders	2663	(0.644)	134	244	515	1090	1980	0.965
Inverse Gaussian	2664	(0.323)	138	237	469	993	1950	0.958
Lognormal	2667	(0.351)	136	255	514	1040	1950	0.958
Exponential	2673	(0.253)	89	243	585	1170	1950	0.975

Table3.4.1: Fort Collins Spring Distributions

In addition, the maximum observed volume for Fort Collins in spring, approximately 3.1 billion liters, corresponds to the 95-95% percentile for each of the candidate distributions. These tables for each combination of season and region can be found in the Appendix. The best model from among the four candidates for each region and season was selected according to the Anderson-Darling test. This test has the null hypothesis that the precipitation volume data come from a specified distribution, and the alternate hypothesis that the data do not come from the specified distribution. The candidate distribution with the largest p-value was selected as the best. Larger p-values indicate a higher probability of observing the given data

assuming they come from the specified distribution. Table 3.4.2 lists the best fit distribution for each region and season with the p-value from the Anderson-Darling test in parentheses.

Boulder	**Model**	**NLogL**	p**-value**	Parameter Estimates
Winter	Birnbaum Saunders	2673	(0.968)	$\beta = 2.453 \times 10^8$ and $\gamma = 1.113$
Spring	Birnbaum Saunders	3843	(0.850)	$\beta = 4.319 \times 10^8$ and $\gamma = 1.151$
Summer	Inverse Gaussian	3525	(0.968)	$\mu = 5.275 \times 10^8$ and $\lambda = 3.361 \times 10^8$
Fall	Generalized Extreme Value	2623	(0.953)	$k = 1.040$, $\sigma = 1.785 \times 10^8$ and $\mu = 1.838 \times 10^8$
Fort Collins	**Model**	**NLogL**	p**-value**	Parameter Estimates
Winter	Loglogistic	1597	(0.9961)	$\mu = 19.498$ and $\sigma = 0.464$
Spring	Birnbaum Saunders	2663	(0.6439)	$\beta = 5.151 \times 10^8$ and $\gamma = 1.131$
Summer	Inverse Gaussian	2629	(0.8915)	$\mu = 7.791 \times 10^8$ and $\lambda = 4.247 \times 10^8$
Fall	Birnbaum Saunders	1977	(0.9447)	$\beta = 5.316 \times 10^8$ and $\gamma = 1.074$
Lakewood	**Model**	**NLogL**	p**-value**	Parameter Estimates
Winter	Loglogistic	2652	(0.9960)	$\mu = 19.390$ and $\sigma = 0.523$
Spring	Birnbaum Saunders	4092	(0.8681)	$\beta = 4.106 \times 10^8$ and $\gamma = 1.176$
Summer	Birnbaum Saunders	3710	(0.9784)	$\beta = 3.701 \times 10^8$ and $\gamma = 1.226$
Fall	Birnbaum Saunders	2833	(0.8146)	$\beta = 3.600 \times 10^8$ and $\gamma = 1.190$

Table 3.4.2: Seasonal Best Fit Distributions

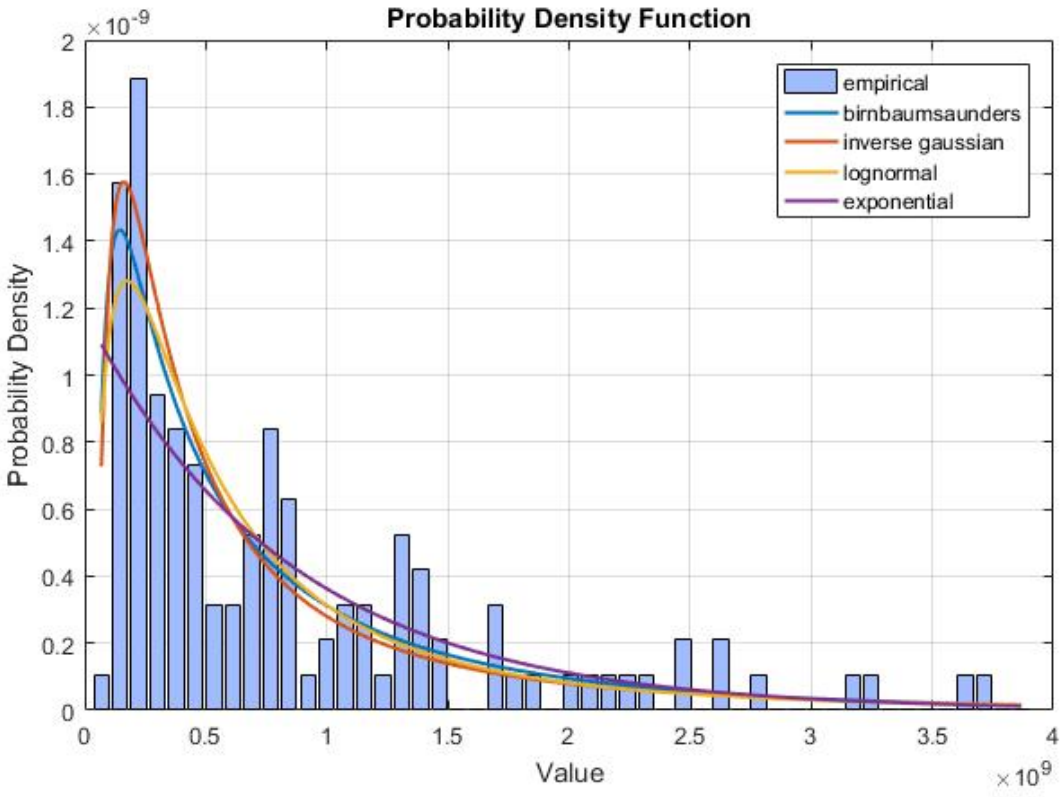


Figure 3.4.6: Fort Collins Spring

4. Conclusion

The distributions developed in the previous section accurately model the seasonal precipitation volumes in the regions of interest. The distributions are skewed far to the right, indicating a rapidly decreasing probability of massive volumes of precipitation in a 24-hour period. Through use of spatially dense data and geostatistical techniques, the methods presented better account for the variability of precipitation depth over a given region during a storm than current methods of smoothing averages relying on few stations. It should be noted that the short time span of available CoCoRaHS data hinders the ability to incorporate certain climatological factors into the model which are known to have meaningful impact on the weather systems in North America. El Niño and La Niña effects can be incorporated into climate models through use of the Southern Oscillation Index, but their effects are more clearly pronounced with a long time horizon. However, the size of the CoCoRaHS network has grown rapidly to 5,177 stations reporting daily across the United States, Puerto Rico, southern Canada, and, even, the Bahamas. The value of this data source should only increase with time, as it becomes more spatially and temporally dense. Going forward, improvements could be made to the process of fitting variograms for the daily precipitation depths. While Fort Collins and Lakewood were both relatively flat regions, Boulder's geography could be better accounted for through the inclusion of data like elevation and proximity from the Flat Iron Foothills. Any improvements to the variogram models will lead to overall increases in accuracy, as they provide the means of inferring precipitation depths at unsampled points. Approaching the variograms from a seasonal perspective could help account for the seasonal variation that was displayed in the estimated volumes. The approach developed to modeling

seasonal 24-hour precipitation volumes provides a novel and solid framework for new analyses of extreme storm events. The need to more accurately model these outlier events will only increase as climate change results in more variable weather systems and reliance on techniques which explicitly smooth out the effects of outliers wanes.

5. References

References

[1] Roger S. Bivand R.S., Pebesma E.J., Gomez-Rubio V., 2013. *Applied spatial data analysis with R*, Second edition. Springer, NY. http://www.asdar-book.org/.

[2] Cressie N.A, 1985. When are relative variograms useful in geostatistics? *Mathematical Geology* **17**: 693–702.

[3] Cressie, N.A. *Statistics for spatial data*. Hoboken, NJ: John Wiley & Sons, Inc, 1993. Print.

[4] Karl, T.R., Knight, R.W., Easterling, D.R., et al., 1996. Indices of climate change for the United States. *Bulletin of the American Meteorological Society*, **77**: 279-292.

[5] Matheron, G, 1963. Principles Of Geostatistics. *Economic Geology* **58.8** 1246-166.

[6] Mattingly, K., Miller, P., Seymour L. (2016). *Estimates of extreme rainfall frequency in urban areas derived from spatially dense rain gauge observations*. Manuscript submitted for publication.

[7] Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**: 683-691.

[8] Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. *R News* **5**: 9-13. `http://cran.r-project.org/doc/Rnews/`.

[9] Press, W.H. *Numerical recipes in C : The Art of Scientific Computing*. Cambridge Cambridgeshire New York: Cambridge University Press, 1992. Print.

[10] United States. Dept. of Agriculture. Hydrology National Engineering Handbook. Washington. March 1993. Web. 23 June 2016. `http://directives.sc.egov.usda.gov/OpenNonWebContent.aspx?content=18383.wba`

[11] United States. NOAA. Point Precipitation Measurement, Areal Precipitation Estimates and Relationships to Hydrologic Modeling. Web. 13 May 2016. `http://www.srh.noaa.gov/abrfc/?n=map`

[12] United States. USGS. Rainfall calculator: How much water falls during a storm? Web. 1 June 2016. `http://water.usgs.gov/edu/activity-howmuchrain.html`

[13] Walter, JF, et al., 2007. Combining Data From Multiple Years Or Areas To Improve Variogram Estimation. *Environmetrics*, **18.6**: 583-598.

[14] Webster R, Oliver M. 2001. *Geostatistics for Environmental Scientists*. Wiley: England.

6. Appendix

6.1 Figures

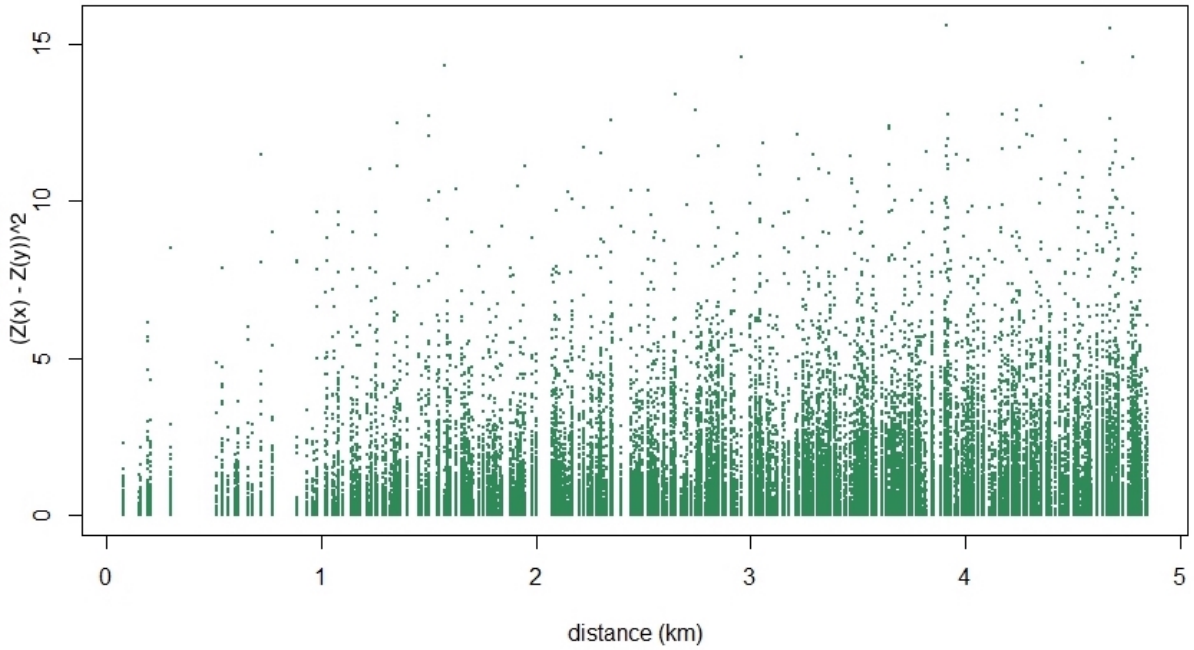


Figure 6.1.1: Fort Collins Combined Variogram Cloud

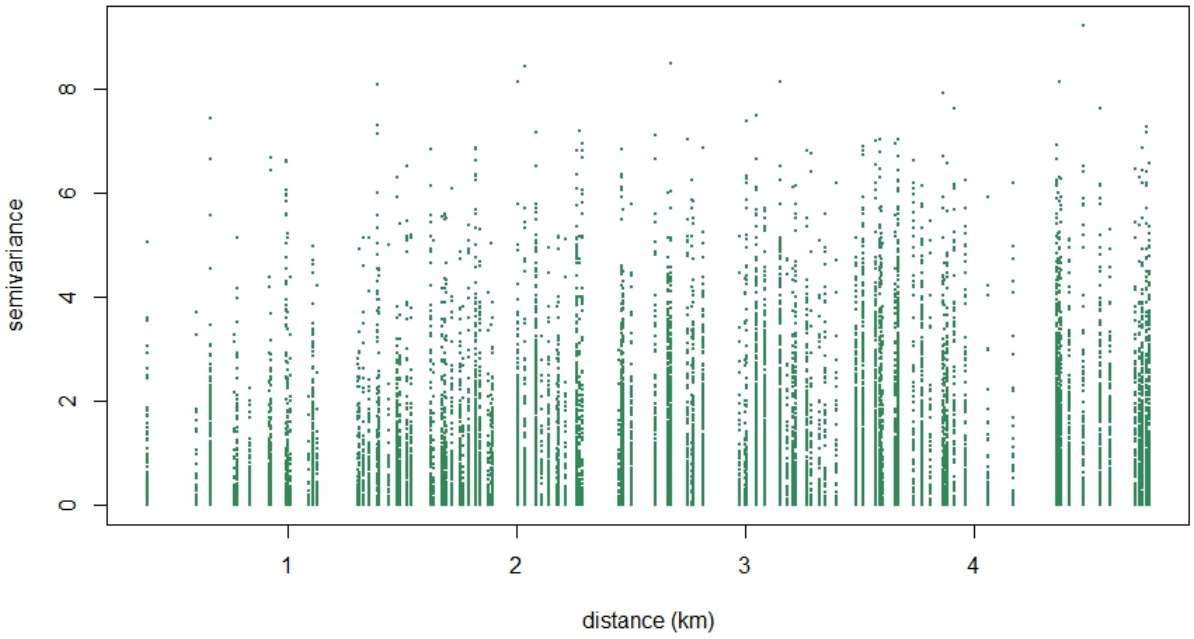


Figure 6.1.2: Boulder Combined Variogram Cloud

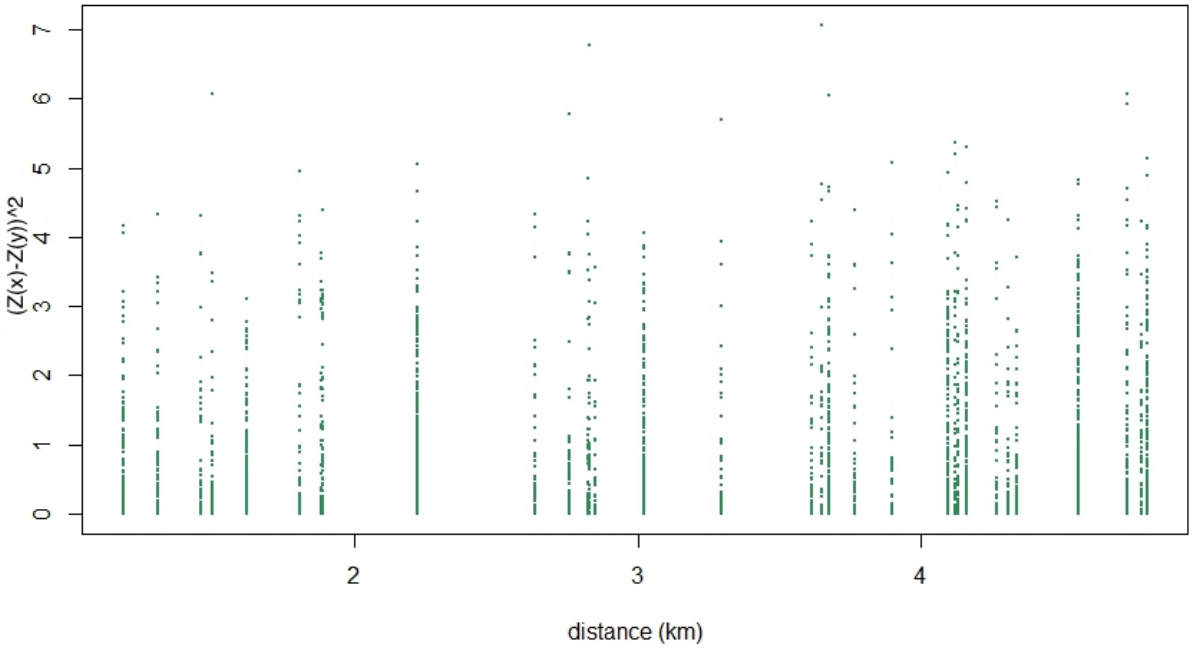


Figure 6.1.3: Lakewood Combined Variogram Cloud

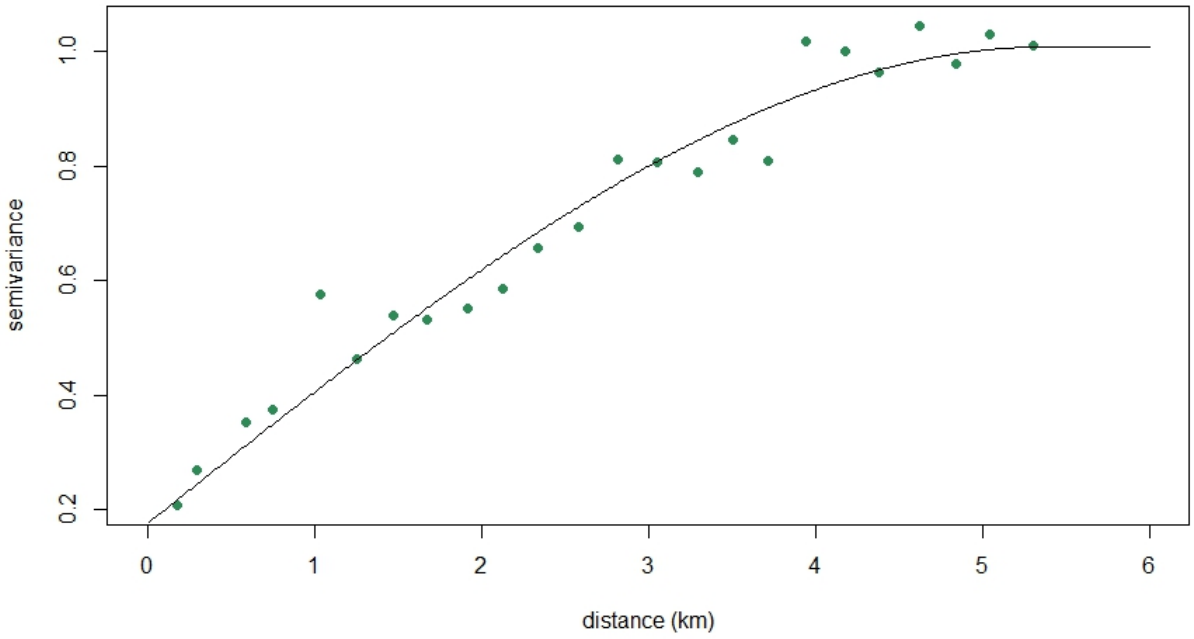


Figure 6.1.4: Fort Collins Empirical and Theoretical Variogram

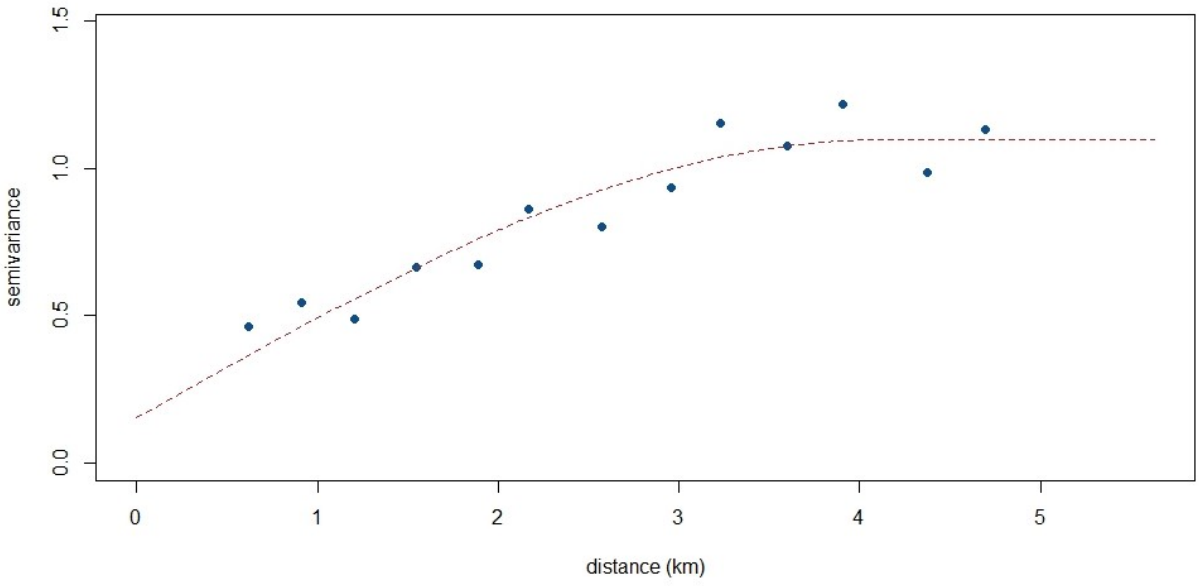


Figure 6.1.5: Boulder Empirical and Theoretical Variogram

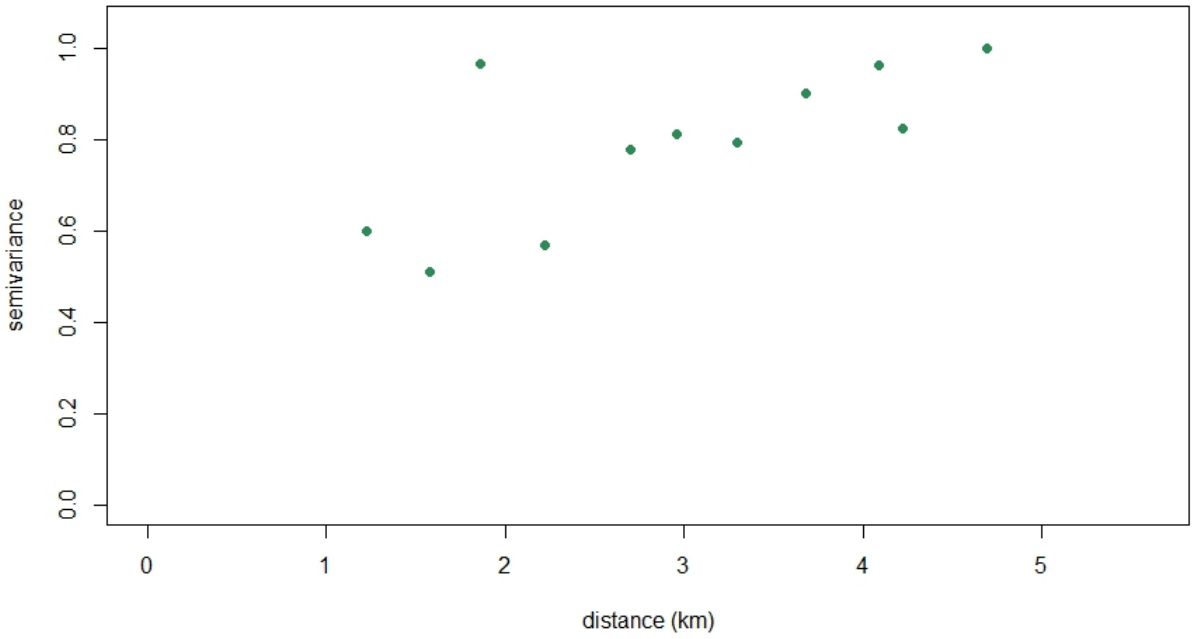
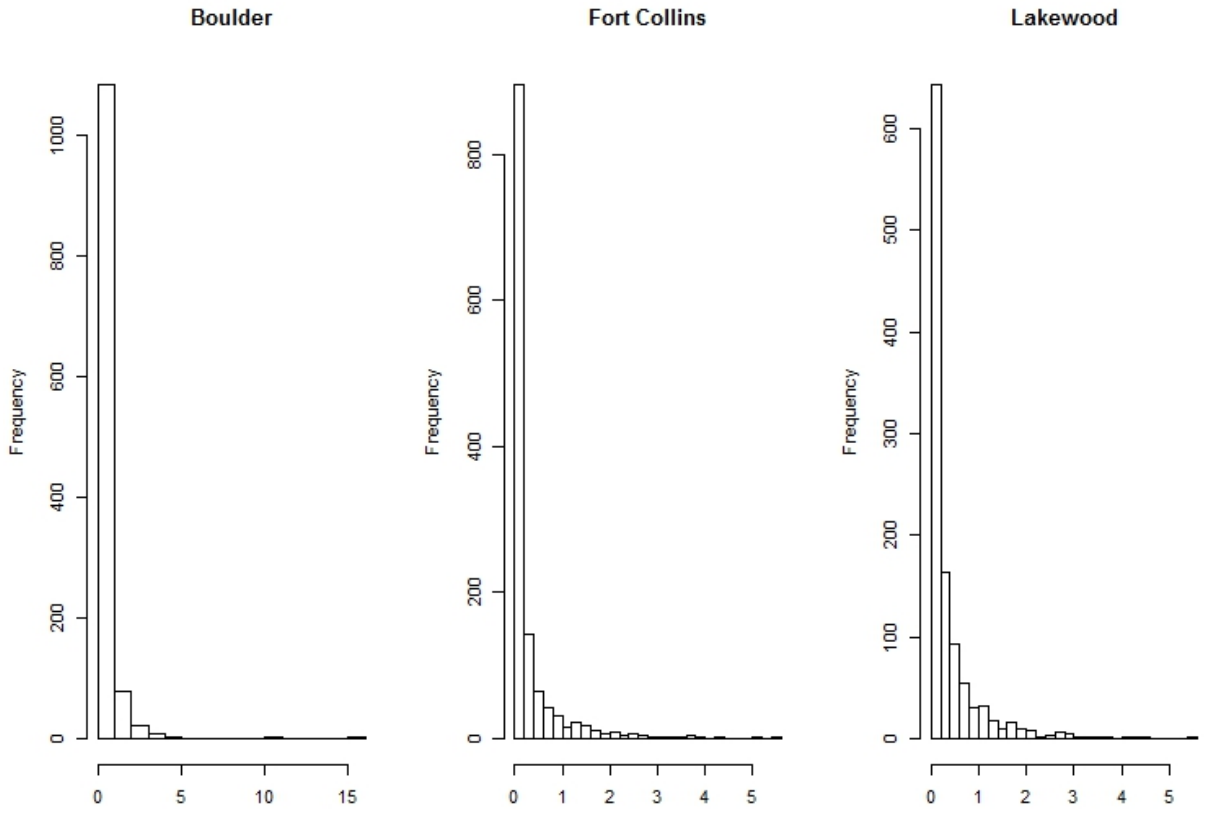


Figure 6.1.6: Lakewood Empirical Variogram



Precipitation Volumes (billion liters): at least one station reporting precipitation

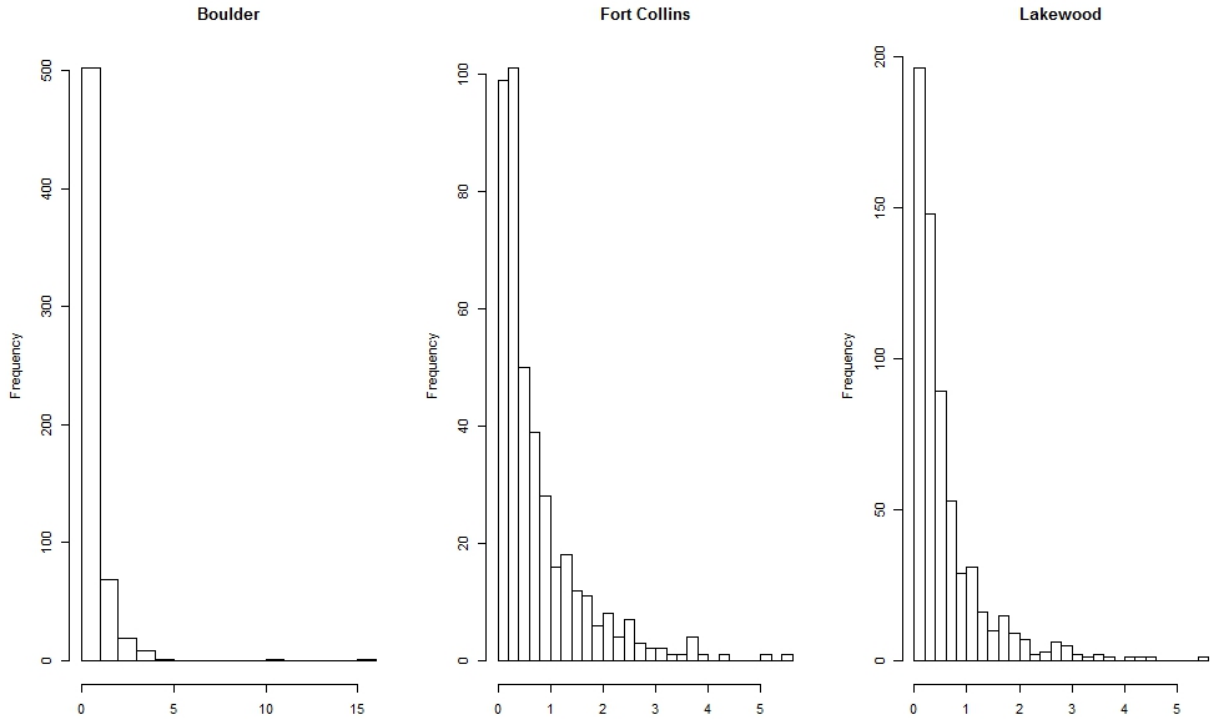


Figure 6.1.7: Precipitation Volumes (billionliters): all stations reporting precipitation

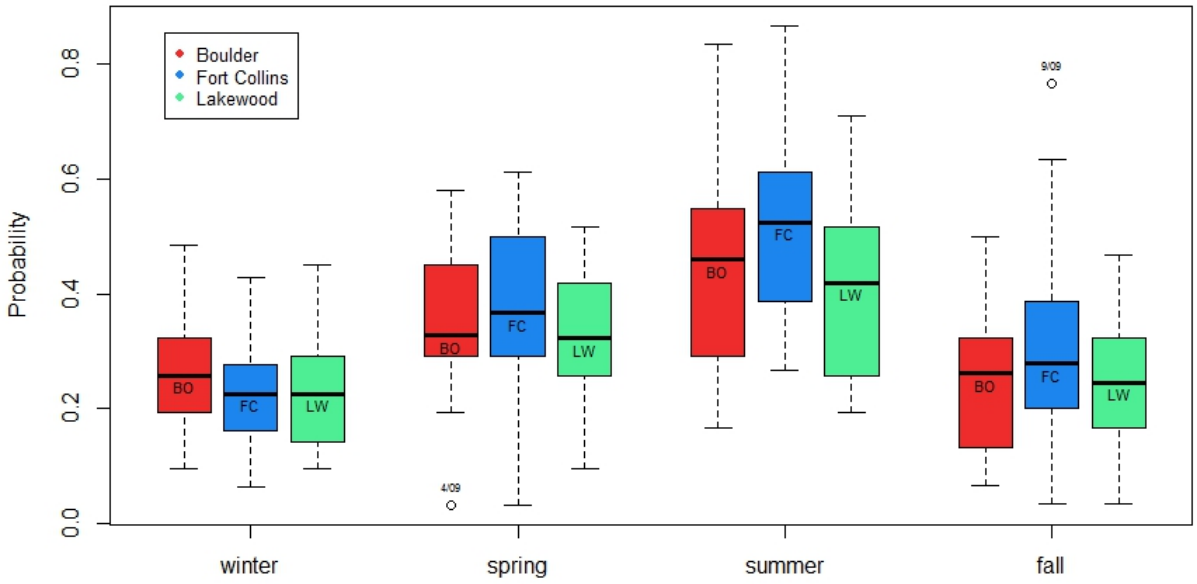


Figure 6.1.8: Seasonal Probability of Precipitation

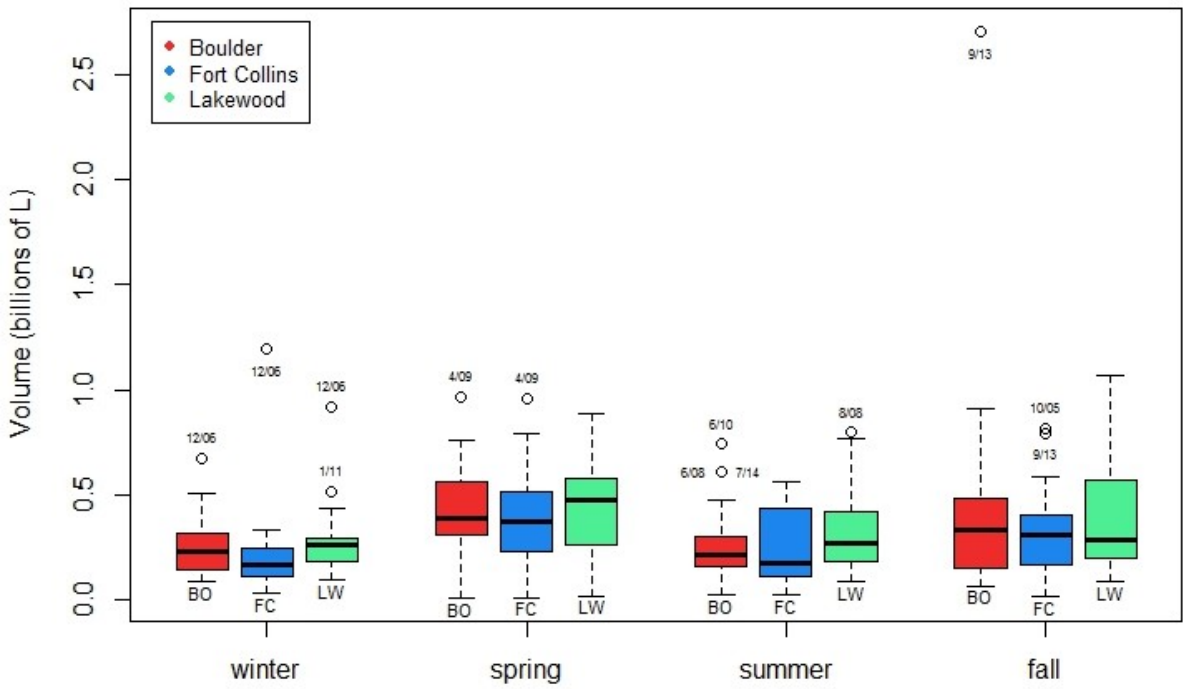


Figure 6.1.9: Seasonal Average Precipitation Volume (L)

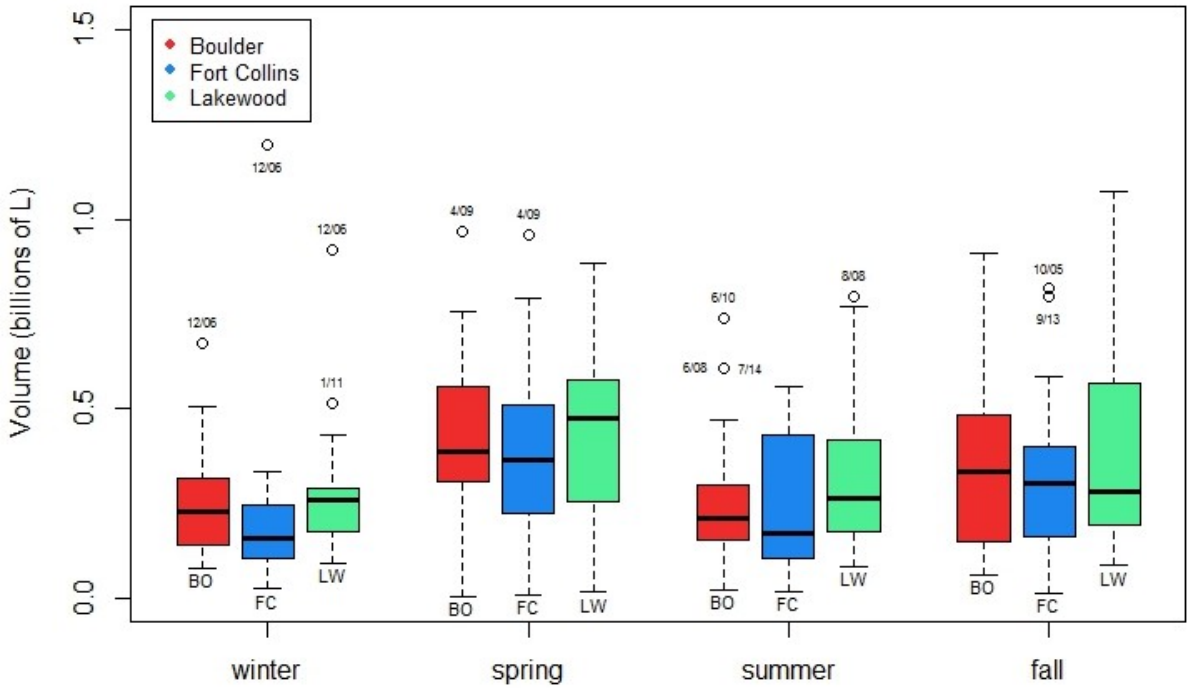


Figure 6.1.10: Seasonal Average Precipitation Volume (L), excluding 2013 floods

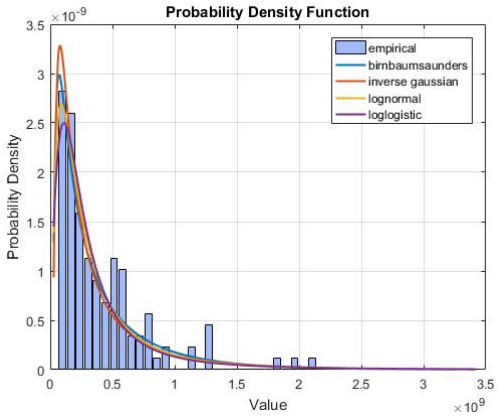


Figure 6.1.11: Boulder Winter

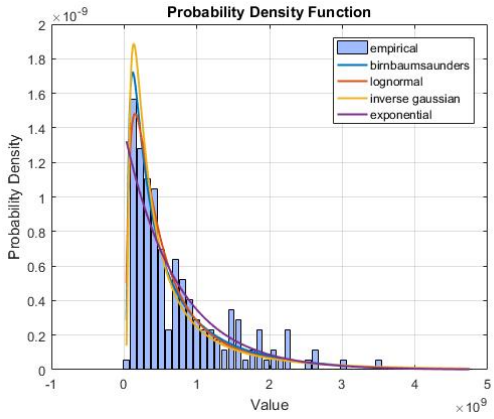


Figure 6.1.12: Boulder Spring

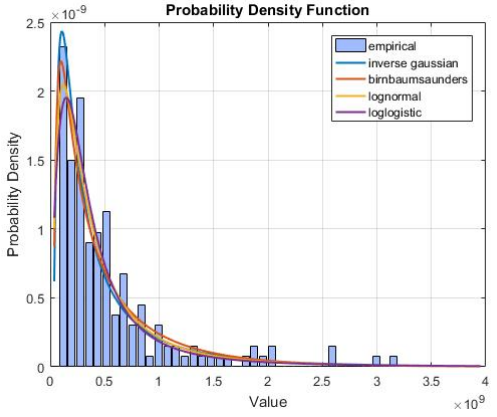


Figure 6.1.13: Boulder Summer

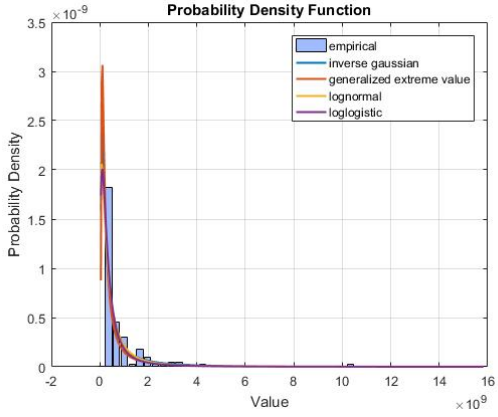


Figure 6.1.14: Boulder Fall

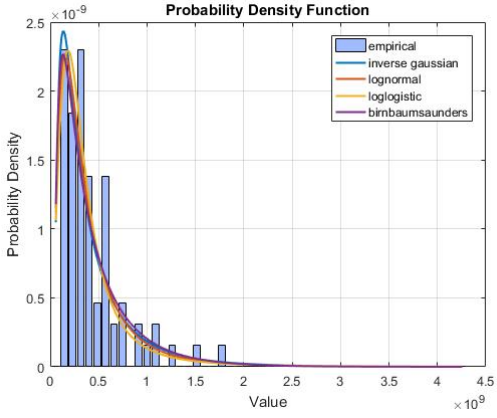


Figure 6.1.15: Fort Collins Winter

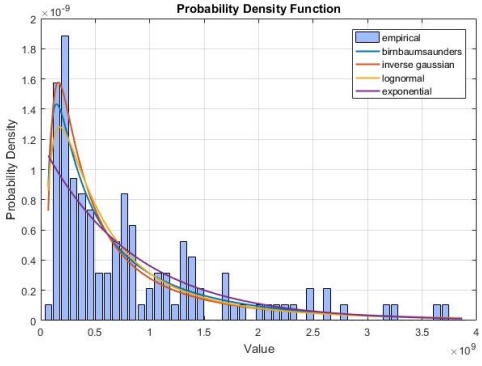


Figure 6.1.16: Fort Collins Spring

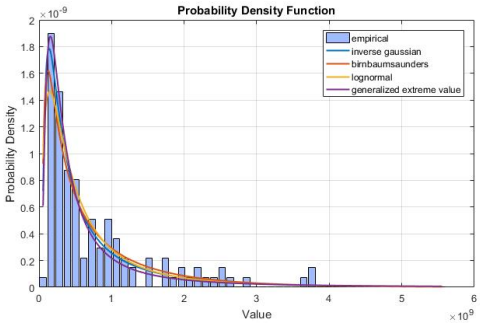


Figure 6.1.17: Fort Collins Summer

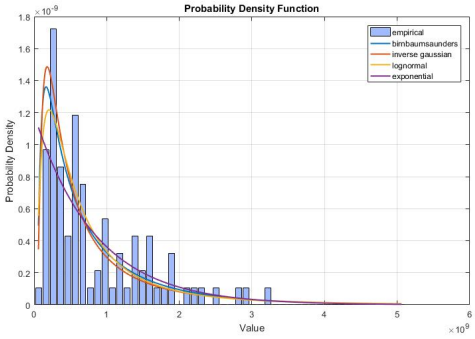


Figure 6.1.18: Fort Collins Fall

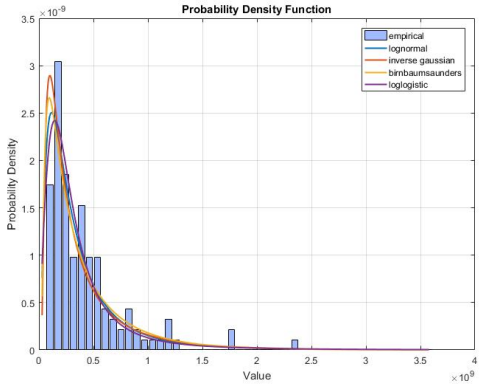


Figure 6.1.19: Lakewood Winter

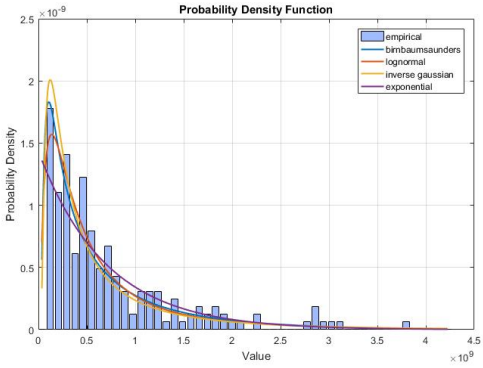


Figure 6.1.20: Lakewood Spring

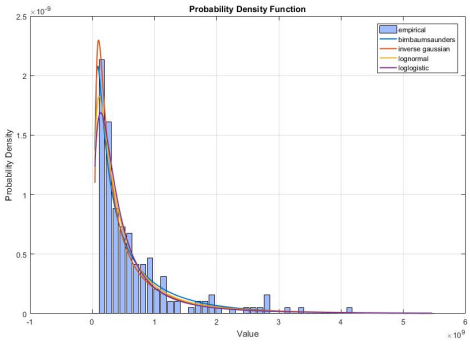


Figure 6.1.21: Lakewood Summer

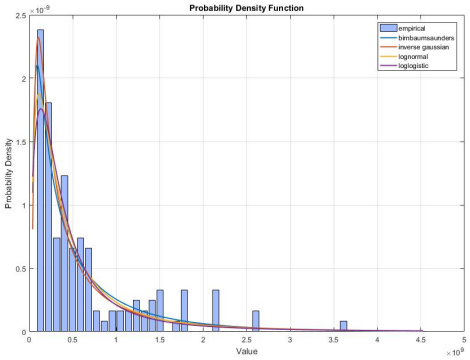


Figure 6.1.22: Lakewood Fall

6.2 Tables

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			101	165	286	508	892	**4260**
Inverse Gaussian	1597	(0.970)	106	168	298	545	933	0.996
Lognormal	1597	(0.995)	104	172	300	522	861	0.999
Loglogistic	1597	(0.996)	106	176	294	489	814	0.997
Birnbaum Saunders	1598	(0.822)	104	171	311	566	932	0.999

Table 6.2.1: Fort Collins Winter Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			120	201	570	1250	2100	**3870**
Birnbaum Saunders	2663	(0.644)	134	244	515	1090	1980	0.965
Inverse Gaussian	2664	(0.323)	138	237	469	993	1950	0.958
Lognormal	2667	(0.351)	136	255	514	1040	1950	0.958
Exponential	2673	(0.253)	89	243	585	1170	1950	0.975

Table 6.2.2: Fort Collins Spring Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			118	194	440	971	2060	**5560**
Inverse Gaussian	2629	(0.891)	120	207	417	904	1820	0.982
Birnbaum Saunders	2629	(0.789)	116	214	463	999	1850	0.974
Lognormal	2632	(0.576)	115	220	449	918	1750	0.974
Generalized Extreme Value	2633	(0.553)	124	205	391	849	2010	0.989

Table 6.2.3: Fort Collins Summer Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			145	241	546	1270	1850	**5060**
Birnbaum Saunders	1977	(0.945)	147	262	532	1080	1920	0.995
Inverse Gaussian	1978	(0.618)	152	255	493	1010	1910	0.989
Lognormal	1978	(0.821)	153	278	540	1050	1900	0.989
Exponential	1983	(0.408)	89	242	583	1170	1940	0.998

Table 6.2.4: Fort Collins Fall Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			60	114	251	516	815	**3420**
Birnbaum Saunders	2673	(0.968)	65	244	245	511	925	0.999
Inverse Gaussian	2674	(0.625)	67	237	224	470	915	0.997
Lognormal	2675	(0.795)	67	255	244	482	888	0.996
Loglogistic	2678	(0.540)	67	243	246	470	900	0.989

Table 6.2.5: Boulder Winter Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			112	206	451	998	1750	**4750**
Birnbaum Saunders	3843	(0.850)	110	202	432	922	1690	0.996
Lognormal	3846	(0.588)	117	220	444	897	1690	0.989
Inverse Gaussian	3847	(0.196)	115	197	394	844	1680	0.990
Exponential	3852	(0.308)	76	208	501	1000	1660	0.999

Table 6.2.6: Boulder Spring Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			91	161	348	683	1210	**3950**
Inverse Gaussian	3525	(0.968)	91	155	302	627	1210	0.995
Birnbaum Saunders	3526	(0.874)	89	159	329	679	1220	0.998
Lognormal	3528	(0.888)	90	164	321	628	1150	0.994
Loglogistic	3531	(0.694)	89	168	316	596	1120	0.988

Table 6.2.7: Boulder Summer Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			74	158	335	688	1480	**15800**
Inverse Gaussian	2625	(0.489)	79	144	322	812	1920	0.999
Generalized Extreme Value	2623	(0.953)	84	134	263	639	1790	0.987
Lognormal	2630	(0.252)	71	148	330	740	1530	0.999
Loglogistic	2631	(0.297)	68	143	303	641	1360	0.997

Table 6.2.8: Boulder Fall Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			92	139	256	483	858	**3580**
Lognormal	2652	(0.996)	82	143	265	489	850	0.998
Inverse Gaussian	2652	(0.911)	81	134	251	491	895	0.998
Birnbaum Saunders	2653	(0.934)	79	137	267	520	898	0.999
Loglogistic	2653	(0.942)	84	149	264	467	827	0.993

Table 6.2.9: Lakewood Winter Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max**/ max%
Observed			93	207	451	880	1680	**4220**
Birnbaum Saunders	4092	(0.868)	102	189	411	890	1650	0.993
Lognormal	4095	(0.755)	108	206	420	857	1630	0.986
Inverse Gaussian	4096	(0.169)	106	184	371	808	1630	0.986
Exponential	4102	(0.273)	74	201	484	968	1610	0.998

Table 6.2.10: Lakewood Spring Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			79	167	371	808	1640	**5460**
Birnbaum Saunders	3710	(0.978)	87	165	370	828	1570	0.998
Inverse Gaussian	3710	(0.717)	91	159	328	736	1530	0.994
Lognormal	3712	(0.889)	89	174	363	759	1470	0.993
Loglogistic	3717	(0.639)	88	178	360	728	1470	0.986

Table 6.2.11: Lakewood Summer Distributions

			Millions of Liters					
Model	**NLogL**	***p*-value**	10%	25%	50%	75%	90%	**max** / max%
Observed			89	156	355	738	1520	**4530**
Birnbaum Saunders	2832	(0.815)	88	165	360	788	1470	0.997
Inverse Gaussian	2833	(0.714)	91	158	322	708	1440	0.992
Lognormal	2836	(0.590)	89	170	352	727	1400	0.991
Loglogistic	2841	(0.001)	33	106	346	1130	3680	0.916

Table 6.2.12: Lakewood Fall Distributions