Analysis of Precipitation in the Northwestern United States

Matthew Wiens*, Kelly Kung

December 21, 2017

Abstract

Predicting precipitation is a key problem for residents of the Pacific Northwest, from its impact on family vacations to landslides. A noteworthy feature of the area is distinct chances of precipitation between the summer and winter due to the topographical differences. There have been previous research regarding precipitation differences in the Northwest, but there is room to build upon their results. In this report, we adopt a Bayesian methodology in order to obtain predictive results and propose two models for differentiating the two seasons and understanding the chance of rain across different climatic areas of the state. The first involves examining the start dates of each season and the second looks at the probability of rain on a given day. Using these models, we find that regions to the west of the Cascade Mountains have generally earlier and longer summers than in the eastern regions, but the eastern regions are generally drier. With these results, we can better understand the precipitation patterns of the areas and can predict optimal times to travel from a rainy region to a dry region.

1 Introduction

Residents and visitors to the Pacific Northwest (Oregon and Washington) notice a distinct pattern year over year: most days are rainy during the winter, and then they feel that there a distinct point during the spring or early summer when the weather becomes mostly dry. During early fall, the nice weather switches back to rain. However, the timing of the switch varies throughout the area, with the prominent Cascade Mountains being a key factor. There are many resorts on the east side of the Cascade Range that cater to residents of the west side looking to escape the rain.

Widmann and Bretherton[1] developed a methodology to control for local variation in topography and precipitation data and generated a dataset containing 46 years of estimated precipitation data in a 50km by 50km grid. In comparison, general models of atmospheric weather operate on the order of hundreds of kilometers, so there is significant room to improve the spatial resolution of weather models. In particular, local topographic features are not captured by general models, which is a limitation in regions like the Pacific Northwest. The temporal and spatial correlation is also highly dependent on the topography and are difficult to model, and therefore there has been a focus on parametrizing models and reducing the dimensionality.

In this report, we analyze the precipitation differences using a Bayesian approach for two reasons: first, to capture prior beliefs from living in the area of one author, and second to be able to discuss distributions around complicated model parameters and probabilistic beliefs about the state of weather on a specific day of the year. Bayesian approaches for daily precipitation data are not new, for example in Olson and Kleiber's work.[2] With a Bayesian approach, we hope to model (1) the beginning of dry and wet seasons and (2) the

*Lived in Seattle for 15 years

[1]M. Widmann and C. S. Bretherton. Validation of Mesoscale Precipitation in the NCEP Reanalysis Using a New Gridcell Dataset for the Northwestern United States. In Journal of Climate, 13(11), 1936-1950. 2000.

[2]B. Olson and W. Kleiber. Approximate Bayesian computation methods for daily spatiotemporal precipitation occurrence simulation,Water Resour. Res., 53, 3352?3372,doi:10.1002/2016WR019741. 2017.

probability of precipitation in the different areas. These models will provide us with tools in order to further analyze and understand the precipitation patterns in the Northwest.

This report is organized as follows. In Section 2, we refine our goals with our analysis, which motivates the creation of several variables in our dataset, which is also outlined in the same section. In Section 3, we discuss the two models that we used in order to analyze the data. We then begin our analysis of the dataset in Section 4, and in Section 5, we conclude our report with summaries of our analysis as well as potential future directions.

2 Background

2.1 Main Goals of Analysis

In their analysis, Widmann and Bretherton found that there was some variability in precipitation levels in the different topographical areas in the Northwestern United States. For this report, we further investigate the variability in the different areas in order to build upon their results. In particular, we are interested in analysis regarding when the dry and wet seasons start and the probability of rain on a given day. By using a Bayesian approach, we are able to calculate posterior probabilities in order to conduct further analysis and answer more questions. For example, which area has the longest summer? Which area is most likely to have the latest winter? What is the chance that someone can escape the bad weather by traveling to another part of the are? With our models and analysis, we may be able to better understand the precipitation patterns in the different topographical areas of the Pacific Northwest.

2.2 Precipitation Dataset

To analyze the precipitation in the Pacific Northwest, we use the dataset generated by Widmann and Bretherton. This data contains 46 years[3] of precipitation logs (measured in millimeters) from 1949 - 1994 that was collected form the National Weather Service and was corrected by the National Climatic Data Center[4][5]. The dataset is a 3-dimensional array with a 2-dimensional 16 (latitude) by 17 (longitude) grid cell[6] for each of the 16,801 days. However, there are approximately 32,000 missing values which correspond to areas with few functional weather stations, such as the ocean and some areas in the southeast. Using this dataset, we then transform and create several variables in order to obtain a model to analyze the precipitation.

2.3 Grouping of Areas

The Pacific Northwest has a varied climate, from temperate rainforests on the Pacific coast to desert in Southeastern Oregon, which has a major impact on the seasons. Therefore, each grid cell is classified into one of nine climatic zones as follows:

1. Washington Coastline (Temperate Rainforest)
2. Oregon Coastline
3. Western Washington / Seattle Metropolitan Area

[3]Note that when creating the indicators for years, we did not take into account of leap years, which resulted in data points that spilled over to the next year. We just take the years in which we have a year's worth of data for.

[4]The data was corrected during the 'Validated Historical Daily Data' project, which aimed to verify the data collected.

[5]T. Reek, S. Doty, and T. Owen. A Deterministic Approach to the Validation of Historical Daily Temperature and Precipitation Data from the Cooperative Network. In Bulletin of the American Meteorological Society 73, no. 6: 753-62. 1992.

[6]Each cell represents an area approximately 50km by 50km.

4. Portland Metropolitan Area / Willamette Valley / Southwestern Oregon
5. Cascade Range
6. Columbia Plateau
7. Selkirk Mountains
8. Blue Mountains
9. Eastern Oregon / High Desert / Great Basin

These zones are decided by considering average precipitation over the time period of the data and the topographic features of the region (Figure 1). In general, the regions correspond to areas with similar topography. For example, region 8, corresponds to the mountain range in northeastern Oregon, which extends into Idaho. However, only data from Oregon is considered. The rainfall for the region was aggregated by considering if there was measurable precipitation (more than 0.1 mm) in each grid by day, and then for each day, if the majority of grid cells reported precipitation, then a value of *rain* was assigned to the area, else it was *dry*.

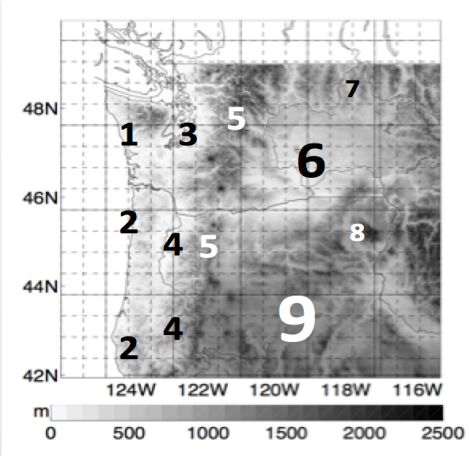

Figure 1: Topography of the Pacific Northwest, with the 9 areas indicated

2.4 Indicators for Dry and Wet Season

Using the area groups and daily precipitation indicators, mentioned in Section 2.3, we then proceeded to create indicators of when the dry season and wet season starts. To do this, we create an algorithm that chooses indicators of start days for the dry season (X_{min}) and wet season (X_{max}) such that it maximizes the number of *dry* days in the dry season and maximizes the number of *wet* days in the wet season, with weights determined by the overall prevalence of dry and wet days. I.e., if there are only a few dry days in the year, such as in region 1, the algorithm puts more weight on classifying the dry days correctly. For example, we expect X_{min} to be approximately 170 and X_{max} to be approximately 260 because these dates correspond to the beginning of summer and the beginning of autumn. Using this algorithm, we find the start dates for the dry and wet seasons for each region for the 46 years.

3 Modeling

3.1 Multivariate Normal Model

Our first model is motivated by the goal of determining when the dry and wet seasons start. We model each of the start dates using a semi-conjugate multivariate normal prior and multivariate normal likelihood. We choose this model because it has a closed form and we are able to sample from the posterior distribution using a Gibbs sampler on the full conditionals. Furthermore, upon initial exploration of the data, we see that the distribution of both the start dates of dry and wet seasons are approximately normal as seen in Figure 2. Thus, we continue with the multivariate normal likelihood with confidence that it represents the data we have.

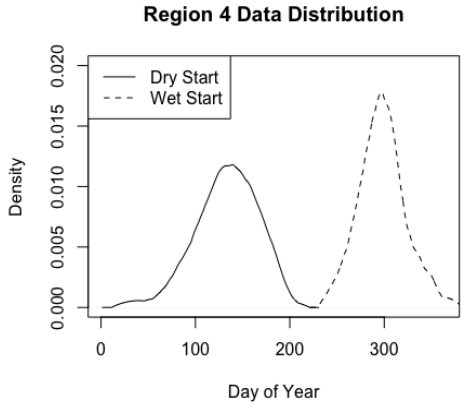

Figure 2: Kernel Density Estimate of the start and end of the dry season for Portland

We first define the semi-conjugate prior distributions as follows.

$$\theta \sim MVN(\mu_0, \Lambda_0)$$
$$\Sigma \sim Inverse - Wishart_{\nu_0}(\Sigma_0^{-1})$$

To determine the prior parameters, we let $\mu_0 = \begin{bmatrix} X_{min} \\ X_{max} \end{bmatrix} = \begin{bmatrix} 170 \\ 260 \end{bmatrix}$ and $\Lambda_0 = \begin{bmatrix} 30^2 & 0 \\ 0 & 30^2 \end{bmatrix}$ for all 9 area groups. The prior means are chosen according to when the summer and autumn season starts and the prior standard deviations of the $\theta's$ were chosen to represent a spread of one month. To set the prior parameter for Σ, we choose the Σ_0 values by mimicking the methods used during our homework, where we let Σ_0 equal the *variance - covariance* matrices for each area group from the data. The parameters for Σ_0 of the nine regions are shown in Table 1. We also let $\nu_0 = 2$ in order to have a relative non-informative prior.

In order to conduct Gibbs sampling, we need full conditionals of θ and Σ, which are shown below. By running Gibbs sampling on these full conditionals, we are effectively sampling from the posterior distribution

1	2	3	4	5
$\begin{bmatrix} 1154.93 & 179.97 \\ 179.97 & 618.95 \end{bmatrix}$	$\begin{bmatrix} 658.26 & 134.48 \\ 134.48 & 423.75 \end{bmatrix}$	$\begin{bmatrix} 1006.46 & 115.78 \\ 115.78 & 434.10 \end{bmatrix}$	$\begin{bmatrix} 1000.96 & 87.99 \\ 87.99 & 575.44 \end{bmatrix}$	$\begin{bmatrix} 993.37 & 236.99 \\ 236.99 & 405.11 \end{bmatrix}$

6	7	8	9
$\begin{bmatrix} 2943.59 & 630.29 \\ 630.29 & 728.15 \end{bmatrix}$	$\begin{bmatrix} 3574.88 & 884.72 \\ 884.72 & 874.46 \end{bmatrix}$	$\begin{bmatrix} 2221.71 & 210.85 \\ 210.85 & 500.14 \end{bmatrix}$	$\begin{bmatrix} 1974.74 & 250.49 \\ 250.49 & 684.75 \end{bmatrix}$

Table 1: Table of prior Σ_0 values for each area group

and so we can obtain posterior estimates of θ and Σ.

$$\theta|\Sigma, y \sim MVN(\mu_n(\Sigma), \Lambda_n(\Sigma)))$$
$$\Sigma|\theta, y \sim Inverse - Wishart_{\nu_n}(\Sigma_n^{-1}(\theta))$$

$$\mu_n(\Sigma) = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$
$$\Lambda_n^{-1}(\Sigma) = \Lambda_0^{-1} + n\Sigma^{-1}$$
$$\nu_n = \nu_0 + n$$
$$\Sigma_n(\theta) = \Sigma_0 + \sum_{i=1}^{n}(y_i - \theta)(y_i - \theta)^T$$

3.2 Daily Precipitation Chance Model

We further develop our analysis by considering the probability of precipitation on each day for the summer (dry) and winter (rainy) seasons. This is motivated by the idea of trying to plan outdoor activities in each area of the Pacific Northwest and deciding if there is value in traveling to a different area to escape rain. Therefore, the following model is proposed based on the observed V shape of the probability of precipitation by day as in Figure 3:

- The data is binary data representing if it rained or not for each day of the year
- The probability of rain on the jth day of the year is given by the following piecewise linear function: $\pi_j = \theta_0 + (a * 1_{j<\theta^*} + b * 1_{j>=\theta^*})(j - \theta^*)$
- θ_0 can be interpreted as the chance of precipitation on the driest day of the year, θ^* is the day number of the driest day, and a and b are the slopes of the springtime daily change in precipitation chance, and fall daily change in precipitation chance, respectively
- Observing precipitation on the jth day of the year is a Bernoulli random variable, with parameter π_j

Furthermore, we note the wintertime (January - March) is has a constant, large chance of rain across the region, as seen in Figure 3, and therefore we choose to focus our analysis on the periods where this chance is changing. We use weakly informative priors in this model: $\theta_0 \sim N(210, 900)$,$\theta^* \sim \beta(5, 20)$, $b \sim -a \sim N(.004, .002^2)$. These priors capture the prior beliefs that the driest days are around the end of July, and we are very sure they are within two months on either side, and that the driest days have about a 20% chance of rain. The slopes follow from noting that the rainiest part of the year has about a 80% chance of rain. We choose a slope that appropriately interpolates the wettest and driest portions of the year, and we choose a variance such that there is only a small chance that the sign of each portion of the slope is contrary to our expectations. The likelihood and full conditionals of the parameters for this model exist in closed form, however they are not informative, so we proceed to use a Metropolis MCMC algorithm to produce posterior distributions.

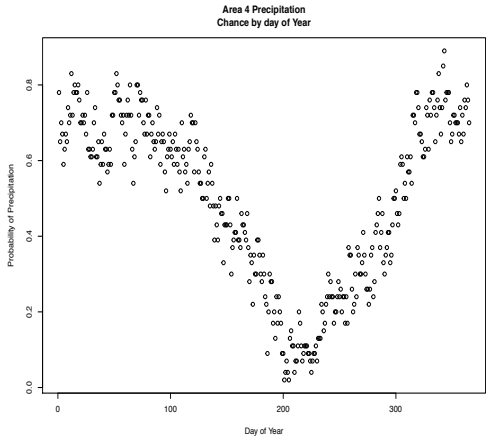

Figure 3: Western Oregon precipitation chance over the year

4 Analysis

4.1 Multivariate Normal Model

Using the semi-conjugate multivariate normal model described in section 3.1, we used a Gibbs sampler to generate samples from the posterior distributions of θ and Σ as well as the posterior predictive distribution $\begin{bmatrix} X_{min} \\ X_{max} \end{bmatrix} \sim MVN(\theta, \Sigma)$. When drawing samples from the posterior predictive distribution, we used the restriction: $(X_{min}, X_{max})\epsilon(0, 365)$ to ensure that the sampled data points represents dates in a year. Because we used Gibbs sampler, we needed to take into account of the dependencies of the Markov Chains. This is important because issues may arise if the chain is stuck in one area of the distribution because of the previous sample. To account for these dependencies, we ran 20,000 simulations in order to obtain high effective sizes. On average, the effective sizes were: $\theta_{X_{min}} : 10,795, \theta_{X_{max}} : 7,229, \Sigma_{X_{min}} : 11,286, \Sigma_{X_{max}} : 8,906$. We then examined the autocorrelation plots and saw that it was fairly high, especially for the lag periods of 1 and 2 iterations. To resolve this issue, we thinned out the data by taking every tenth entry which resulted in 2,000 sampled values. Figure 4 shows the autocorrelation plots before and after thinning the data. We then analyzed the data using this thinned out dataset.

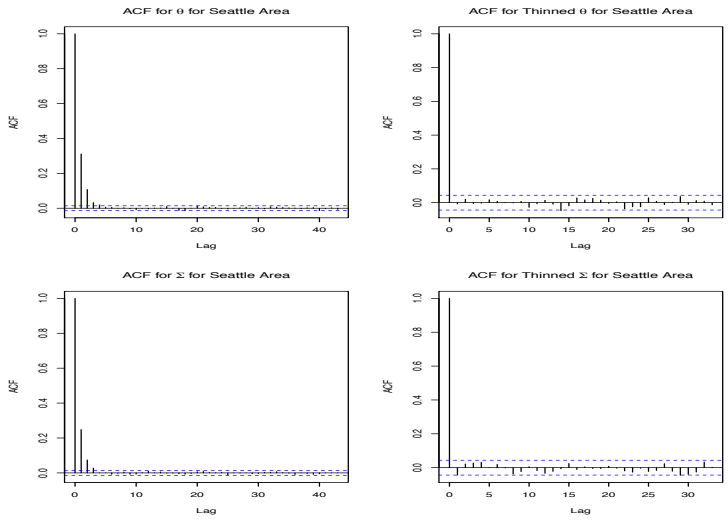

Figure 4: Autocorrelation plots for parameters before and after thinning data

We report summary statistics of θ, the posterior mean of the start date of the dry and west seasons, in Table 2. In particular we note the variability in the mean start dates of the dry season between the regions, which ranges from 149 - 162, and the increase in the mean start date as we move towards the eastern regions. On the other hand, the end date of the dry season is much more consistent across regions, ranging from 280 to 289.

Region	Start Date Mean	Start Date Variance	End Date Mean	End Date Variance
1	152	24	280	17
2	149	20	288	18
3	153	22	284	15
4	154	23	282	22
5	153	17	289	12
6	158	30	283	27
7	158	28	284	28
8	162	25	289	20
9	160	26	284	30

Table 2: Posterior Mean and Variance of θ

Using the predictive posterior sample, we also generated probabilities that an area has the earliest dry season and latest wet season in a given year. The probabilities are shown in Table 3. Looking at the probabilities, we see that the region 6 and 7 (Columbia Plateau and Selkirk Mountains) are most likely to have the earliest dry season and latest wet season. This aligns with the climate of the area since the Columbia Plateau is somewhat arid because of the rain shadow of the Cascade Range and the Selkirk Mountains are located in the generally drier, East. Since the two regions are relatively close to each other, it makes sense that they share the same precipitation patterns. However, neighboring regions, such as region 5, the Cascade Range, have fairly low probabilities of having the earliest dry season. This suggests that the precipitation patterns do vary across the Northwest region, and it is likely due to the topographical makeup of each area. This is because although the Cascade Range and Columbia Plateau are only approximately 320 km apart, the Cascade Range does not share similar patterns as the Columbia Plateau and Selkirk Mountains which suggests that there is another underlying reason for the differences in precipitation patterns.

Looking at the lengths of the dry season in Table 3, there is quite some variability, with a difference of 20 days of the longest and shortest periods. The Oregon Coastline (region 2) and Cascade Range (region 5) have the longest dry season while region 9 (Eastern Oregon border area) has the shortest dry season. Although

this is not as we first expected, it makes sense according to how we defined the start dates and the climate of the regions. In the western regions, we expected a clear distinction between the dry and wet seasons whereas the eastern regions are generally drier and may not have a clear distinction of the two seasons. Therefore, we are most likely distinguishing between *very dry* days from the normally *dry* days in areas such the Columbia Plateau, Selkirk Mountains, and the desert/basin area. Thus, we are essentially comparing the lengths of very dry seasons in the eastern regions and dry seasons in the western regions, which may be an explanation of the results. However, the results still lead us to believe that there are varying precipitation patterns across the regions.

Region	Earliest Dry Season	Latest Wet Season	Length of Dry Season
1	0.099	0.074	129
2	0.074	0.032	138
3	0.091	0.057	122
4	0.087	0.081	127
5	0.073	0.068	136
6	0.180	0.198	123
7	0.179	0.200	121
8	0.109	0.145	129
9	0.110	0.147	119

Table 3: Area with probabilities of earliest and latest start times as well as length of dry season

4.2 Daily Precipitation Chance Model

The model described in section 3.2 was implemented with a Metropolis MCMC algorithm for each of the nine areas, with a thinning by a factor of 50 to reduce autocorrelation. Each of the nine areas was again treated independently, because at the scale of which we defined the regions, there is minimal correlation in daily precipitation. As expected, the driest day of the year in all the regions is in early August. All the regions on the west side of the Cascade Range had the driest day around day number 220, while those east of the mountains had the driest day around day number 230, which is analogous to the results from the first model we explored and reported in Table 2. These posterior distributions are reported in Figure 5, and it highlights the clear difference between the two geographical areas. However, the chance of rain on the driest day of the year varied significantly more. In region 1, the temperate rainforest and coast, the minimum chance was 25%, while in region 9, which contains part of the Great Basin and other steppe climates, the minimum chance was 6%. The estimated chance of rain for each day and each region is reported in Figure 6.

Posterior estimates of the mean of the slopes are all between 0.003 and 0.006 for the fall, and the negative, but the same range for the springtime. The natural interpretation is that for each day that passes, the probability of rain changes by 0.3% to 0.6%. However, there is little correlation of the means (-0.07) between how dry the driest day is, and how rapidly the chance of precipitation changes. We suspect that this is due to the variety of topographical features that obscure any correlation.

Furthermore, we use this model to analyze the possibility of leaving the wet Seattle (region 3) climate for the drier climate to the east of the Cascade Mountains (region 6). These results are found in Figure 7, which maintains the V shape of the model. We find that during the summer, the chance of escaping Seattle rain is high if it rains, but such a need is rare because Seattle is dry as well in the summer. However, if one takes the perspective of a resort manager in the drier region, outside of the summertime, the chance of an influx of business from Seattle residents searching for drier weather is approximately constant. There is a balance between the higher demand from Seattle because of the increased chance of precipitation in the west and the leaving of the clientele because of the increased chance of precipitation in the drier climate as well.

This piecewise linear model gives additional insight into how the probability of precipitation changes throughout the year. It confirms the conclusion of the multivariate normal model outline in section 3.1, and

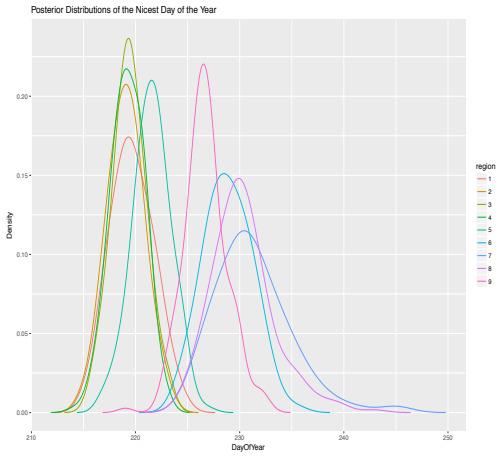

Figure 5: Distributions of the day with minimum chance of rain

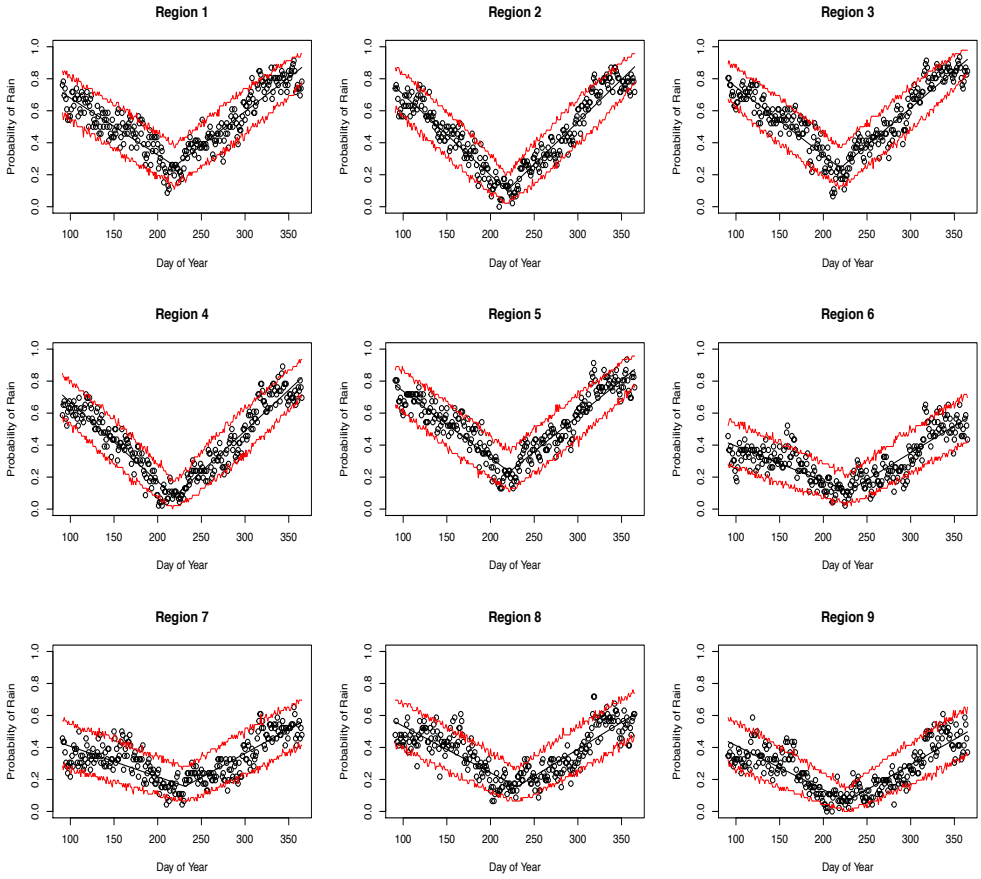

Figure 6: Data Points for each region, with mean estimate and 95% credible interval for the observed chance of rain over 46 years. In most regions, approximately 95% of the data is inside the credible interval, indicating a valid model

shows how precipitation data of this sort yields the conclusions of the simpler model.

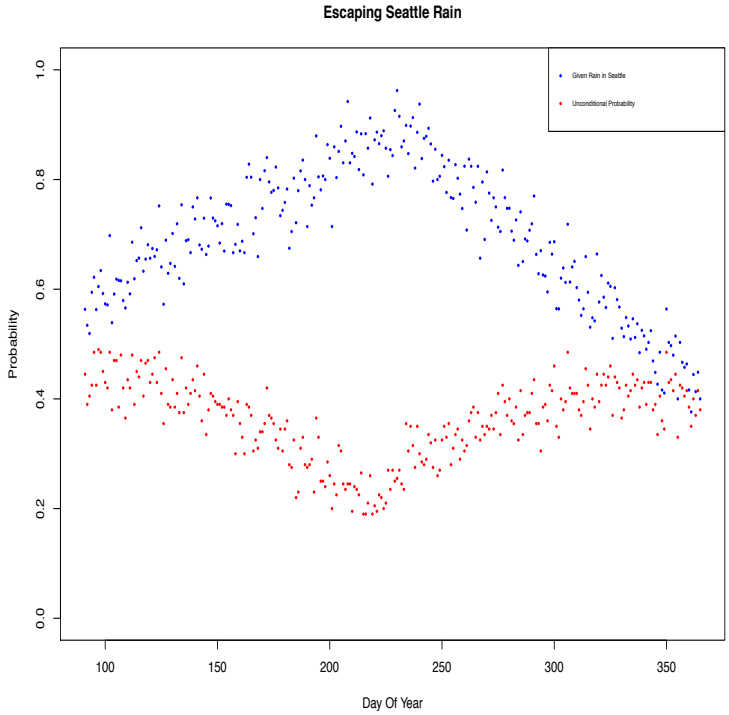

Figure 7: Daily chance of escaping precipitation in Seattle. In blue, is the probability that that Region 6 will be dry given that Region 3 rains. In red, is the probability that Region 3 is rainy and Region 6 is dry

5 Conclusion

In this report, we analyzed the precipitation in the Pacific Northwest by diving the area into 9 distinct regions and using 46 years of data with 30km by 30km resolution to approximate the region-wide precipitation. Two models were proposed to understand how the chance of precipitation changes over the year. In the first, we divide the year into a rainy season and a dry season and model the changeover point as a proxy for residents' experience with the weather patterns. Based on observed data, we proposed a semi-conjugate multivariable normal model for the changeover points. There was some correlation within regions, but assumed independence between regions. Subtle correlation between the regions and within smaller subregions may be an avenue for further exploration with this data set. Within this model, we also explored how the duration of the dry season varied across the regions. The Oregon Coast and Cascade mountains had the longest dry season, and we suggested that this is due to the large difference between the winter and summer in precipitation chance, so even moderately dry weather is classified as the dry season. The second model we proposed seeks to understand how the probability of precipitation changes throughout the year on a daily level, as opposed to describing broad yearly climatic patterns. This model proposes a constant rate of decrease in the probability of precipitation during the springtime, and also a constant rate of increase during autumn, with some minimum chance of rain in the summer. In this model, we found that the driest days of the year are earlier on western regions of the Pacific Northwest, however the eastern regions are drier overall, and also have a much drier summer. We also examined the opportunity for Seattle residents to escape rainy weather by traveling a short distance to the Columbia Plateau, and the model validates the tourism industry in the Columbia Plateau targeting Seattle residents.

These models could be extended to include local measures of correlation and to find in which areas is precipitation locally correlated and which neighboring areas have low correlation. For example, in the 100 miles east of Seattle, there are 3 distinct climatic areas, which likely have low correlation, however 100 miles of coastline may be strongly correlated. Further investigation is warranted into the weather patterns that

cause all the regions east of the Cascade range to have a later driest day, and relatively drier weather later into the fall.

Appendix

Link to our Git Repository

https://github.com/kkung111/MA578_Project.git

R Code for the Two Models

```
#-----------------------------#
# Load in the data            #
#-----------------------------#
library(ncdf4)
weather<-nc_open("pnwrain.50km.daily.4994.nc")
#details about the dataset
print(weather)
weatherDat<-ncvar_get(weather, attributes(weather$var)$names[1])
weatherDat
dim(weatherDat) #latitude (17), longitude (16), time (16801 in days since 1949)
#each cross section shows the mm/day amount of rainfall
#32767 missing data points

# Define the 9 regions
finer_grids <-
  c( 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
     0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
     0,0,0,0,0,3,5,5,5,6,6,6,6,7,7,7,7,
     0,0,0,0,0,3,5,5,6,6,6,6,6,7,7,7,7,
     0,0,1,1,3,3,5,5,6,6,6,6,6,6,6,6,6,
     0,0,0,3,3,3,5,5,6,6,6,6,6,6,6,6,6,
     0,0,0,3,3,3,5,5,6,6,6,6,6,6,6,6,6,
     0,0,0,3,3,3,5,5,6,6,6,6,6,6,6,6,6,
     0,0,0,2,3,3,5,6,6,6,6,6,8,8,8,8,8,
     0,0,0,2,3,3,5,9,9,9,9,9,8,8,8,8,8,
     0,0,0,2,4,4,5,9,9,9,9,9,8,8,8,8,8,
     0,0,0,2,4,4,5,9,9,9,9,9,9,9,9,9,9,
     0,0,0,2,4,4,5,9,9,9,9,9,9,9,9,9,9,
     0,0,0,2,4,4,9,9,9,9,9,9,9,9,9,9,9,
     0,0,2,2,4,4,9,9,9,9,9,9,9,9,9,9,9,
     0,0,2,2,4,4,9,9,9,9,9,9,9,9,0,0,0,
     0,0,0,2,4,4,9,9,9,9,9,9,0,0,0,0,0
  )
grid2 <- t(matrix(finer_grids,ncol=17,byrow=T)[,-17])

#
#-------------------------------------------------------------#
# Model 3.1                                                   #
# Multivariate Normal Gibbs Sampler Model                     #
#-------------------------------------------------------------#

library(MCMCpack)
library(mvtnorm)
```

```r
#hold our values for each area
output_th <- list(rep(0,9))
output_s2 <- list(rep(0,9))
output_yt <- list(rep(0,9))

#helper functions from class
rmv<-function(n,mu,Sigma){    # samples Y~MVN(mu,Sigma)
  cm<-chol(Sigma);d<-dim(Sigma)[1]
  Y0<-matrix(rnorm(n*d),nrow=d)
  t(cm)%*%Y0 + mu
}
riw<-function(n,nu0,Sm){ # Sigma~IW(nu0,Sigma^(-1)); requires rmv
  m<- solve(Sm)
  sapply(1:n,function(i)
    solve(crossprod(t(rmv(nu0*n,0,m))[(i-1)*nu0+1:nu0,])), simplify = 'array')
}

for (k in seq(1,9)) {
precipData<-as.matrix(read.table(paste("Area",k,"MinMaxDataV2.tsv"), sep = "\t"))

precipDataCov<-cov(precipData)

#find the prior means
precipDataMinMu0<-170
precipDataMaxMu0<-260

#store them into a matrix
precipDataMu0<-matrix(c(precipDataMinMu0, precipDataMaxMu0), nrow = 2)


#prior standard deviations of theta
precipDataMins20<-30
precipDataMaxs20<-30

#prior Sigma0
precipDatas20<-matrix(c(precipDataMins20, 0, 0, precipDataMaxs20), nrow = 2)


#Gibbs Sampler

#Starting values
ybW<-apply(precipData, 2, mean)
nW<-dim(precipData)[1]
nu0<-2 #note: had to change this because was giving me errors with 1
nun<-nu0 + nW

set.seed(1234+k)
nSim<-20000

SigmaW<-riwish(nu0, precipDataCov)
thetaW<-rmvnorm(1, precipDataMu0, precipDatas20)
```

```r
#initiate the matrices to hold our values
THW<-S2W<-YtW<-NULL

for(i in 1:(nSim+1000)){
  #sample Sigma
  LnW<-precipDataCov + crossprod(precipData - outer(rep(1, nrow(precipData)), c(thetaW)))
  SigmaW<-riw(1, nun, LnW)[,,1]


  #sample theta
  LN<-solve(solve(precipDatas20) + nW *solve(SigmaW))
  munW<-LN%*%(solve(precipDatas20)%*%precipDataMu0 + nW*solve(SigmaW)%*%ybW)
  thetaW<-rmv(1, munW,LN)


  #prediction
  #we didn't take into account of the first 1000 simulations during the "warm up" period"
  if(i > 1000) {
  yPred<-rmv(1, thetaW, SigmaW)
  #restrict predictions to be between 0 and 365
  while(yPred[1]<0 | yPred[2]>365){yPred<-rmv(1, thetaW, SigmaW)}
  ytW<-yPred
  THW<-cbind(THW, thetaW)
  S2W<-cbind(S2W, c(SigmaW))
  YtW<-cbind(YtW, ytW)
  }

}

rowMeans(THW)
rowMeans(S2W)
rowMeans(YtW)

output_th[[k]] <- t(THW)
output_s2[[k]] <- t(S2W)
output_yt[[k]] <- t(YtW)
}


#assumption checking:
#plot Seattle Data
plot(density(as.matrix(read.table(paste("Area",4,"MinMaxDataV2.tsv"), sep = "\t"))[,1],
     kernel = "epanechnikov"),lty = 1,xlim=c(1,365),main = "Region 4 Data Distribution",
     xlab = "Day of Year",ylim = c(0,1/50))
points(density(as.matrix(read.table(paste("Area",4,"MinMaxDataV2.tsv"), sep = "\t"))[,2],
     kernel = "epanechnikov"),lty = 2,type="l")
legend("topleft", c("Dry Start","Wet Start"),lty = c(1,2))


#Model 3.1 Analysis


#thin out the data and keep copies of the original data
```

```
seqLength<-seq(1, dim(output_yt[[1]])[1], 10)
tempoutput_yt<-output_yt
tempoutput_th<-output_th
tempoutput_s2<-output_s2

output_yt<-lapply(output_yt, function(x){x[seqLength,]})
output_th<-lapply(output_th, function(x){x[seqLength,]})
output_s2<-lapply(output_s2, function(x){x[seqLength,]})

#find the effectice Sizes and acf
#effective size calculation
effSizeth<-matrix(unlist(lapply(tempoutput_th, function(x){effectiveSize(x)})), ncol = 2, byrow = T)
effSizes2<-matrix(unlist(lapply(tempoutput_s2, function(x){effectiveSize(x)})), ncol = 2, byrow = T)

colMeans(effSizeth)
colMeans(effSizes2)


#plot some of the acf plots
par(mfrow=(c(2,2)))
acf(tempoutput_th[[3]][,1], main = expression(paste("ACF for ", theta, " for Seattle Area")))
acf(output_th[[3]][,1], main = expression(paste("ACF for Thinned ", theta, " for Seattle Area")))
acf(tempoutput_s2[[3]][,1], main = expression(paste("ACF for ", Sigma, " for Seattle Area")))
acf(output_s2[[3]][,1], main = expression(paste("ACF for Thinned ", Sigma, " for Seattle Area")))

# Theta Means and Variance
round(matrix(unlist(lapply(output_th,function(x) {apply(x,2,mean)})),ncol=2,byrow=T))
round(matrix(unlist(lapply(output_th,function(x) {apply(x,2,var)})),ncol=2,byrow=T))

# First/Last Dates for each area
start_dates <- matrix( unlist(lapply(output_yt,function(x) {x[,1]} )),
              nrow = length(seqLength),ncol=9,byrow=F)
table(apply(start_dates,1,which.min))/length(seqLength)

end_dates <- matrix( unlist(lapply(output_yt,function(x) {x[,2]} )),
             nrow = length(seqLength),ncol=9,byrow=F)
table(apply(start_dates,1,which.max))/length(seqLength)

# Length of summer
summer_length <- matrix( unlist(lapply(output_yt,function(x) {x[,2]-x[,1]} )),
                 nrow = length(seqLength),ncol=9,byrow=F)
round(apply(summer_length,2,mean))
round(sqrt(apply(summer_length,2,var)))

#-------------------------------------------------------------#
# Model 3.2                                                   #
# Implemented using Metropolis for each of the 9 areas        #
#-------------------------------------------------------------#

# Piecewise Linear Metropolis

# Metropolis for refined model
library(coda)
```

```
#tf_weather_data2 <- weatherDat >= .5
# see code in create_9part_dataset.R to manipulate this into a list of lists.
# this uses an intermediate result - i.e. a list of 9 [ list of 46 [vector of 275 T/F]]
# Parameterize it by mindate,minval, a, b
drop_days <-90
mindate_all <- list(rep(0,9))
minval_all<-list(rep(0,9))
a_all <- list(rep(0,9))
b_all <- list(rep(0,9))
obsrain_all <- list(rep(0,9))
prain_all <- list(rep(0,9))
posterior_means <- matrix(nrow=9,ncol=4)
for (l in seq(1,9)){
  data_met <- tf_data2[[l]]
  year_aggregates <-  apply(matrix(as.numeric(unlist(data_met)),ncol=365,byrow=T),2,sum)
                [(drop_days+1):365]
  #priors
  mindate_mean <- mindate <- 210-drop_days
  mindate_sd <- 30
  minval_a  <- 5
  minval_b <- 20
  a_mean <-  -.6/150
  a_sd <- abs(a_mean)/2
  b_mean <- b <- .6/150
  b_sd <- abs(b_mean)/2
  minval <- minval_a / (minval_a + minval_b)

  S <- 10000
  MINDATE <- MINVAL <- A <- B <- rep(0,S)
  OBSRAIN <- PRAIN <- matrix(ncol=365-drop_days,nrow=S)
  accept_probs <- rep(0,S+500)

  delta <- .095
  llike <- function(y,theta_min,theta_val,a,b){
    # takes in the list of 365 data points
    after_min <- c(rep(0,floor(theta_min)),rep(1,365-drop_days-floor(theta_min)))
    j <- seq(1,(365-drop_days))
    prob <- theta_val + (j-theta_min)*(a*(1- after_min)+ b*after_min)
    prob <- sapply(prob,function(x) {max(min(x,1),0)})
    llike <- sum(dbinom(y,46,prob,log=T))
    return(llike)
  }

  prob_by_day <- function(theta_min,theta_val,a,b)
  {
    after_min <- c(rep(0,floor(theta_min)),rep(1,365-drop_days-floor(theta_min)))
    j <- seq(1,(365-drop_days))
    prob <- theta_val + (j-theta_min)*(a*(1- after_min)+ b*after_min)
    prob <- sapply(prob,function(x) {max(min(x,1),0)})
    return(prob)
  }

  accept <- 0
```

```
for( i in seq(1,(S+500))){ # 500 warm up iterations
  mindate.star <- rnorm(1,mindate,mindate_sd*delta)
  minval.star <- rnorm(1,minval,sqrt(minval_a*minval_b/((minval_a+minval_b)^2*
                (minval_a+minval_b+1)))*delta)
  a.star <- rnorm(1,a,delta*a_sd)
  b.star <- rnorm(1,b,delta*b_sd)

  log.r <- llike(year_aggregates,mindate.star,minval.star,a.star,b.star)
         -llike(year_aggregates,mindate,minval,a,b) +
         dnorm(a.star,a_mean,a_sd,log=T) + dbeta(minval.star,minval_a,minval_b,log=T) +
         dnorm(a.star,a_mean,a_sd,log=T) + dnorm(b.star,b_mean,b_sd,log=T) -
         dnorm(a,a_mean,a_sd,log=T) - dbeta(minval,minval_a,minval_b,log=T) -
         dnorm(a,a_mean,a_sd,log=T) - dnorm(b,b_mean,b_sd,log=T)
  accept_probs[i] <- exp(log.r)
  if(log(runif(1))<log.r){
    if(i > 500) accept <- accept+1
    a <- a.star
    b <- b.star
    minval <- minval.star
    mindate <- mindate.star
  }
  output_filter <- 50
  if((i > 500) ) {
    A[i-500] <- a
    B[i-500] <- b
    MINDATE[i-500] <- mindate
    MINVAL[i-500] <- minval
    PRAIN[i-500,] <- prob_by_day(mindate,minval,a,b)
    OBSRAIN[i-500,] <- rbinom(365-drop_days,46,prob_by_day(mindate,minval,a,b))/46
  }
}
MINDATE <- MINDATE[seq(1,S,by=50)]
MINVAL <- MINVAL[seq(1,S,by=50)]
A <- A[seq(1,S,by=50)]
B <- B[seq(1,S,by=50)]
OBSRAIN <- OBSRAIN[seq(1,S,by=50),]
PRAIN <- PRAIN[seq(1,S,by=50),]

sum(accept)/S
c(effectiveSize(MINDATE),effectiveSize(MINVAL),effectiveSize(A),effectiveSize(B))
c(mean(MINDATE),mean(MINVAL),mean(A),mean(B))
c(var(MINDATE),var(MINVAL),var(A),var(B))

plot(density(MINDATE+drop_days,adjust = 1.5),xlab = "Day of Year", m
     ain = "Distribution of Nicest day of the Year")
plot(seq(drop_days+1,365),year_aggregates/46,ylab = "Probability of Rain",
     xlab = "Day of Year",main = "Probability of Rain by Day\nWith 95% Credible
     Interval")
points(seq(drop_days+1,365),prob_by_day(mean(MINDATE),mean(MINVAL),mean(A),mean(B)),
       type="l")
points(seq(drop_days+1,365),apply(OBSRAIN,2,quantile,.975),type="l",col="red")
points(seq(drop_days+1,365),apply(OBSRAIN,2,quantile,.025),type="l",col="red")
```

```
  mindate_all[[l]] <- MINDATE
  minval_all[[l]] <- MINVAL
  a_all[[l]] <- A
  b_all[[l]] <- B
  obsrain_all[[l]] <- OBSRAIN
  prain_all[[l]] <- PRAIN
  posterior_means[l,] <- c(mean(MINDATE),mean(MINVAL),mean(A),mean(B))
}

# Model 3.2 Analysis

# Plot the distributions of the means
library(ggplot2)
plt_df <- data.frame(region = rep("1",512),DayOfYear =
          density(mindate_all[[1]]+drop_days,adjust = 1.5)$x,Density =
          density(mindate_all[[1]]+drop_days,adjust = 1.5)$y)
for(l in seq(2,9)){
  plt_df <- rbind(plt_df,data.frame(region = rep(paste(l),512),DayOfYear =
            density(mindate_all[[l]]+drop_days,adjust = 1.5)$x,Density =
            density(mindate_all[[l]]+drop_days,adjust = 1.5)$y))
}

ggplot(data=plt_df, aes(x=DayOfYear, y=Density, group=region) )
      +geom_line(aes(color=region)) +
      ggtitle("Posterior Distributions of the Nicest Day of the Year")

# Credible Intervals
par(mfrow = c(3,3))
for (l in 1:9) {
  plot(seq(drop_days+1,365),apply(matrix(as.numeric(unlist(tf_data2[[l]])),ncol=365,
      byrow=T),2,sum)[(drop_days+1):365]/46,ylab = "Probability of Rain",xlab =
      "Day of Year",main = paste("Region",l))
  points(seq(drop_days+1,365),prob_by_day(mean(mindate_all[[l]]),mean(minval_all[[l]]),
      mean(a_all[[l]]),mean(b_all[[l]])),type="l")
  points(seq(drop_days+1,365),apply(obsrain_all[[l]],2,quantile,.975),type="l",col="red")
  points(seq(drop_days+1,365),apply(obsrain_all[[l]],2,quantile,.025),type="l",col="red")
}
par(mfrow = c(1,1))

#Escape Seattle Rain
escape_p <- seq(1,365-drop_days)
num_escape <- seq(1,365-drop_days)
for( i in seq(1,365-drop_days)){
  searain <- rbinom(length(prain_all[[3]][,i]),1,prain_all[[3]][,i])
  eastdry <- rbinom(length(prain_all[[3]][,i]),1,1-prain_all[[6]][,i])
  escape_p[i] <-  sum(searain*eastdry)/sum(searain)
  num_escape[i] <- sum(searain*eastdry)/length(searain*eastdry)
}
plot(seq(drop_days+1,365),escape_p,ylim = c(0,1),col="blue",cex=.5,
    xlab = "Day Of Year",ylab = "Probability",main = "Escaping Seattle Rain",pch=16)
points(seq(drop_days+1,365),num_escape,col="red",cex=.5,pch=16)
legend("topright",legend = c("Given Rain in Seattle","Unconditional Probability"),
       col = c("blue","red"),cex = .5,pch=c(16,16))
```