Analysis of Climate Change and Flood Risk. Opportunities for Insurance Companies, Local Governments, and Individuals

Team #7780

April 14th, 2021

Contents

1 Executive Summary

Floods have many points of origin and are chronic due to climate change and sea-level rise. Many different variables can affect the severity of a flood. Understanding these variables and their impact is vital to understanding flooding. Like the National Flood Insurance Program(NFIP) and the Federal Emergency Management Agency(FEMA), government organizations attempt to understand the origins of floods and how to mitigate their losses. Our task is to analyze and model flooding to give recommendations on mitigating its losses for the states of North Carolina and Missouri.

We researched historical climate data from The National Oceanic and Atmospheric Administration (NOAA). We also garnered historical claims due to flooding and Flood Hazard Maps from FEMA. To estimate the flood probability, we use the occurrence of floods, climate factors, and data from the Flood Hazard Maps. Our climate factors consist of minimum temperature, maximum temperature, snow, and precipitation, all from NOAA's databases. The Flood Hazard Maps indicates data such as the number of levees, number of dams, number of channels, and Special Flood Hazards Areas(SFHA). SFHA are zones where the flood probability is higher than one percent.

We developed a logistic regression model to estimate the flood probability for a particular county in the next year. The model provides information on how temperature and precipitation data relate to flood probability. We also built a Knn-means regression model to estimate the Expected Percentage Payment (EPP). EPP is the ratio of the paid claim amount to the insurance coverage. We then use the output of the model to estimate the expected claim payment or flood loss.

After analyzing the results, the highest flood probability counties are in the north-western region and the south-eastern region for North Carolina and Missouri respectively. North Carolina's most impacting variables are the number of SFHA and precipitation change in the spring lag one year, while Missouri's are the number of SFHA and precipitation change in spring lag two years. Geographical risk groups are available from this information. SFHA form an issue in terms of misinformation to the individuals in these zones. Individuals outside the SFHA believe they are not at risk. Individuals near the outer edge of the SFHA are in a risk group as well. SFHA is expanding because of climate change, so this risk group is also increasing. Other risk groups include individuals not being able to afford insurance.

Recommendations for this issue can be a variety of items. One of the most important ones is properly informing possible policyholders. If we can adequately inform all policyholders, more of them will receive coverage, and the insurers will increase penetration. Another approach is to bundle flood into homeowner policies. This approach may be the most efficient economic method to cover the risk for the insurer. Insurers can encourage investments in cost-effective loss reduction measures in return for premium discounts. One of these implications could be between local governments and insurers. Insurers can provide products in severely adverse local government situations if insurers encourage investments in cost-effective loss reduction measures.

2 Background Information

There are few places on Earth where flooding is not a concern. Any region where there is precipitation is susceptible to floods. Floods are the most common natural disaster in the United States. Each year billions of dollars are lost due to flooding across the United States. In 2019 alone, $3.75 billion was lost. However, This is not the most extensive annual flood loss the United States has suffered. In 2017, $60 billion was lost to flooding. To understand how these large losses develop, we must first understand how flooding can ensue. See Table A.1 in Appendix A for Estimated amount of damage incurred by flood.

Flooding can have many points of origin. These can vary from natural disasters, such as hurricanes or "100-year storms", to extreme climate conditions, such as abnormal precipitation. While these threats seem to be perpetual, climate change and sea-level rise may cause severe flooding even in the absence of natural disasters. According to Coastal Climate Solutions, flooding from annual King Tides has increased year-over-year for the past decade. They estimate that coastal properties will be subjected to multiple days of increasingly more severe King Tide flooding each year. They also forecast flooding, solely from King Tides, could increase by more than four times by 2030.

Flooding may also occur from human-made phenomena such as dams, pipelines, levee failure, and sewer system failure. Often flooding caused by human-made phenomena ultimately leads to what is known as urban flooding. Urban flooding can result from two different types of flooding: surface water flooding and sewerage flooding. Surface water flooding occurs when the drainage system cannot keep up with the speed that cumulative rainfall hits it, and flooding spreads from water collecting on roads and pavements. Intense rainstorms are becoming more frequent because of climate change. That, along with inadequate maintenance of drains and inadequate capacity of drains, are imperative factors. These factors also contribute to what is known as sewerage flooding. The overflow of combined sewage usually overflows causes sewerage flooding.

Government organizations have been attempting to understand the origins of floods and how to mitigate their losses for over half a century. The National Flood Insurance Program (NFIP), established by Congress in 1968, operates under the jurisdiction of the Federal Emergency Management Agency (FEMA). The function of the NFIP is both to offer primary flood insurance to properties with significant flood risk and reduce flood risk through the adoption of floodplain management standards. FEMA is responsible for the United States' support for disasters. FEMA manages a Risk Mapping, Assessment, and Planning (Risk MAP) process to produce Flood Insurance Rate Maps (FIRMs).

NFIP provides the FIRMs that function as the basis for setting insurance premiums. FIRMs define a high-risk area designated as a Special Hazard Flood Area (SFHA) where the annual probability of flooding is estimated to be greater than 1-in-100. Regions outside the SFHA are considered to have an annual flood probability of less than 1-in-100. Very few homeowners purchased coverage during the initial five years in either zone. This caused Congress to pass the federal Flood Protection Act of 1973. The act requires all properties located in a SFHA to purchase flood insurance if they have a mortgage or loan

from a federally backed or regulated lender. Property owners are not required to purchase flood insurance in areas that FEMA has designated as having an annual flood probability of less than 1-in-100.

This separation of SFHA and non - SFHA caused what is known as the protection gap. Homeowners outside the SFHA are misled to believe that they are not at risk. It seems like a bonus to them for not being forced to pay for flood coverage. The reality is that homes outside the SFHA may also be at risk for severe damage. FEMA estimated that before Hurricane Harvey (2017), only 15 percent of residents in Harris County, Texas had flood insurance, and fewer than half of homeowners in Florida were protected against the losses they experienced from Hurricane Irma (2017). Less than 1 percent of Puerto Rico households had NFIP flood insurance when Hurricane Maria (2017) devastated the island.

Another NFIP's shortfall is its relative lack of choice. The NFIP lacks variety both in terms of alternative products and policies. This lack of variety in policies may increase the protection gap. One way to expand the risk pool and eliminate adverse selection is to make flood insurance compelling enough for all homeowners to consider buying it seriously. One approach is to bundle flood into homeowner policies. This may be the most efficient economic method to cover the risk for insurers and deliver value and peace of mind to policyholders. Risk pricing is also another option. Risk pricing will change price based on the risk of flooding. Risk pricing will not always incentivize the homeowner to buy coverage.

Insurances can spread the event's cost across all policyholders, each of whom will only be paying a relatively small premium so insurers can cover the large losses that are only experienced by a few. Minimizing the protection gap would spread the cost of the event across more homeowners. This increase in penetration may lead to a decrease in premiums.

Insurers can encourage investments in cost-effective loss reduction measures in return for premium discounts. These cost-effective loss reduction measures are not only applicable to homeowners. Miami Beach and the state of Florida are currently investing billions of dollars in mitigation efforts, which include scores of industrial-level sump pumps. The 2016 King Tide season caused floods to reach heights beyond what was expected. Miami Beach was able to activate many of the pumps, and the damage was kept in check for the most part. This illustrates how local and state governments can promote policy coverage by lowering the likelihood of a flood. A primary concern expressed is that the pumps alone will not be enough to prevent future damage from relatively moderate King Floods or worse.

Karen Clark of Applied Insurance Research (now AIR Worldwide) introduced the first commercially available hurricane model in 1987. This model evolved into what is known as a storm surge model. A storm surge model uses available meteorological data, such as air pressure, wind speed, storm size, speed, and track combined with the localized water bathymetry, tide estimates, and land elevation to estimate flood inundation and velocity. The vulnerability component of the model then estimates how the water depth and velocity translate into property damage. This would not be the only model used to

analyze floods and their potential losses in the future, but it was a start. The difficulty in modeling inland floods rests in simulating all the various forms of inland flooding on top of the United States' massive landmass.

A successful model must keep track of all the numerous types of storms, covering various lengths of time, and must even consider winter snowpack, river icing, and premature thawing. A successful model must also simulate how and where excess water will flow, how the excess water will change the elevation of rivers, lakes, and streams, and how the hydrology of the discharge is changed. The model must also have a database of the elevation of the entire landmass and accurately describe how floodwaters will traverse across both flood plains and beyond. The model should also determine the ability to mitigate damage using levees, sewers, and other techniques.

We have been assigned to analyze how floods caused by many factors can lead to severe loss in North Carolina, Missouri, and Michigan. The task is not easy and has been continuously attempted for over half a century. However, we can use all of the previous information compiled together for the development of our model. We were able to read *A Methodological Approach for Pricing Flood Insurance and Evaluating Loss Reduction Measures: Application to Texas* from the Wharton University of Pennsylvania. This paper was written by Jeffrey Czajkowski, Howard Kunreuther, and Erwann Michel-Kerjan. We were able to contact Jeffrey Czajkowski and speak to him about his process and experience to hopefully learn from it and use what we learned in our models. Along with this, we know that we will need an immense amount of data to make a proper and useful model. George E. P. Box once said, "All models are wrong, but some are useful." We understand that our model will not be perfect, but we will work to be useful. The model will be useful if it helps us better understand how a flood can arise and how its severity can impact the losses of where it strikes. From there, we will use the results of the model to provide useful recommendations to hopefully mitigate both the likelihood of the flood and the amount of losses.

3 Model Development and Results

The frequency and severity of an event are two critical elements to look at when estimating the risk at what we are exposed to. By combining these two elements we get an estimate for the distribution of potential losses. Based on this idea, we developed the following models:

1. Flood Probability model (FP). This model estimates the probability that a particular county will experience a flood event in the near future.
2. Expected Percentage Payment model(EPP). We use this model to estimate the potential payment or loss per policy in case of a flood event.

We elaborate on each of these models in Section 3.2.2.

Table 3.1.1: NOAA - Climate Variables

Element	Description	Units
PRCP	Precipitation	tenths of mm
SNOW	Snowfall	mm
TMAX	Maximum temperature	tenths of degrees C
TMIN	Minimum temperature	tenths of degrees C

3.1 Data Methodology

3.1.1 Data Identification

We identified the following data sources for the development of our models.
To built the flood probability model:

1. We used climate factors such as precipitation, snow, and temperature as predictors to the flood probability model since floods may result from extreme precipitation or substantial winter snow accumulations. Also, these factors are a crucial indicators of climate change.
2. Special Flood Hazard Areas, Structural and Topographic data that characterize the county are used as predictors. These predictors provide information on how flooding damage may occur close to water bodies and result from heavy precipitation that overwhelms the existing sewer infrastructures.
3. Historical data of the flood event per county and year is used as a response variable to predict the probability of flood.

To estimate the expected claim amount:

1. Yearly amount paid in claims per county provide us information on the impact flood has had on insurance companies' losses.
2. Yearly Insurance Coverage on buildings provide us information total flood loss by the insurance company

3.1.2 Data Reliability

We retrieve data from The National Oceanic and atmospheric administration' Climate Data Online (NOAA). The Global Historical Climate Network (GHCN) includes quality and controlled daily summaries from land surface stations in the US (See climate variables in Table 3.1.1). NOAA's Storm Events Database is essential in identifying historical flood events. NOAA is an American agency and a reliable source for environmental information.

The Federal Emergency Management Agency's Datasets (FEMA) provide historical data on losses of insurance companies due to flood (See Table 3.1.2). The National Flood Hazard Layer database contains current sufficient flood hazard data, including flood zone ratings, structural and topographic information that characterize project areas. FEMA'S

Table 3.1.2: FEMA - FIMA NFIP Redacted Claims

Variable	Description
Claims	Total claims by county per year
Amount Paid in claims	Total amount paid on building, contents and increased cost of compliance per year
Insurance Coverage	Total insurance coverage on building and contents per year

Table 3.1.3: FEMA - National Flood Hazard Layer Variables

Predictor	Description
F_Aqueduct	Aqueducts found in the county
F_Channel	Channels found in the county
F_DAM	Dams found in the county
F_Pipeline	Pipelines found in the county
F_LeveeCenterLine	Levee Center Lines found in the county
F_Area not included in Flood Zone	Area not included in Flood Zone
F_minElev	Minimum elevation in the county
F_0.2 percentage annual chance flood zone	Areas rated of minimal flood hazard
FH_A_FLD_ZONE (SFHA)	Special Flood Hazard Area(SFHA),1% chance of flood
F_Open Water Flood Zone	Flood Hazard Area rated

Flood maps are the basis of the National Flood Insurance Program (NFIP) regulations and flood insurance requirements. As a federal agency, FEMA is a reliable source of information (See Table 3.1.3.).

3.1.3 Data Transformations and Formulas

We computed yearly average for maximum and minimum temperatures, and the cumulative value for precipitation and snow. We also calculated the seasonal values for each variable (Winter (Wn): January, February and March; Spring (Sp): April, May and June; Summer (Sm): July, August and September; and Fall (Fll): October, November and December). We found that some counties have more than one station. To aggregate the variables per county, we calculated the Haversine distance between the stations and the county centroid. See equation 1.

$$d = 2r\arcsin(\sqrt{sin^2(\frac{\phi_2 - \phi_1}{2}) + cos(\phi_1)cos(\phi_2)sin^2(\frac{\lambda_2 - \lambda_1}{2})}) \quad (1)$$

We then estimated each climate variable's values at the county centroid by applying an inverse-distance weighting formula. As a result, we computed a weighted average for a county. See equation 2.

Table 3.1.4: Lag Climate Variables

Predictor	Description	Example
VarP_{t-1}	Variable Value lag 1 year	TmaxP1, TminSpP1
VarP_{t-2}	Variable Value lag 2 year	PrcpP2, PrcpSmP2
VarP_{t-3}	Variable Value lag 3 year	SnowP3, SnowWnP2
$\delta\text{VarP}_{t-1,t-2}$	Percentage change between VarP_{t-2} and VarP_{t-1}	$dTmaxP12 = 100 * \frac{TmaxP1-TmaxP2}{abs(TmaxP2)}$
$\delta\text{VarP}_{t-1,t-3}$	Percentage change between VarP_{t-3} and VarP_{t-1}	$dPrcpP13 = 100 * \frac{PrcpP1-PrcpP2}{abs(PrcpP2)}$

$$Z_{county} = \frac{\sum_{i=1}^{n} \frac{Z_i}{d_i}}{\sum_{i=1}^{n} d_i} \tag{2}$$

Where Z_i is the value of any climate variable of the station in the model and d_i is the distance from the station to the county centroid.

We applied transformations to the climate variables to use them as predictors for the flood probability model. We lag the yearly and seasonal values and computed percentange change between them as shown in Table 3.1.4.

After analyzing the data, we found a scarcity data problem for Michigan. There are many missings in the structural and topographic data. This problem affects the performance of the models for Michigan. Due to a lack of time and resources, we decided to omit this location from the analysis. We noticed that Michigan's data have the highest missing rate for several variables, for some them it reaches 40%. Especially for total Claims the missing rate is more than 60%. See Table B.2 in Appendix for more details.

3.2 Mathematics Methodology

3.2.1 Assumptions and justification

The development of any mathematical model requires the identification of modeling assumptions and the justification to make these assumptions. The disclosure of these assumptions allows potential users to understand the model's limitations. Below we describe the most relevant assumptions for our models.

MA1 The estimation of an average for the climate-related factors as presented in Table 3.1.1 is a good proxy for the overall climate condition for the entire county. The Haversine distance to average the information is a well-established approach used to get aggregated data for a particular geographical area.

MA2 The use of yearly data is flexible enough to identify how climate-related factors (See Table 3.1.1) and the Flood Hazard data (See Table 3.1.3) relate to the flood probability, and losses distributions.

MA3 The Flood indicator does not differentiate based on flood's root cause (flash flood, coastal flood, etc.). This helps to increase the number of events for the model to

learn from. It will also likely imply the estimation of higher flood probabilities.

3.2.2 Model Development

Flood Probability Model (FP)

Our first model aims to answer the question; what is the probability that a particular county will experience a flood event in the near future? This problem deals with the estimation of a probability of an event.

Logistic regression models are frequently used when modeling events probability. The logistic model has high interpretability making the review of the model easy to understand and easy to validate against common intuition.

We frequently reference the odds of an event happening rather than their probability. An example is a way how bets operate in horse racing competitions. You may have heard a certain horse having 2:1 odds of winning. This is a good chance of winning. Odds and probability are related as presented by the equations 3a and 3b. It is equivalent to talking in terms of odds or probabilities. We can easily move from odds to probabilities and vice-versa.

$$Odds = \frac{Probability}{1 - Probability} \quad (3a) \qquad Probability = \frac{1}{1 + Odds} \quad (3b)$$

The logistic regression model assumes there is a linear relationship between the $log-odds$ of an event and a set of predictors. This is presented in equation 4a. Then we can estimate the corresponding probability using equation 4b.

$$\log odds = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (4a)$$

$$probability = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}} \quad (4b)$$

We estimated two logistic regression models. One for North Carolina counties and one for Missouri counties.

The predictors we tested include all climate-related factors presented in Table 3.1.1. We also incorporated Flood Hazard data shown in Table 3.1.3. We also considered different transformations introduced in Table 3.1.4 to find a better fit model for each of these predictors.

The expectation is that the climate-related factors are able to capture information on climate change and climate trends. On the other hand, we are interested in how the Flood Hazard data factors can help estimating flood probabilities because these predictors allow for potential human intervention. Adding these predictors gives the model some flexibility that will allow for scenario analysis. This opens the door to potential analysis of mitigation

actions.

The variable selection is a challenging and time-consuming process and requires several iterations. The variable selection was initially led by a Stepwise procedure. Under the Stepwise procedure, predictors are being added and removed to the model based on an objective function. We used the Akaike information criterion (AIC) as our objective function (See equation 5). The AIC rewards goodness of fit measured by the likelihood function($\hat{L}$) though also penalizing the model based on the number of predictors(k) in the model. The goal of the AIC is to discourage model overfitting. Given a set of candidate models, the preferred model is the one with the minimum AIC value.

$$AIC = 2k - 2\ln(\hat{L}) \tag{5}$$

Once the Stepwise procedure has selected a candidate set of predictors, we conduct a review of the model configuration to validate interpretability and significance. We want to ensure the contribution of each predictor is intuitive and with statistical significance. An example of a not intuitive model is one that associates more precipitation with less flood probability. This is counterintuitive. In these cases, we decided to drop one at a time all variables that the contribution to the model was not intuitive. We stopped once we have a set of factors where the relationship with flood risk makes sense.

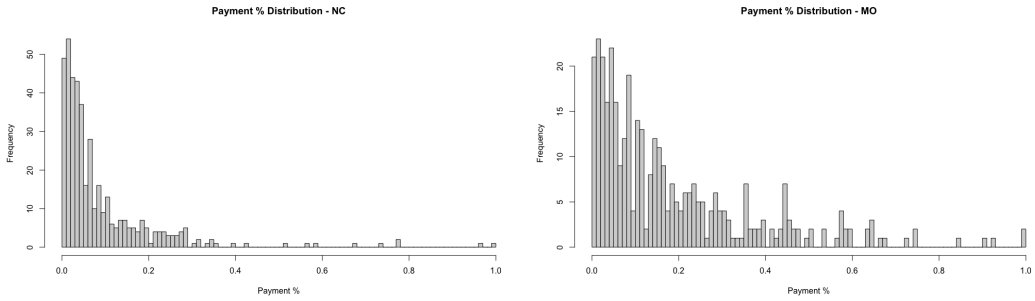
Along this process, we also looked at statistical significance for each predictor. Statistical significance tells us if one predictor helps increase the model's performance or if its contribution is limited, given that other variables are already incorporated into the model. We evaluated variable significance by inspecting the *p-value* for each predictor in the model. If the *p-value* for a particular variable is more extensive than 0.05, we concluded that this predictor is not statistically significant while still considering all other predictors already accounted for in the model. We dropped non-significant variables during the variable selection process.

Expected Percentage Payment Model (EPP)

Our second model aims to estimate the expected claim amount as a percentage of the insurance coverage. In combination with our FP model, the two models will be used to estimate the expected payout amount for an individual flood insurance policy in a county. The combined outputs of these models can be used to estimate required reserves to cover the expected claim amounts.

The estimation of these model requires information on actual claims, paid amounts, and information on insurance coverage amount. The predictor's candidates for this model were limited to the topographical and structural variables identified in Table 3.1.3.

A limited number of records had information on all required fields. This can be attributed to potential gaps in the data and limited flood insurance policy penetration. Each state reported less than 1,500 counties with all variables correctly populated.



(a) North Carolina (b) Missouri

Figure 3.2.1: Expected Percentage Payment

Figures 3.2.1a and 3.2.1b show the Expected Percentage Payment distributions for North Carolina and Missouri respectively. Both distributions are heavily skewed to the right. The distribution for NC reports most values from 0% to 20%. The Expected Percentage Payment for MO accumulates most values under 20% though a good portion of values goes from 20% to 60%.

We considered few alternatives to develop out EPP model. These include Classification and Regression Tree models (CART), Beta Regression models, and Knn-means regression.

CART models are very intuitive and usually with good fitting qualities. CART models operate as decision trees making easy to follow and use in practice. The Beta regression models are frequently used when estimating a model where the output is a number in the interval (0,1), and this is the case of our expected percentage payment.

The Knn-means regression models are also very intuitive and are based on the idea of close neighbors to estimate then make a prediction for the response variable. This algorithm is quite flexible and can account for a non-linear relationship between the predictors and the response variable. The downside for the Knn-means regression model is that this may not perform well in high dimensions (many predictors). Our set of predictors is limited, so we concluded The Knn-means regression model could also be tested to estimate the Expected Percentage Payment.

3.3 Results and Model Analysis

3.3.1 Flood Probability

Table 3.3.1 shows the list of selected predictors for the FP model for North Carolina. Meanwhile, Table 3.3.2 presents the list of variables used by Missouri FP model.

All predictors for both models reported p-values lower than 0.05. With this we reject the null hypothesis and conclude that every predictor is statistically significant, and therefore the predictors help improve model performance. See Figures C.1a and C.1b in Appendix C for more details.

Table 3.3.1: North Carolina - Flood Probability Model

Predictor	Description
dPrcpFllP13	Percentage Change in Precipitation Values During the Fall of Lag 3 years and Lag 1 year
TmaxSpP1	Average Maximum Temperature Value during the Spring of lag 1 year
TminSpP1	Average Minimum Temperature Value during the Spring of lag 1 year
PrcpSpP1	Cumulative Precipitation Value during the Spring of lag 1 year
TmaxFllP1	Average Maximum Temperature Value during the Fall of lag 1 year
FH_A_FLD_ZONE	Number of Areas Rated as Special Flood Hazard Area (SFHA), Defined as Area that will be Inundated by the Flood Event having a 1-Percent Chance of being Equaled or Exceeded in any Given year

Table 3.3.2: Missouri - Flood Probability Model

Predictor	Description
PrcpSpP2	Cumulative Precipitation Value during the Spring of lag 2 years
TminSpP3	Average Minimum Temperature Value during the Spring of lag 3 years
FH_A_FLD_ZONE	Number of Areas Rated as Special Flood Hazard Area (SFHA), Defined as Area that will be Inundated by the Flood Event having a 1-Percent Chance of being Equaled or Exceeded in any Given year
dPrcpFllP13	Percentage Change in Precipitation Values During the Fall of lag 3 years and lag 1 year
dTmaxSmP12	Percentage Change in Maximum Temperature Values During the Summer of lag 2 years and lag 1 year
dTminSmP12	Percentage Change in Minimum Temperature Values During the Summer of lag 2 years and lag 1 year

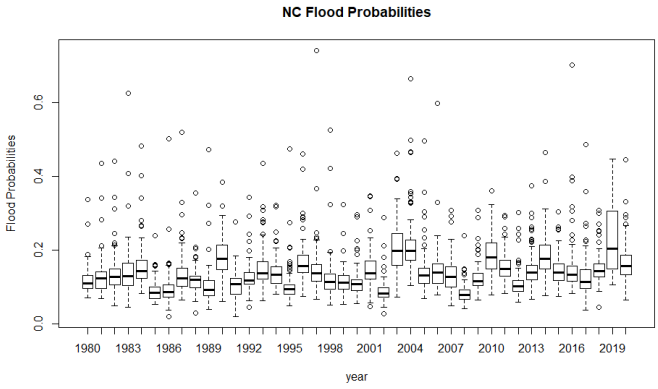


Figure 3.3.1: North Carolina - Estimated Flood Probability Distribution

We evaluate the model's performance using the Receiver Operating Characteristics curves (ROC) and associated Area Under the Curve(AUC) statistic. The AUC measures the discriminatory power of the model to predict the response variable correctly. In this case, the event is a county registering a flood event based on historical data. The AUC range is from 0.5 to 1, where 0.5 represents a model with no discriminatory power and one corresponds to a model that can correctly predict an event's occurrence (flood).

The AUC for FP model for North Carolina is 0.6547, and the AUC for FP model for Missouri is 0.675. Both models report what can be considered low discriminatory power. See Figure D.1 in Appendix D for more details on ROC curves and AUC statistics.

Figure 3.3.1 presents box plots by year for the estimated flood probabilities for all North Carolina counties. The box plots show how the flood probabilities change over time based on the information captured by each set of predictors. See Table 3.3.1 for a list of the final set of predictors.

Most of the estimated flood probabilities for North Carolina range from 0.1 to 0.2 though some years report estimated probabilities from 0.2 to 0.3. This is the case of the years 2003 and 2019.

Some concurrent years show large changes in the flood probabilities from one year to another. We decided to look at the predictor's distribution by year to get a better idea of the driver for such variation. Our analysis indicated that predictors dPrcpFllP13 *(Percentage Change in Precipitation Values During the Fall of lag three years and lag one year)* and PrcpSpP1 *(Cumulative Precipitation Value during the Spring of lag one year)* present the largest shifting in their distribution over time explaining the shifting in the distribution of flood probability. See Figure E.1 in Appendix E for box plots presenting the predictors distributions over time.

As for Missouri FP model, Figure 3.3.2 presents the box plots by year for the estimated flood probabilities. The estimated flood probabilities for Missouri seem to be larger than the estimated probabilities for North Carolina. Most of the estimates are between 0.2 and

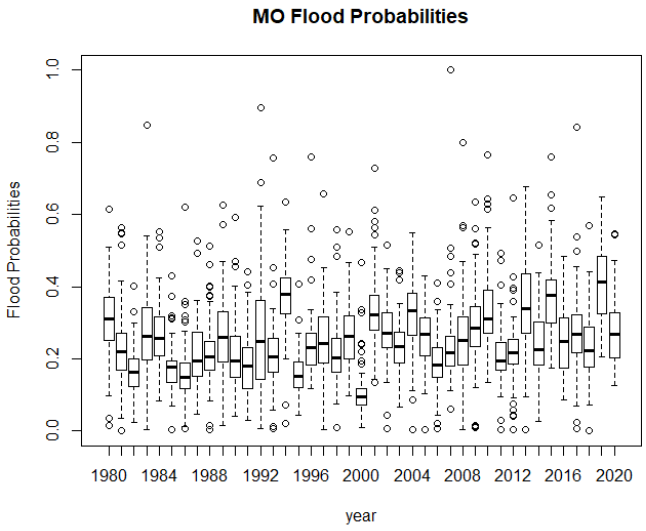


Figure 3.3.2: Missouri - Estimated Flood Probability Distribution

0.4. Some years, however, the report estimated probabilities higher than 0.4. This is the case of years 2013, 2015, and 2019.

Predictors PrcpSpP2 *(Cumulative Precipitation Value during the Spring of lag two years)*, dPrcpFllP13 *(Percentage Change in Precipitation Values During the Fall of lag three years and lag one year)*, and dTmaxSmP12 *(Percentage Change in Maximum Temperature Values During the Summer of lag two years and lag one year)* showed the most variation in their distribution over time. This shifting drives the change in the estimated flood probability as presented in Figure 3.3.2. See Figure E.2 in Appendix E for box plots presenting the predictors distributions over time.

3.3.2 Expected Percentage Payment

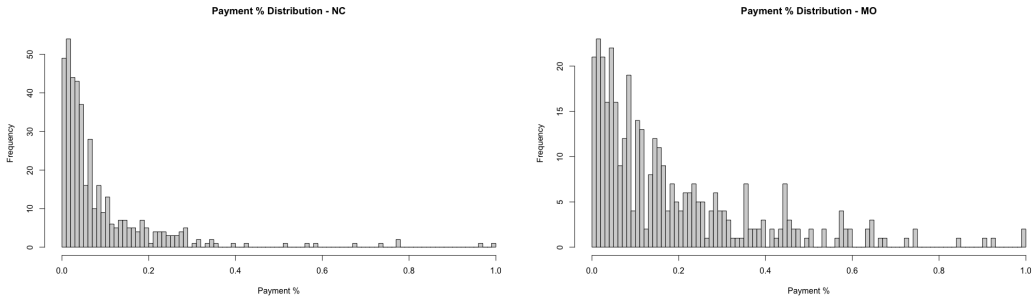
We tested three modeling approaches to estimate the EPP models.

- Classification and Regression Tree models (CART)
- Beta Regression
- Knn-means Regression

The distributions of Expected Percentage Payment for both states are right-skewed with most values from 0% to 20%. See Figure 3.3.3.

CART

The CART model looks for the predictors that better stress differences in the average response based on some binning or bucketing. Figure 3.3.3 shows the inter-quantile range is narrow for both distributions though North Carolina has the narrowest one. The less



(a) North Carolina (b) Missouri

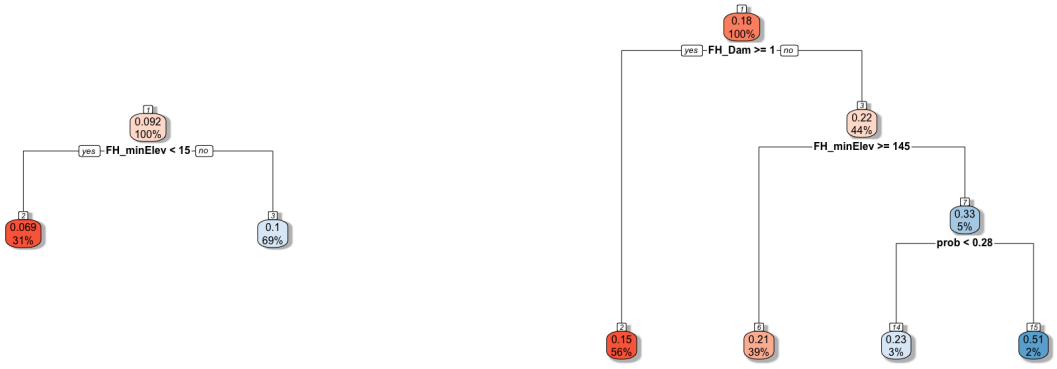
Figure 3.3.3: Expected Percentage Payment

the variability or dispersion in the response variable, the more challenging to identify predictors that can differentiate in the response variable.

Figure 3.3.4 shows the CART models for North Carolina (3.3.4a) and Missouri (3.3.4b).

The CART model for NC only uses the predictor **F_minElev** to estimate the average Expected Percentage Payment. Policies in a county where the **F_minElev** is less than 15 feet are estimated to have an average Expected Percentage Payment of 6.9% otherwise the average Expected Percentage Payment is 10.0%.

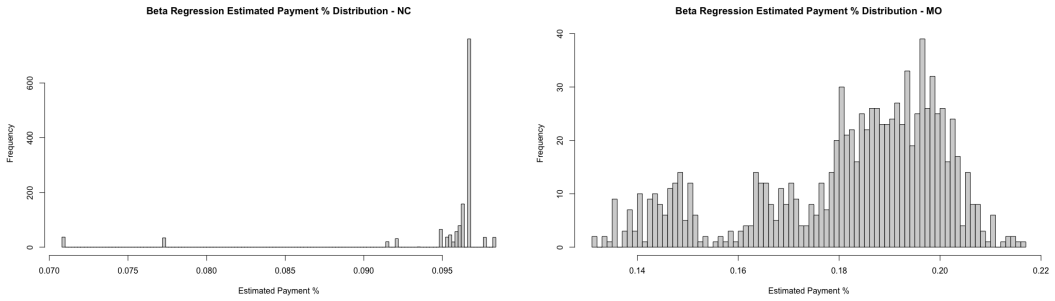
The CART model for MO uses a combination of the predictor **F_minElev, F_DAM, and Flood Probability** to estimate the average Expected Percentage Payment. The estimated average Expected Percentage Payment under this model are 15%, 21%, 23%, and 51%. The average depends on each of the predictors' values as presented in Figure 3.3.4.



(a) North Carolina (b) Missouri

Figure 3.3.4: Classification and Regression Tree models

Unfortunately, both CART models lack enough granularity to estimate the Expected Percentage Payment. This translates into a step function with only a few possible outcomes.



(a) North Carolina (b) Missouri

Figure 3.3.5: Beta Regression Distributions

Therefore, we decided to test other modeling approaches.

Beta Regression

The Beta Regression model assumes the response variable or a transformation of it can be approximated by a linear combination of the predictors in addition to the response variable is expected to follow Beta Distribution.

Figure 3.3.5 shows the predicted distribution for Expected Percentage Payment for North Carolina (3.3.5a) and Missouri (3.3.5a). The distributions of estimated Expected Percentage Payment for both states are left-skewed. The location of these distributions is significantly different from actual Expected Percentage Payment distributions (Figure 3.3.3).

The Beta regression models showed poor performance; therefore, we decided to look at other modeling alternatives.

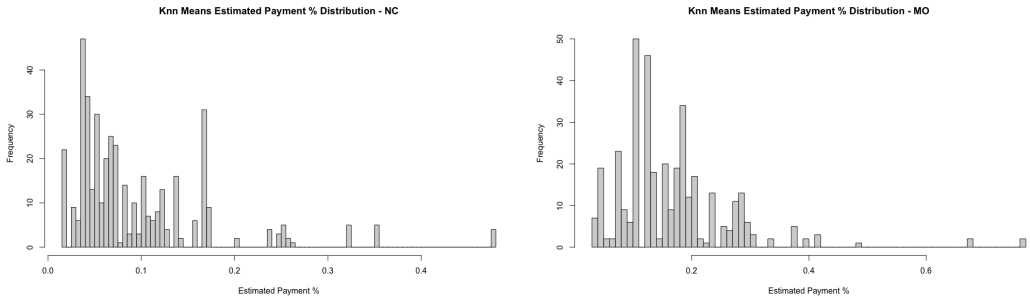
Knn-means Regression

The Knn-means regression results are also dependent on the number of close neighbors we want to look at. We tested K values from 2 to 10 and found 5 to produce the more intuitive results.

Figures 3.3.6a and 3.3.6b show the predicted distribution for Expected Percentage Payment for North Carolina and Missouri. The distribution for North Carolina is most concentrated from 0% to 15% while the distribution for Missouri shows most values from 0% to 25%

Compared to the results presented for CART and Beta Regression, the estimated distributions for Expected Percentage Payment using the Knn-means Regression are more consistent with the empirical data.

All considered, we opted for using the Knn-means Regression as the final model to estimate Expected Percentage Payment.



(a) North Carolina (b) Missouri

Figure 3.3.6: Knn-means Regression

3.3.3 Models Outputs Commentary

The Flood Probability models for North Carolina and Missouri showed limited discriminatory power. The AUC statistics are 0.6547 and 0.675 for North Carolina and Missouri, respectively. The closer the AUC to 0.5, the weaker the discriminatory power of our models.

The low model performance may relate to data scarcity in terms of the list of candidate predictors and the county and year levels' aggregation step. Developing other models using monthly data through the modeling approach will have to be adjusted to account for larger autocorrelation in the data. Other alternatives include Linear Mixed models.

Figures 3.3.1 and 3.3.2 show the box plots by year for the estimated flood probabilities. At first sight, these estimates look quite high, with estimates larger than 10% for most records. This seems counterintuitive considering that the number of claims is consistently under 10 for the vast majority of the counties (See Figures F.1 and F.2 in Appendix F). We have to remember our modeling decision to flag all counties reporting any flood event in a particular year. We did this to provide the model with enough information to learn from all of these events. This decision leads to a higher frequency and to higher probabilities. The downside in this approach is that we could be adding noise to the development data by considering minor flood events that are poorly correlated with our predictors. Later we could enhance the development data set by defining a set of criteria or materiality to meet in order for an event to be considered significant and therefore to be added to the estimation of our logistic regression model.

Nevertheless, the flood probability models provide valuable information on how the combination of precipitation data differentiating between spring and summer relates to the probability of a flood event. The variation in the amount of precipitation over time is also identified as a predictor to estimate the flood probability. See Tables 3.3.1 and 3.3.2.

All considered the flood probability models are tools that can be used to rank-order counties based on the likelihood of registering a flood event. This information can be used by insurance companies, local governments, and individuals to revisit contingency

plans, change penetration plans, and establish a need for different insurance products, etc.

Modeling the Expected Percentage Payment presented a challenge due to data scarcity and to distributions with low variability, plus the distributions are heavily right-skewed. We attempted different modeling approaches, including Classification and Regression models (CART), Beta Regression, and Knn-means Regression.

The CARTs models showed limited granularity, and the Beta Regression produced estimated distributions that differ significantly from the empirical data. The Knn-means Regression is a non-parametric model that provides enough flexibility to capture non-linear relationships in the data. Based on the results presented, we selected the Knn-means Regression as the best alternative to estimate Expected Percentage Payment.

4 Analysis and Conclusions

4.1 Risk Analysis

Although our models looked at our information at the state level, we decided to run an analysis at the county level for further comparison. Running analysis at the county level would allow us to see how the variables we chose to predict the flood probability model play out per county. We could then see what variables are impacting the probability of the flood occurring in that county. We first found the specific variable values for the top 5 highest probability counties in each state for 2020. We also found the average specific variable values across the whole state for the year 2020. We then compare the specific variable values of each top 5 counties to the state's specific variable values as a whole. We would end up with a ratio of:

$$\Delta\text{Ratio} = \frac{\text{County Value}}{\text{State Average}}$$

We will have the counties in order of highest probability to lowest and their actual predictor's values alongside them. We will also see how well they correlate to a flood's probability and their variables. These values will be listed in a table format to have a visual comparison. Table 4.1.1 will be comparing the values for North Carolina. Table 4.1.3 will be for comparing the values for Missouri.

Table 4.1.1 shows that the top 5 highest probability counties in North Carolina are Forsyth, Caldwell, Brunswick, Guilford, and Watauga. Brunswick stands out for its location. Brunswick is located right near the southeast coast of North Carolina. This makes sense for why Brunswick is in the top 5. It differs most from the state average in the number of Special Flood Hazard Areas(SFHA)/FH_A_FLD_ZONE. Brunswick is expected to have many SFHA due to being on the coast. The other 4 of the top 5 can be found around the northwestern part of North Carolina. While this may seem odd at first, it is important to see how their variables deviate from the state average. Forsyth has

(a) Number of Special Flood Hazard Areas

	SFHA	ΔRatios	Prob.
State Average	6.12		
Forsyth	0	0	0.44
Caldwell	2	0.33	0.33
Brunswick	135	22.06	0.30
Guilford	34	5.56	0.29
Watauga	0	0	0.27

(b) Last Year Spring Precipitation

	PrcpSpP1	ΔRatios	Prob.
State Average	2887.45		
Forsyth	4688.93	1.62	0.44
Caldwell	6911.45	2.39	0.33
Brunswick	3857.44	1.34	0.30
Guilford	4508.49	1.56	0.29
Watauga	5397.36	1.87	0.27

Table 4.1.1: NC - Flood Probability - Factor Benchmark

the highest flood probability(0.44), yet it does not seem to possess any SFHA. However, Forsyth's PrcpSpP1 value was more than 1.5 times the state average. Variable PrcpSpP1 represents the cumulative precipitation value during the Spring lag one year. It can be said that the precipitation in Forsyth plays a huge role in its high flood probability(0.44) and its expected number of claims. Caldwell county is number two, and it follows the same pattern as Forsyth. Caldwell's PrcpSpP1 value is more than 2.2 times the state average. Caldwell also only reports two SFHA, which makes it similar to Forsyth. This scenario follows for both Watauga and Guilford. The county with the lowest PrcpSpP1 value is Brunswick. We have also put a table to show counties that have been in the top 5 from 2016 to 2020 for North Carolina. See Table 4.1.2a

It seems that Brunswick is the most consistent when it comes to being in the top 5 most likely to have a flood. We assume it is very important to take into account the location of Brunswick. Since Brunswick is very close to the coast, it has many SFHA. Caldwell is the 2nd most consistent in the top 5. It also has the highest PrcpSpP1 value among all the other top 5s in 2020. From our previous studies, it was clear to us that spring was the most flood-prone season, which PrcpSpP1 has a big influence on. From all of this, we are able to differentiate counties based on these important variables that we have found. North Carolina counties with a high number of SFHA or high PrcpSpP1 value are at high risk due to having a high flood probability.

From Table 4.1.3 we can see that the top 5 highest probability counties in Missouri are Jasper, Clinton, Scott, Butler, and Taney. The county with the highest flood probability(.55) is Jasper. An important fact about Jasper county is that a river runs through its center from west to east. Jasper's SFHA is more than four times the state average. Clinton does not have any rivers going through it but a mildly severe river running along the side. Clinton's number of SFHA is more than 1.7 times the state average. However, Clinton's PrcpSpP2 value is almost 1.3 times that of the state average. Jasper's PrcpSpP2 value was almost the same as the state average. This trend of a high PrcpSpP2 value is present in all of the top 5 counties except Jasper. One thing that is important to understand the data is that Missouri is full of rivers. Butler itself has three rivers running through it. However, while these rivers may be present, they might not be a true threat as they might be little streams, and the water might not be moving at such a fast pace. For that reason, Jasper might have more SFHA than other counties due to the strength

(a) North Carolina

County	2016	2017	2018	2019	2020
Avery				1	
Brunswick	5	2	3		3
Caldwell			1	3	2
Davie				5	
Forsyth			5		1
Guilford					4
Hyde	4	1			
Jackson		4			
Lenoir			2		
Lincoln		3			
Pender		5			
Rutherford				2	
Stanly	1				
Swain	2				
Watauga			4	4	5
Wayne	3				

(b) Missouri

County	2016	2017	2018	2019	2020
Butler		3	3		4
Camden	2			5	
Girardeau	5				
Clinton					2
Cole				4	
Jackson			2	3	
Jasper	1	2	1	2	1
Jefferson		5			
Miller				1	
Ozark		4			
Platte		1			
Scott	4		4		3
St.Louis	3				
Taney					5
Vernon			5		

Table 4.1.2: Counties with the highest predictions of flood probability

(a) Number of Special Flood Hazard Areas

	SFHA	ΔRatios	Prob.
State Average	68.24		
Jasper	280	4.1	0.55
Clinton	118	1.73	0.54
Scott	35	0.51	0.47
Butler	23	0.34	0.45
Taney	90	1.32	0.43

(b) Spring Precipitation Lag 2 years

	PrcpSpP2	Δ Ratios	Prob.
State Average	2136.16		
Jasper	2193	1.03	0.55
Clinton	2775	1.3	0.54
Scott	3911.52	1.83	0.47
Butler	3109.4	1.46	0.45
Taney	3075.55	1.44	0.43

Table 4.1.3: MO - Flood probability - Factor benchmark

or severity of the river that runs through it. We have also put a table to show counties that have been in the top 5 from 2016 to 2020 for Missouri. See Table 4.1.2b

When we look at the top 5 consistent counties, we realize that the average number of SFHA is a significant variable. Jasper county is by far the most consistent of them all. Jasper has been the first three times when it comes to the highest predicted flood probability from 2016-2020. Jasper's number of SFHA is more than four times the state average, and it only has one river running through it. This is an important concept to realize as this river alone may be causing all of those SFHA. The second most consistent for Missouri is Butler county. Butler County does not have many SFHA. Instead, Butler's PrcpSpP2 value is almost 1.5 times more than the state average. We saw before how there seems to be a trend where the top 5 have a significantly high PrcpSpP2 value. This was true for all except Jasper. So another vital variable that can affect the probability of flooding can be the PrcpSpP2 value. From all of this, we can differentiate counties based on these critical variables that we have found. Missouri counties with a high number of SFHA or high PrcpSpP2 value are at high risk due to having a high flood probability.

Our quantitative analysis shows several risk groups or organizations. The primary persons at risk are individuals living within the SFHA in North Carolina and Missouri. Our model suggests this. The model shows a positive correlation between SFHA and the flood probability. SFHA represents the number of areas rated as a Special Flood Hazard Area. A positive correlation would mean the more SFHA areas you have, the higher the flood probability. Individuals within the SFHA are directly impacted by flooding and resulting losses.

Within the SFHA are specific subsets that can be identified. These specific subsets are based on environmental variables that were analyzed in our model. Our model advocates that individuals in North Carolina (already within SFHA) are more susceptible to flooding if they live in a region with higher precipitation during the spring or fall. Our flood probability model for North Carolina shows a positive correlation for both PrcpSpP1 and dPrcpFllP13. If either of them increases, our model shows that the flood probability should increase as well. PrcpSpP1 represents the cumulative precipitation value during the Spring lag one year. dPrcpFllP13 represents the change in percentage during the fall between the precipitation value lag three years and lag one year. Missouri shows similar traits. Our flood probability model for Missouri suggests that Missouri individuals (already within SFHA) are more susceptible to flooding if they live in a region with more precipitation during the spring and fall. Variables PrcpSpP2 and dPrcpFllP13 both show a positive correlation with the flood probability. PrcpSpP2 represents the cumulative precipitation value during the Spring lag two years. dPrcpFllP13 represents the change in percentage during the fall between the precipitation value lag three years and lag one year.

The size and severity of floods are believed to increase over the years and, therefore, may increase the SFHA size. This, along with population increase, may lead to more exposed assets in this expanding SFHA. Individuals that are presumed to be within the SFHA are also at risk. These individuals can be under the impression that they are not

currently at risk. Individuals outside the SFHA are not forced to buy coverage. This may mislead them into believing they are not at risk. However, climate change is increasing the likelihood and severity of floods. So individuals not in the SFHA but near the edge of the SFHA are also at risk.

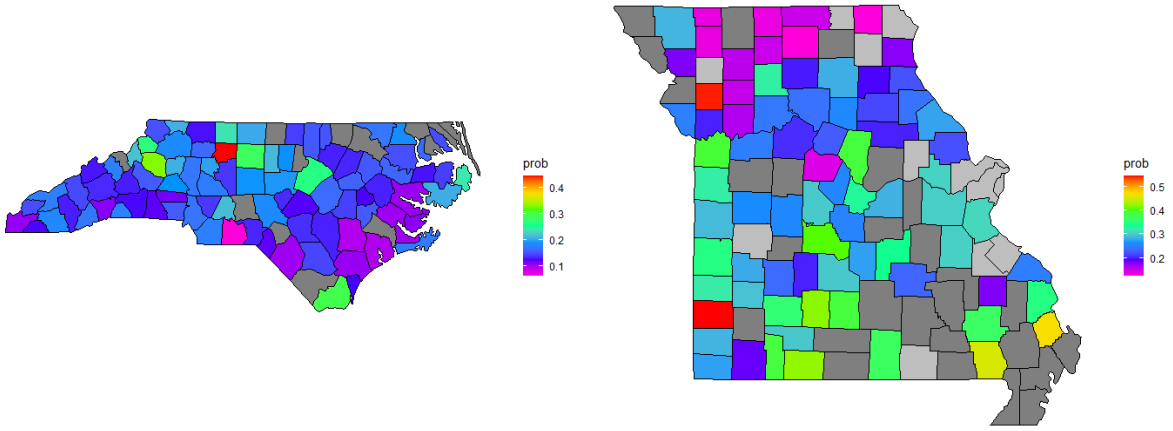
These individuals can be separated into subsets by Socioeconomic variables. Individuals near the outside edge of the SFHA are not forced to buy coverage. If the SFHA increases, eventually, these individuals will also be forced to buy coverage. Socioeconomic variables may be the difference between being able to afford coverage. Without coverage, these disadvantaged individuals are always susceptible to flooding loss and do not have a safety net to provide them peace of mind, unlike their wealthier counterparts. This can cause an economic cycle as lower economic status groups being engulfed into the SFHA cannot afford coverage and will keep falling behind due to flood losses.

It is essential to emphasize environmental variables that can be changed. Giving importance to environmental variables that cannot be changed is not useful as these cannot be used to reduce the flood probability or flood losses. We were able to read *A Methodological Approach for Pricing Flood Insurance and Evaluating Loss Reduction Measures: Application to Texas* from the Wharton University of Pennsylvania and written by Jeffrey Czajkowski, Howard Kunreuther, and Erwann Michel-Kerjan. We reached out to Dr. Czajkowski personally and got some insight on their study. Dr. Czajkowski informed us that they attained infrastructure data on specific building material, age of construction, and base elevation. They found that base elevation influences number of claims therefore and is a possible mitigation strategy. Base elevation can also help create subgroups inside and outside the SFHA.

Our analysis of risks in North Carolina can be better seen through visual representations of our data. From our Flood Probability model, we were able to find the flood probability per county in North Carolina and Missouri. We computed these probability values and colored a map of the counties of North Carolina and Missouri accordingly. See Figure 4.1.1a for North Carolina and Figure 4.1.1b for Missouri.

We have already looked at the riskiest counties in both North Carolina and Missouri. Now we will look at how these can change based on variable changes. We will be changing both climate variables and infrastructure variables. We will be changing each variable by its standard deviation independently and simultaneously. We will do this for one and two standard deviations. Figure 4.1.2a is one Standard Deviation Change for North Carolina. Figure 4.1.2b is a two Standard Deviation Change for North Carolina. Figure 4.1.3a is one Standard Deviation Change for Missouri. Figure 4.1.3b is a two Standard Deviation Change for Missouri. We will then compare these to our original probabilities (Figure 4.1.1a for NC and Figure 4.1.1b for MO).

When we look at Figure 4.1.2a, we can see how the increase by one standard deviation is enough to increase the flood probabilities by a significant amount from our original values. Figure 4.1.2a can be used to create even more risk groups just based on the likelihood of an extreme event. It can be seen that the counties with the highest probabilities are in



(a) North Carolina (b) Missouri

Figure 4.1.1: Flood Probability Maps

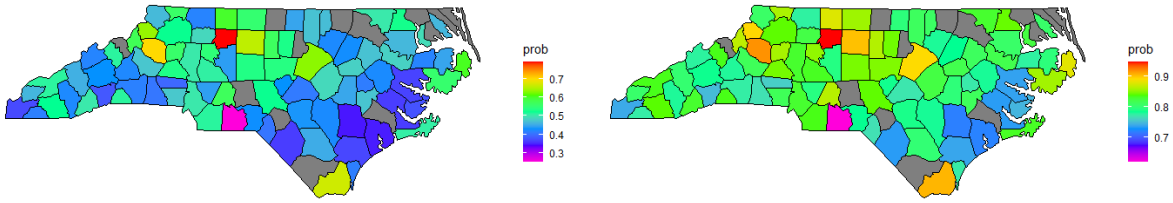
the northeastern part of North Carolina. This specific region can be considered its risk group.

When we look at Figure 4.1.2b we can see how the increase by two standard deviations is enough to increase the flood probabilities by a significant amount from our original values. Figure 4.1.2b shows us what the probabilities would be like in the likelihood of a catastrophic event. We can then see what areas will be riskiest during these catastrophic events. The northeastern counties in North Carolina still seem to be the highest risk region, but it seems as if the gap between the northeastern and the rest of the counties is closing.

When we look at Figure 4.1.3a we can see how the increase by one standard deviation is enough to increase the flood probabilities by a significant amount from our original values. Figure 4.1.3a can be used to create even more risk groups just based on the likelihood of an extreme event. In Missouri, we can see that the closer an individual is to the southeastern part of Missouri, the higher the risk of flooding. However, we do some outliers were the riskiest seems to be in the northeast part of Missouri.

When we look at Figure 4.1.3b, we can see how the increase by two standard deviations is enough to increase the flood probabilities by a significant amount from our original values. Figure 4.1.3b shows us what the probabilities would be like in the likelihood of a catastrophic event. We can then see what areas will be riskiest during these catastrophic events. The change from Figure 4.1.3a to Figure 4.1.3b is not much. We can still see that the southeast region of Missouri does seem to contain overall riskier zones. Both Figure 4.1.3a and Figure 4.1.3b can be used to create region risk zones.

During the analysis we could identify the individuals within the SFHA as the groups at the highest risk. Everything in those areas, such as homes, businesses, schools, hospitals,



(a) 1 STD (b) 2 STD

Figure 4.1.2: North Carolina - Flood Probability Maps

crops, and buildings, is at risk because of flooding. Also as a consequence insurance companies are at risk of a big amount of claims due to flooding. Over time the SFHA will increase and at one point it will engulf the individuals near the edge, we believe these individuals belong to a future risk group.

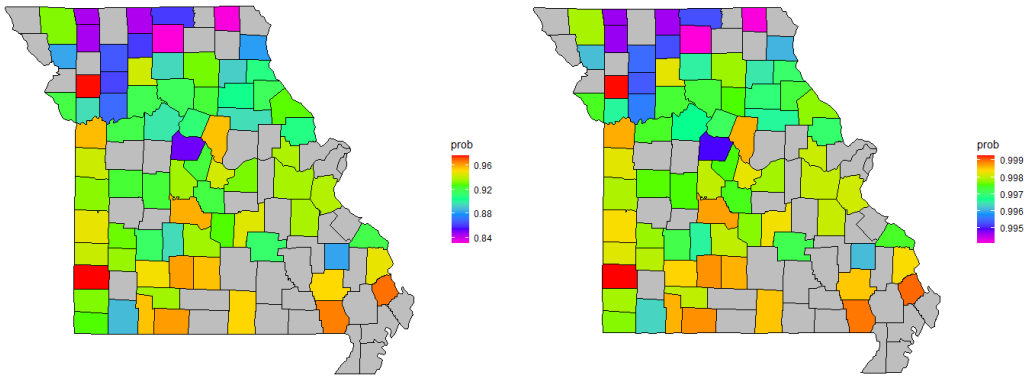
4.2 Recommendations

Using our expected percentage payment model (Section 3.2.2), we were able to find the expected percentage payment for each county in both North Carolina and Missouri for insurers. We created a colored map according to expected percentage payment values as we did with the flood probability. This can be seen in Figure 4.2.1a for Missouri and Figure 4.2.1b for North Carolina.

Figure 4.2.1a shows that this map corresponds with all of our flood probability maps. Our regions of higher probability will also seem to have higher expected percentage payment values. The insurance companies in Missouri can use this to estimate their reserve amounts for contingencies better. This will help them to understand where most of their reserves will be going.

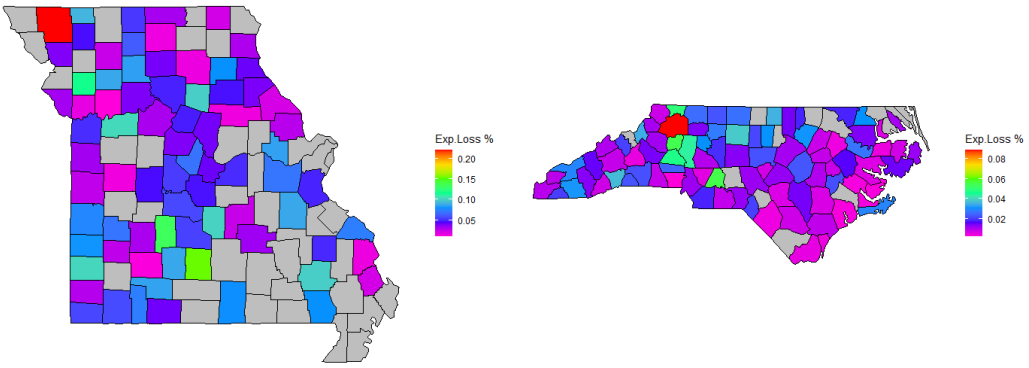
Figure 4.2.1b shows that our northeastern counties in North Carolina seem to have a higher expected percentage payment which was also the case for our flood probability maps. Using this map can be useful for creating reserves for contingencies. Insurance companies can then see where most of the reserves will be used based on the map's information.

A critical issue addressed in Section Risk (Section 4.1) Analysis was the protection gap issue between SFHA and Non-SFHA. Recommendations for this issue can be a variety of items. One of the most important ones is properly informing possible policyholders.



(a) 1 STD (b) 2 STD

Figure 4.1.3: Missouri - Flood Probability Maps



(a) Missouri (b) North Carolina

Figure 4.2.1: Expected Percentage Payment

Hopefully, if we can adequately inform all policyholders, more of them will receive coverage, and the insurers will increase penetration.

Another approach is to bundle flood into homeowner policies. This may be the most efficient economic method to cover the risk for insurers and deliver value and peace of mind to policyholders. This would again increase our penetration and therefore lowering the protection gap.

We read *A Methodological Approach for Pricing Flood Insurance and Evaluating Loss Reduction Measures: Application to Texas* from the Wharton University of Pennsylvania. This paper was written by Jeffrey Czajkowski, Howard Kunreuther, and Erwann Michel-Kerjan. We talked to Dr. Czajkowski personally and got insight on their study. They found that base elevation influences the number of claims and can be seen as a possible mitigation strategy. The recommendation coming from here would be that raising the base elevation of current or future buildings would, in return, lower both the likelihood of flooding and the number of claims.

Insurers can encourage investments in cost-effective loss reduction measures in return for premium discounts. These cost-effective loss reduction measures are not only applicable to homeowners. These encouragements from the insurance companies can persuade possible policyholders and, in return, lower the flood probability. One possible implication could be between local governments and insurers. If insurers encourage investments in cost-effective loss reduction measures, Insurers can provide products in severely adverse local government situations. These investments in cost-effective loss reduction measures can include maintenance of drains and increasing the capacity of drains to meet demands.

The most important thing to come from all of this is that we now understand that actions must be taken. These actions can be behavior changes like insurance companies properly informing house owners about raising their base elevations to lower flood risk in return for premium discounts. Other behavior changes can include government efforts to adequately prepare sewerage systems for flooding and encouraging investments in cost-effective loss reduction measures in return for premium discounts. These behavior changes will hopefully lower flood probability and flood losses.

5 Acknowledgments

First of all, thank you to The Actuarial Foundation for creating The Modeling the Future Challenge Program and giving me the opportunity to be a part of this learning experience.

I am so grateful for my mentor, Mr. Ryan Benitez, for all his support and help during this project.

Thanks to Dr. Jeffrey Czajkowski and Dr. Howard Kunreuther for responding to our communication and providing additional sources.

Finally, Thank you to my coach/mother, Samara Carmona, and my coach/father Ricardo Reyes for your support.

6 References

1. National Oceanic and Atmospheric Administration (NOAA). Climate Data Online. GHCN-Daily dataset retrieved from https://www.ncdc.noaa.gov/cdo-web/datasets.
2. National Oceanic and Atmospheric Administration (NOAA). The Storm Events Database retrieved from https://www.ncdc.noaa.gov/stormevents.
3. FEMA, "OpenFEMA". Dataset of historic claims. https://www.fema.gov/about/reports-and-data/openfema.
4. FEMA, "National Flood Hazard Layer" retrieved from https://www.fema.gov/ /flood-maps/national-flood-hazard-layer.
5. Kunreuther, H., J. Dorman, S. Edelman, C. Jones, M. Montgomery, and J. Sperger (2018). *Structure Specific Flood Risk Based Insurance*. World Scientific Publishing Company.
6. Czajkowski, J., H. Kunreuther and E. Michel-Kerjan (2012). *A Methodological Approach for Pricing Flood Insurance and Evaluating Loss Reduction Measures: Appliccation to Texas*. Center for Risk Management and Decision Processes. The Wharton School, University of Pennsylvania, January 2012.
7. Altmaier, D., A. Case, M. Chaney, N. Dolese, J. Donelon, R. Farmer, D. Jones, D. Karapiperis, C.Kousky, H. Kunreuther, N. Lamparelli, S. Larkin-Thorne, I. Maddox, T. Miller, P. Patel, B. Stringer, S. Surminski, and T. Travis (2017). *CIPR STUDY Flood Risk and Insurance* Nacional Association of Insurance Commissioners and The Center for Insurance Policy and Research, April 2017.
8. Climate Science Special Report. Chapter 8: Droughts, Floods, and Wildfire. https://science2017.globalchange.gov/chapter/8/
9. Hosmer, D., S. Lemeshow (2000). *Applied Logistic Regression* (Second Edition). Wiley-Interscience Publication.
10. Dobson, A., A. Barnett (2008). *An Introduction to Generalized Linear Models* (Third Edition). Taylor & Francis Group LLC.
11. Hastie, T., R. Tibshirani, J. Friedman (2013). *The Elements of Statistical Learning* (Second Edition). Springer Science+Business Media.
12. Python, "Tutorials and Documentation". https://www.python.org/doc.
13. R-project, "Contributed Documentation". https://cran.r-project.org/other-docs.html.
14. The LATEX Project, "Tutorials". https://www.latex-project.org.

7 Appendix

A Estimated Amount of Damage

Table A.1: Estimated amount of damage incurred by flood.

Year	Michigan	Missouri	North Carolina
1996	29,939,700	496,000	41,109,000
1997	9,300,000	667,000	13,636,300
1998	3,925,000	39,984,550	16,125,000
1999	355,000	10,354,000	103,050,000
2000	24,660,000	109,735,000	7,550,000
2001	8,454,000	1,775,000	11,755,000
2002	18,857,000	25,629,000	3,097,000
2003	16,006,000	751,000	21,836,000
2004	129,868,000	2,732,000	201,486,900
2005	919,000	1,600,000	1,734,000
2006	3,066,000	24,187,030	3,046,000
2007	777,000	535,102,000	278,000
2008	37,625,000	143,571,000	18,519,000
2009	45,125,000	1,169,000	16,898,250
2010	5,050,000	4,376,000	76,449,000
2011	10,619,000	345,865,000	7,648,500
2012	10,744,000	132,000	1,466,000
2013	101,757,500	22,039,000	23,657,000
2014	707,617,000	859,000	2,859,000
2015	100,000	324,993,000	2,402,500
2016	7,343,000	5,118,000	892,307,000
2017	211,335,000	138,717,000	1,548,500
2018	118,215,000	162,000	663,759,310
2019	102,719,000	28,755,000	3,710,000
2020	258,415,000	366,500	12,472,300
Grand Total	1,862,791,200	1,769,135,080	2,148,399,560

B Data Analysis - Missing Data

Table B.1: Missing data for North Carolina

Variable	**% Missings**	**Variable**	**% Missings**
dTmaxP12	9.54	TminFllP1	8.94
dTmaxP13	10	PrcpFllP1	8.94
dTmaxP14	10.44	TmaxP2	9.41

Continued on next page

Table B.1: Missing data for North Carolina

Variable	**% Missings**	**Variable**	**% Missings**
dTminP12	9.56	TminP2	9.41
dTminP13	10.03	PrcpP2	9.41
dTminP14	10.47	SnowP2	9.41
dPrcpP12	10.18	TmaxWnP2	9.41
dPrcpP13	10.6	TminWnP2	9.41
dPrcpP14	10.98	PrcpWnP2	9.41
dSnowP12	41.23	SnowWnP2	9.41
dSnowP13	41.74	TmaxSpP2	9.41
dSnowP14	42.13	TminSpP2	9.41
dTmaxWnP12	10.08	PrcpSpP2	9.41
dTmaxWnP13	10.55	SnowSpP2	9.41
dTminWnP12	10.11	TmaxSmP2	9.41
dTminWnP13	10.57	TminSmP2	9.41
dPrcpWnP12	10.78	PrcpSmP2	9.41
dPrcpWnP13	11.19	TmaxFllP2	9.41
dSnowWnP12	43.09	TminFllP2	9.41
dSnowWnP13	43.5	PrcpFllP2	9.41
dTmaxSpP12	9.98	TmaxP3	9.85
dTmaxSpP13	10.44	TminP3	9.85
dTminSpP12	10	PrcpP3	9.85
dTminSpP13	10.47	SnowP3	9.85
dPrcpSpP12	10.67	TmaxWnP3	9.85
dPrcpSpP13	11.09	TminWnP3	9.85
dSnowSpP12	93.93	PrcpWnP3	9.85
dSnowSpP13	94	SnowWnP3	9.85
dTmaxSmP12	9.87	TmaxSpP3	9.85
dTmaxSmP13	10.39	TminSpP3	9.85
dTminSmP12	9.9	PrcpSpP3	9.85
dTminSmP13	10.42	SnowSpP3	9.85
dPrcpSmP12	10.55	TmaxSmP3	9.85
dPrcpSmP13	11.01	TminSmP3	9.85
dTmaxFllP12	9.69	PrcpSmP3	9.85
dTmaxFllP13	10.24	TmaxFllP3	9.85
dTminFllP12	9.69	TminFllP3	9.85
dTminFllP13	10.24	PrcpFllP3	9.85
dPrcpFllP12	10.39	TmaxP4	10.26
dPrcpFllP13	10.91	TminP4	10.26
TmaxP1	8.94	PrcpP4	10.26
TminP1	8.94	SnowP4	10.26
PrcpP1	8.94	Flood	8.35

Continued on next page

Table B.1: Missing data for North Carolina

Variable	**% Missings**	**Variable**	**% Missings**
SnowP1	8.94	FH_2PctAnChanceFlood	8.35
TmaxWnP1	8.94	FH_Aqueduct	8.35
TminWnP1	8.94	FH_Channel	8.35
PrcpWnP1	8.94	FH_Dam	8.35
SnowWnP1	8.94	FH_Pipeline	8.35
TmaxSpP1	8.94	FH_LeveeCenterline	8.35
TminSpP1	8.94	FH_A_FLD_ZONE	8.35
PrcpSpP1	8.94	FH_OPEN_WATER	8.35
SnowSpP1	8.94	FH_AREANOTINCLUD	8.35
TmaxSmP1	8.94	FH_minElev	8.35
TminSmP1	8.94	FH_maxElev	8.35
PrcpSmP1	8.94	C_Claims_total	47.61
TmaxFllP1	8.94		

Table B.2: Missing data for Michigan

Variable	**% Missings**	**Variable**	**% Missings**
dTmaxP12	6.83	TminFllP1	6.37
dTmaxP13	7.14	PrcpFllP1	6.37
dTmaxP14	7.42	TmaxP2	6.68
dTminP12	6.83	TminP2	6.68
dTminP13	7.14	PrcpP2	6.68
dTminP14	7.42	SnowP2	6.68
dPrcpP12	7.57	TmaxWnP2	6.68
dPrcpP13	7.82	TminWnP2	6.68
dPrcpP14	8.03	PrcpWnP2	6.68
dSnowP12	10.28	SnowWnP2	6.68
dSnowP13	10.49	TmaxSpP2	6.68
dSnowP14	10.65	TminSpP2	6.68
dTmaxWnP12	7.63	PrcpSpP2	6.68
dTmaxWnP13	7.94	SnowSpP2	6.68
dTminWnP12	7.63	TmaxSmP2	6.68
dTminWnP13	7.94	TminSmP2	6.68
dPrcpWnP12	8.34	PrcpSmP2	6.68
dPrcpWnP13	8.58	TmaxFllP2	6.68
dSnowWnP12	11.66	TminFllP2	6.68
dSnowWnP13	11.94	PrcpFllP2	6.68
dTmaxSpP12	7.42	TmaxP3	6.89
dTmaxSpP13	7.78	TminP3	6.89
dTminSpP12	7.42	PrcpP3	6.89

Continued on next page

Table B.2: Missing data for Michigan

Variable	**% Missings**	**Variable**	**% Missings**
dTminSpP13	7.78	SnowP3	6.89
dPrcpSpP12	8.18	TmaxWnP3	6.89
dPrcpSpP13	8.46	TminWnP3	6.89
dSnowSpP12	37.66	PrcpWnP3	6.89
dSnowSpP13	38.25	SnowWnP3	6.89
dTmaxSmP12	7.38	TmaxSpP3	6.89
dTmaxSmP13	7.75	TminSpP3	6.89
dTminSmP12	7.38	PrcpSpP3	6.89
dTminSmP13	7.75	SnowSpP3	6.89
dPrcpSmP12	8.06	TmaxSmP3	6.89
dPrcpSmP13	8.34	TminSmP3	6.89
dTmaxFllP12	7.08	PrcpSmP3	6.89
dTmaxFllP13	7.45	TmaxFllP3	6.89
dTminFllP12	7.08	TminFllP3	6.89
dTminFllP13	7.45	PrcpFllP3	6.89
dPrcpFllP12	7.78	TmaxP4	7.17
dPrcpFllP13	8.09	TminP4	7.17
TmaxP1	6.37	PrcpP4	7.17
TminP1	6.37	SnowP4	7.17
PrcpP1	6.37	Flood	5.82
SnowP1	6.37	FH_2PctAnChanceFlood	38.89
TmaxWnP1	6.37	FH_Aqueduct	38.89
TminWnP1	6.37	FH_Channel	38.89
PrcpWnP1	6.37	FH_Dam	38.89
SnowWnP1	6.37	FH_Pipeline	38.89
TmaxSpP1	6.37	FH_LeveeCenterline	38.89
TminSpP1	6.37	FH_A_FLD_ZONE	38.89
PrcpSpP1	6.37	FH_OPEN_WATER	38.89
SnowSpP1	6.37	FH_AREANOTINCLUD	38.89
TmaxSmP1	6.37	FH_minElev	38.89
TminSmP1	6.37	FH_maxElev	38.89
PrcpSmP1	6.37	C_Claims_total	65.05
TmaxFllP1	6.37		

Table B.3: Missing data for Missouri

Variable	**% Missings**	**Variable**	**% Missings**
dTmaxP12	10.98	TminFllP1	9.84
dTmaxP13	11.63	PrcpFllP1	9.84
dTmaxP14	12.17	TmaxP2	10.6

Continued on next page

Table B.3: Missing data for Missouri

Variable	**% Missings**	**Variable**	**% Missings**
dTminP12	10.98	TminP2	10.6
dTminP13	11.63	PrcpP2	10.6
dTminP14	12.17	SnowP2	10.6
dPrcpP12	11.58	TmaxWnP2	10.6
dPrcpP13	12.17	TminWnP2	10.6
dPrcpP14	12.66	PrcpWnP2	10.6
dSnowP12	21.05	SnowWnP2	10.6
dSnowP13	21.52	TmaxSpP2	10.6
dSnowP14	21.88	TminSpP2	10.6
dTmaxWnP12	12.23	PrcpSpP2	10.6
dTmaxWnP13	12.83	SnowSpP2	10.6
dTminWnP12	12.23	TmaxSmP2	10.6
dTminWnP13	12.83	TminSmP2	10.6
dPrcpWnP12	13.01	PrcpSmP2	10.6
dPrcpWnP13	13.55	TmaxFllP2	10.6
dSnowWnP12	25.22	TminFllP2	10.6
dSnowWnP13	25.54	PrcpFllP2	10.6
dTmaxSpP12	11.88	TmaxP3	11.14
dTmaxSpP13	12.52	TminP3	11.14
dTminSpP12	11.88	PrcpP3	11.14
dTminSpP13	12.52	SnowP3	11.14
dPrcpSpP12	12.5	TmaxWnP3	11.14
dPrcpSpP13	13.08	TminWnP3	11.14
dSnowSpP12	92.77	PrcpWnP3	11.14
dSnowSpP13	94.02	SnowWnP3	11.14
dTmaxSmP12	11.56	TmaxSpP3	11.14
dTmaxSmP13	12.34	TminSpP3	11.14
dTminSmP12	11.61	PrcpSpP3	11.14
dTminSmP13	12.39	SnowSpP3	11.14
dPrcpSmP12	12.23	TmaxSmP3	11.14
dPrcpSmP13	12.95	TminSmP3	11.14
dTmaxFllP12	11.32	PrcpSmP3	11.14
dTmaxFllP13	12.28	TmaxFllP3	11.14
dTminFllP12	11.34	TminFllP3	11.14
dTminFllP13	12.3	PrcpFllP3	11.14
dPrcpFllP12	12.08	TmaxP4	11.58
dPrcpFllP13	12.97	TminP4	11.58
TmaxP1	9.84	PrcpP4	11.58
TminP1	9.84	SnowP4	11.58
PrcpP1	9.84	Flood	8.64

Continued on next page

Table B.3: Missing data for Missouri

Variable	% Missings	Variable	% Missings
SnowP1	9.84	FH_2PctAnChanceFlood	26.21
TmaxWnP1	9.84	FH_Aqueduct	26.21
TminWnP1	9.84	FH_Channel	26.21
PrcpWnP1	9.84	FH_Dam	26.21
SnowWnP1	9.84	FH_Pipeline	26.21
TmaxSpP1	9.84	FH_LeveeCenterline	26.21
TminSpP1	9.84	FH_A_FLD_ZONE	26.21
PrcpSpP1	9.84	FH_OPEN_WATER	26.21
SnowSpP1	9.84	FH_AREANOTINCLUD	26.21
TmaxSmP1	9.84	FH_minElev	26.21
TminSmP1	9.84	FH_maxElev	26.21
PrcpSmP1	9.84	C_Claims_total	61.92
TmaxFllP1			

C Flood Probability Model Configurations

Figure C.1a and C.1b present the model summary for the logistic model for North Carolina and Missouri respectively. The last column in both summaries shows the p-value corresponding to the hypothesis testing to investigate if each factor has a linear relationship with the response variable while other predictors are already part of the model. With p-values lower than 0.05, we reject the null hypothesis and conclude that every predictor is statistically significant, and therefore the predictors help improve model performance.

```
> summary(model)

Call:
glm(formula = Flood ~ +dPrcpFllP13 + TmaxSpP1 + TminSpP1 + PrcpSpP1 +
    TmaxFllP1 + FH_A_FLD_ZONE, family = "binomial", data = MIdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6434  -0.5697  -0.4961  -0.4140   2.4859

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.636e-01  4.705e-01   0.773  0.43969
dPrcpFllP13    2.258e-03  3.869e-04   5.836 5.34e-09 ***
TmaxSpP1      -9.877e-03  2.646e-03  -3.732  0.00019 ***
TminSpP1       1.689e-02  3.145e-03   5.370 7.85e-08 ***
PrcpSpP1       9.113e-05  4.228e-05   2.156  0.03112 *
TmaxFllP1     -1.307e-02  2.194e-03  -5.957 2.57e-09 ***
FH_A_FLD_ZONE  5.070e-03  1.988e-03   2.551  0.01075 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2778.7  on 3398  degrees of freedom
Residual deviance: 2672.1  on 3392  degrees of freedom
AIC: 2686.1

Number of Fisher Scoring iterations: 4
```

(a) North Carolina

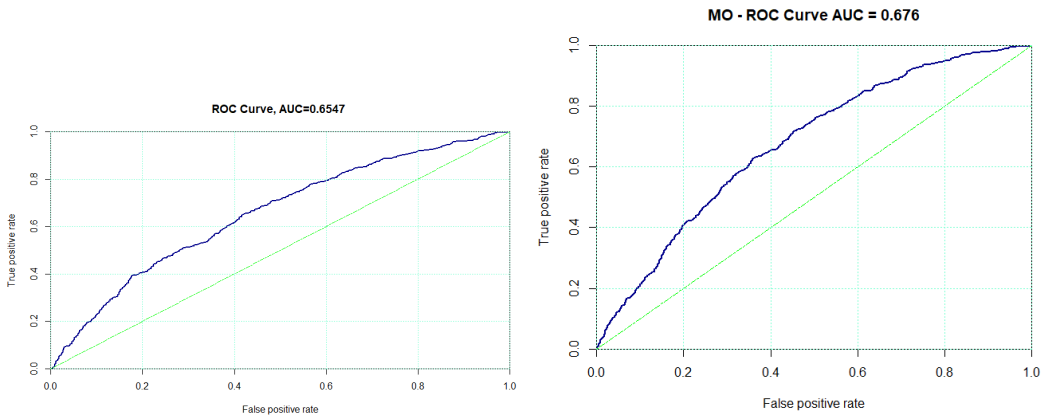
```
> summary(modelaic)

Call:
glm(formula = Flood ~ PrcpSpP2 + TminSpP3 + FH_A_FLD_ZONE + dPrcpFllP13 +
    dTmaxSmP12 + dTminSmP12, family = "binomial", data = MIdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1248  -0.7899  -0.6356   1.0361   3.4049

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.980e+00  3.956e-01 -15.114  < 2e-16 ***
PrcpSpP2       1.574e-04  3.307e-05   4.760 1.94e-06 ***
TminSpP3       3.437e-02  3.000e-03  11.456  < 2e-16 ***
FH_A_FLD_ZONE  2.871e-03  7.225e-04   3.973 7.10e-05 ***
dPrcpFllP13    1.867e-03  3.668e-04   5.089 3.60e-07 ***
dTmaxSmP12    -2.516e-02  6.945e-03  -3.623 0.000291 ***
dTminSmP12     2.393e-02  6.487e-03   3.689 0.000225 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3490.3  on 3072  degrees of freedom
Residual deviance: 3278.1  on 3066  degrees of freedom
AIC: 3292.1

Number of Fisher Scoring iterations: 5
```

(b) Missouri

Figure C.1: Flood Probability Model

D Flood Probability Model Performance

Figure D.1 shows the ROC curves and AUC statistics for both regression models. The AUC for North Carolina's model is 0.6547, and the AUC for Missouri's model is 0.675.



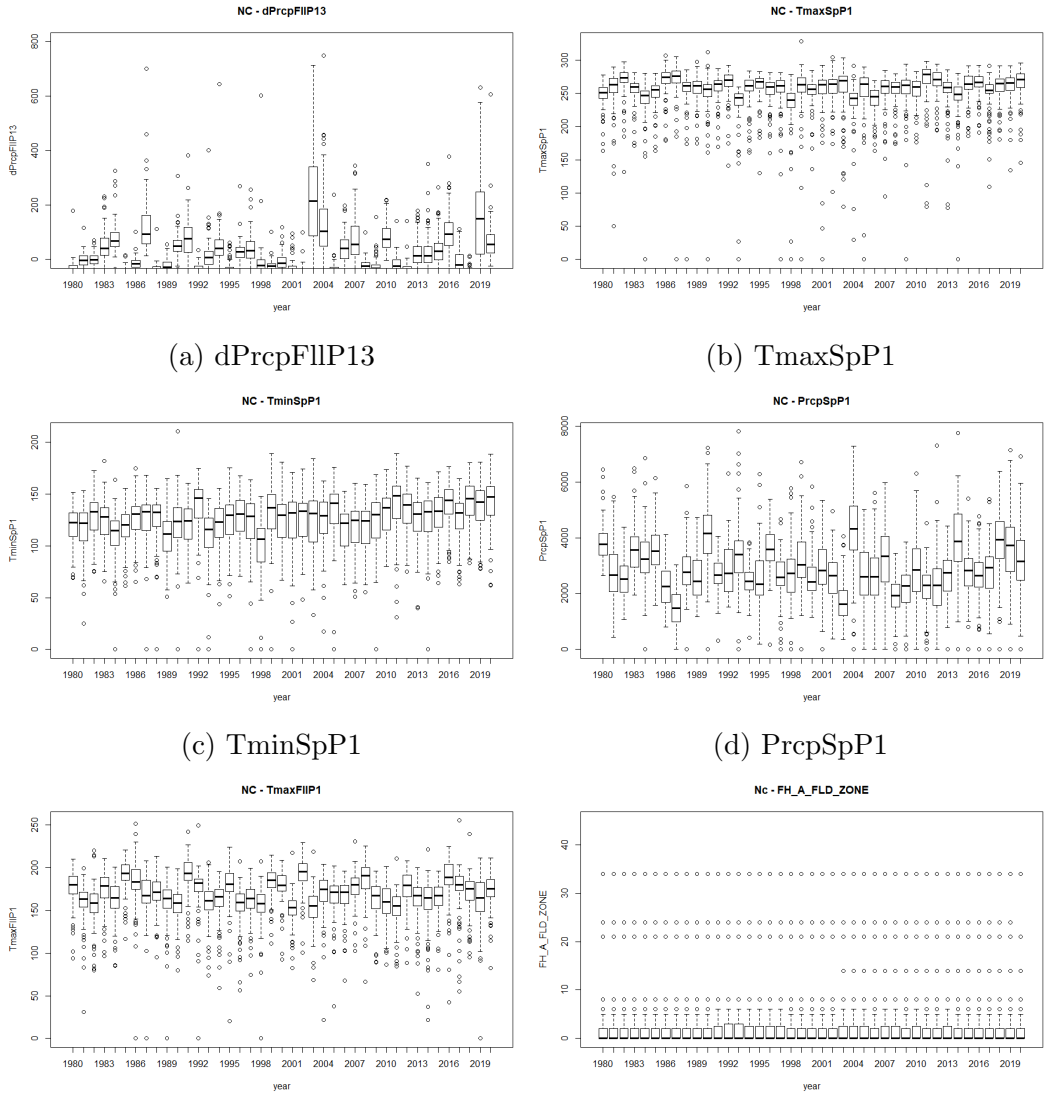
(a) North Carolina (b) Missouri)

Figure D.1: ROC Curve and Corresponding AUC Statistics

E Flood Probability Model - Factors Distributions

Figure E.1 presents the box plots for predictors in the North Carolina FP model.

Factors dPrcpFllP13 *(Percentage Change in Precipitation Values During the Fall of lag three years and lag one year)* and PrcpSpP1 *(Cumulative Precipitation Value during the Spring of lag one year)* present the largest shifting in their distribution over time explaining the shifting in the distribution of flood probability.



(a) dPrcpFllP13 (b) TmaxSpP1

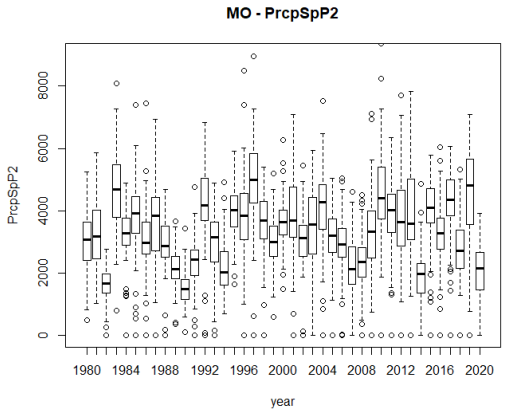
(c) TminSpP1 (d) PrcpSpP1

(e) TmaxFllP1 (f) FH_A_FLD_ZONE

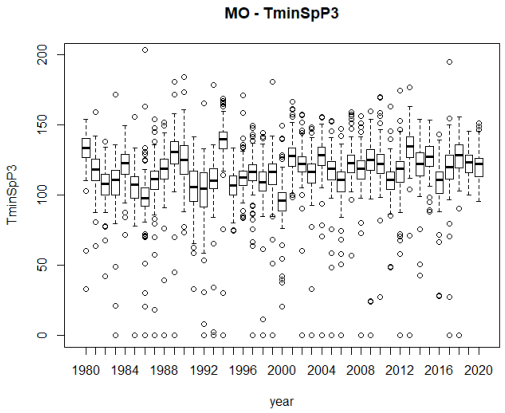
Figure E.1: North Carolina - Predictors Box Plots

Figure E.2 presents the box plots for predictors in the Missouri FP model.

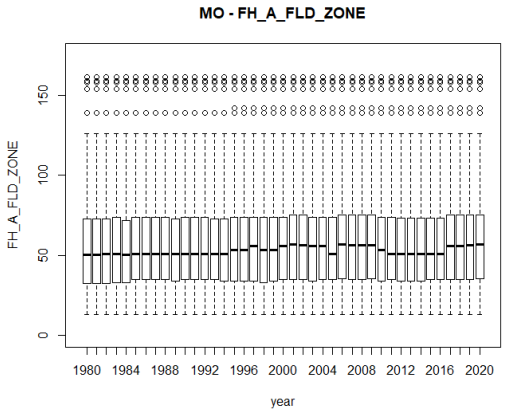
Predictors PrcpSpP2 *(Cumulative Precipitation Value during the Spring of lag two years)*, dPrcpFllP13 *(Percentage Change in Precipitation Values During the Fall of lag three years and lag one year)*, and dTmaxSmP12 *(Percentage Change in Maximum Temperature Values During the Summer of lag two years and lag one year)* show the most variation in their distribution over time. This shifting drives the change in the estimated flood probability forMissouri.



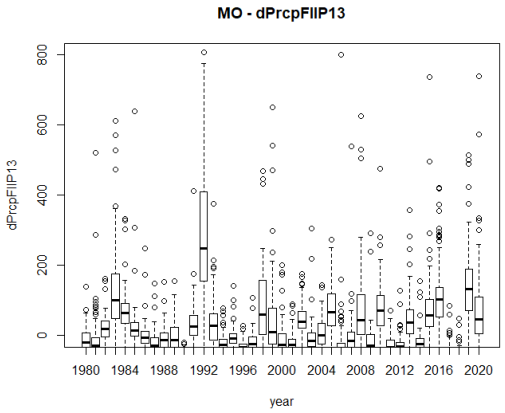
(a) PrcpSpP2



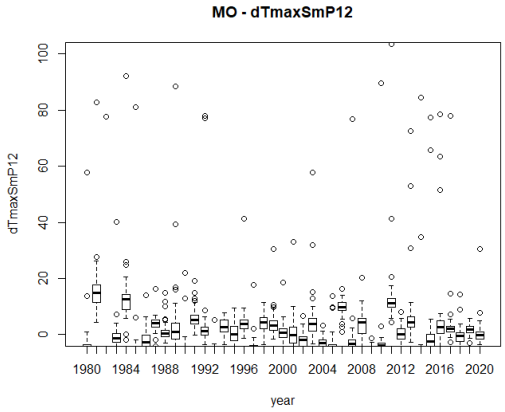
(b) TminSpP3



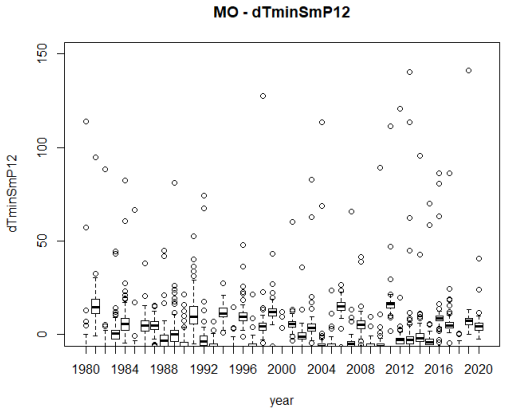
(c) FH_A_FLD_ZONE



(d) dPrcpFllP13



(e) dTmaxSmP12

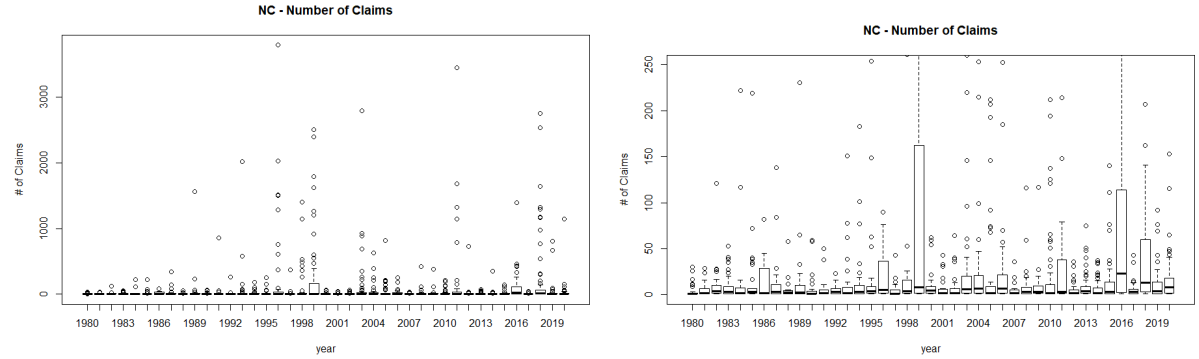


(f) dTminSmP12

Figure E.2: Missouri - Predictors Box Plots

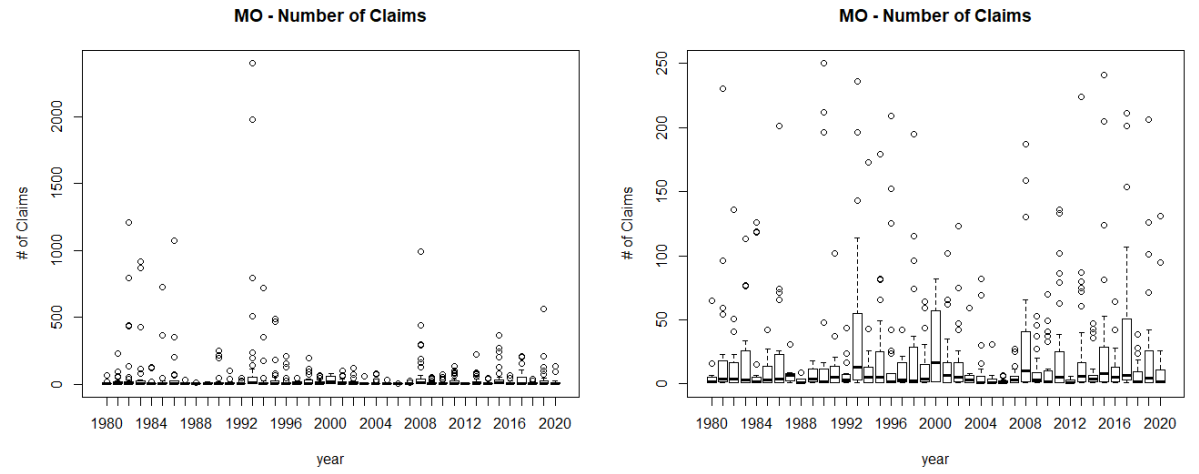
F Number of Claims

We present the box plots for the number of claims by year for North Carolina and Missouri. Figures F.1 shows information corresponding to North Carolina and Figure F.2 shows data related to Missouri.



(a) All Claims (b) All Claim - Zoom In (0 to 250)

Figure F.1: North Carolina - Number of Claims



(a) All Claims (b) All Claim - Zoom In (0 to 250)

Figure F.2: Missouri - Number of Claims

G R Code - North Carolina Flood Probability Model

```
# LOAD LIBRARY
library(MASS)
library(ROCR)
library(StepReg)
```

```
library(tidyr)
library(corrplot)

# SETS WORKING DIRECTORY
setwd('D:/MTFC/MTFC/Data')

# DELETE ROWS WITH MISSING VALUES IN VARIABLES
delete.na <- function(DF, n=0) {
DF[rowSums(is.na(DF)) <= n,]
}

# LOADS DATA
data <- as.data.frame(read.csv('MTFCData.csv', header = TRUE))

# CLIMATE VARIABLES
variables <- c( )

# STRUCTURE VARIABLES
str_variables <- c( )

# CLIMATE AND STRUCTURE VARIABLES TOGETHER
all_var <- c(variables,str_variables)

# DATA FOR NC and year > 1980
MIdata <- data[which(data$state=='NC' & data$year>=1980),all_var]
head(MIdata)
summary(MIdata)

# DELETE  MISSINGS
MIdata <- delete.na(MIdata)

# MODEL FULL AND NULL
full <- glm(Flood ~ . , data = MIdata, family = "binomial")
summary(full)
null <- glm(Flood ~ 1 , data = MIdata, family = "binomial")
summary(null)

modelaic <- step(null, list(lower=formula(null),upper=formula(full)),
direction="both",trace=0) # AIC: 2355.6
summary(modelaic)

modelaic <- update(modelaic, ~ . -FH_LeveeCenterline)
modelaic <- update(modelaic, ~ . -dTminP14)
modelaic <- update(modelaic, ~ . -dPrcpWnP13)
```

```
modelaic <- update(modelaic, ~ . -FH_Culvert - TminFllP2)
modelaic <- update(modelaic, ~ . -FH_countElev)
modelaic <- update(modelaic, ~ . -FH_OPEN_WATER_FLD_ZONE)
modelaic <- update(modelaic, ~ . -PrcpWnP1)
modelaic <- update(modelaic, ~ . -FH_Aqueduct)
modelaic <- update(modelaic, ~ . -FH_Pipeline)
modelaic <- update(modelaic, ~ . -PrcpWnP3)
modelaic <- update(modelaic, ~ . -TminWnP1)
modelaic <- update(modelaic, ~ . -TmaxWnP1)
modelaic <- update(modelaic, ~ . -TminFllP3)
modelaic <- update(modelaic, ~ . -TminSpP3)
modelaic <- update(modelaic, ~ . -TmaxSmP2)
modelaic <- update(modelaic, ~ . -TminSmP2)
modelaic <- update(modelaic, ~ . -PrcpSpP3)
modelaic <- update(modelaic, ~ . -PrcpSpP3)
modelaic <- update(modelaic, ~ . -TmaxFllP2)
modelaic <- update(modelaic, ~ . -TmaxSmP3)
modelaic <- update(modelaic, ~ . -dTminSmP13)
modelaic <- update(modelaic, ~ . -TmaxSpP1)
modelaic <- update(modelaic, ~ . -PrcpSmP3)
modelaic <- update(modelaic, ~ . -dPrcpWnP12)
modelaic <- update(modelaic, ~ . -PrcpWnP2)
modelaic <- update(modelaic, ~ . -TmaxFllP3)
modelaic <- update(modelaic, ~ . -PrcpFllP3)
modelaic <- update(modelaic, ~ . -FH_Bridge)
modelaic <- update(modelaic, ~ . -FH_Dam)
modelaic <- update(modelaic, ~ . -FH_Channel)
modelaic <- update(modelaic, ~ . -PrcpFllP2)
modelaic <- update(modelaic, ~ . -FH_maxElev)
modelaic <- update(modelaic, ~ . -FH_minElev)
summary(modelaic)

model <- glm(Flood ~
+ dPrcpFllP13
+ TmaxSpP1
+ TminSpP1
+ PrcpSpP1
+ TmaxFllP1
+ FH_A_FLD_ZONE
,
data = MIdata, family = "binomial")
summary(model)

# > summary(model)
```

```
# Call:
#   glm(formula = Flood ~ +dPrcpFllP13 + TmaxSpP1 + TminSpP1 + PrcpSpP1 +
#         TmaxFllP1 + FH_A_FLD_ZONE, family = "binomial", data = MIdata)
#
# Deviance Residuals:
#   Min       1Q   Median       3Q      Max
# -1.6434  -0.5697  -0.4961  -0.4140   2.4859
#
# Coefficients:
#   Estimate Std. Error z value Pr(>|z|)
# (Intercept)    3.636e-01  4.705e-01   0.773  0.43969
# dPrcpFllP13    2.258e-03  3.869e-04   5.836 5.34e-09 ***
#   TmaxSpP1      -9.877e-03  2.646e-03  -3.732  0.00019 ***
#   TminSpP1       1.689e-02  3.145e-03   5.370 7.85e-08 ***
#   PrcpSpP1       9.113e-05  4.228e-05   2.156  0.03112 *
#   TmaxFllP1     -1.307e-02  2.194e-03  -5.957 2.57e-09 ***
#   FH_A_FLD_ZONE  5.070e-03  1.988e-03   2.551  0.01075 *
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
# Null deviance: 2778.7  on 3398  degrees of freedom
# Residual deviance: 2672.1  on 3392  degrees of freedom
# AIC: 2686.1

 class <- model$y
 score <- model$fitted.values
 logit_scores <- prediction(predictions=score, labels=class)
 logit_auc <- performance(logit_scores, "auc")
 as.numeric(logit_auc@y.values) #

 logit_perf <- performance(logit_scores, "tpr", "fpr")
 plot(logit_perf,col = "darkblue",lwd=2,xaxs="i",yaxs="i",tck=NA,
  main="ROC Curve, AUC=0.6547 ")
 box()
 abline(0,1, lty = 300, col = "green")
 grid(col="aquamarine")
 #[1] 0.6547718
```

H R Code - Missouri Flood Probability Model

```
library(MASS)
library(ROCR)
```

```
library(StepReg)
library(tidyr)
library(corrplot)

setwd('D:/MTFC/MTFC/Data')

delete.na <- function(DF, n=0) {
DF[rowSums(is.na(DF)) <= n,]
}

data <- as.data.frame(read.csv('MTFCData.csv', header = TRUE))

variables <- c( )

str_variables <- c( )

#all_var <- variables
all_var <- c(variables,str_variables)

MIdata <- data[which(data$state=='MO' & data$year>=1980),all_var]
head(MIdata)
summary(MIdata)

# DELETE MISSINGS
MIdata <- delete.na(MIdata)

full <- glm(Flood ~ . , data = MIdata, family = "binomial")
summary(full)
null <- glm(Flood ~ 1 , data = MIdata, family = "binomial")
summary(null)

modelaic <- step(null, list(lower=formula(null),upper=formula(full)),
direction="both",trace=0)
summary(modelaic)

#drop1(update(modelaic, ~ . -FH_LeveeCenterline), test = "LRT")

modelaic <- update(modelaic, ~ . -FH_LeveeCenterline)
modelaic <- update(modelaic, ~ . -SnowWnP1)
modelaic <- update(modelaic, ~ . -PrcpFllP3)
modelaic <- update(modelaic, ~ . -PrcpP3)
modelaic <- update(modelaic, ~ . -SnowWnP2 )
modelaic <- update(modelaic, ~ . -PrcpFllP2 )
modelaic <- update(modelaic, ~ . -FH_ControlStructure )
```

```
modelaic <- update(modelaic, ~ . -SnowP4 )
modelaic <- update(modelaic, ~ . -PrcpFllP1 )
modelaic <- update(modelaic, ~ . -FH_Weir )
modelaic <- update(modelaic, ~ . -dPrcpWnP13)
modelaic <- update(modelaic, ~ . -SnowWnP3)
modelaic <- update(modelaic, ~ . -TminFllP1)
modelaic <- update(modelaic, ~ . -FH_X_FLD_ZONE)
modelaic <- update(modelaic, ~ . -dTmaxP14)
modelaic <- update(modelaic, ~ . -TminWnP2)
modelaic <- update(modelaic, ~ . -dPrcpSmP12)
modelaic <- update(modelaic, ~ . -TmaxSpP1)
modelaic <- update(modelaic, ~ . -TminWnP1)
modelaic <- update(modelaic, ~ . -SnowP1)
modelaic <- update(modelaic, ~ . -SnowFllP2)
modelaic <- update(modelaic, ~ . -FH_AREA_NOT_INCLUDED_FLD_ZONE)
modelaic <- update(modelaic, ~ . -FH_FloodwayContainedInStructure)
modelaic <- update(modelaic, ~ . -FH_Footbridge)
modelaic <- update(modelaic, ~ . -dTminFllP13)
modelaic <- update(modelaic, ~ . -dTmaxWnP12)
modelaic <- update(modelaic, ~ . -dPrcpSpP13)
modelaic <- update(modelaic, ~ . -dTminP14)
modelaic <- update(modelaic, ~ . -SnowSpP3)
modelaic <- update(modelaic, ~ . -TminFllP3)
modelaic <- update(modelaic, ~ . -TmaxSpP2)
modelaic <- update(modelaic, ~ . -FH_AH_FLD_ZONE)
modelaic <- update(modelaic, ~ . -PrcpSmP3)
modelaic <- update(modelaic, ~ . -dTminSpP13)
modelaic <- update(modelaic, ~ . -TminSpP2)
modelaic <- update(modelaic, ~ . -PrcpSpP1)
modelaic <- update(modelaic, ~ . -PrcpSpP3)
modelaic <- update(modelaic, ~ . -TmaxP4)
modelaic <- update(modelaic, ~ . -TmaxP1)
modelaic <- update(modelaic, ~ . -TmaxP2)
modelaic <- update(modelaic, ~ . -dTminWnP12)
modelaic <- update(modelaic, ~ . -TmaxWnP1)
modelaic <- update(modelaic, ~ . -TminP4)
modelaic <- update(modelaic, ~ . -TmaxSmP2)

summary(modelaic)
# > summary(modelaic)
# Call:
#   glm(formula = Flood ~ PrcpSpP2 + TminSpP3 + FH_A_FLD_ZONE + dPrcpFllP13 +
#         dTmaxSmP12 + dTminSmP12, family = "binomial", data = MIdata)
#
```

```
# Deviance Residuals:
#    Min       1Q   Median       3Q      Max
# -2.1248  -0.7899  -0.6356   1.0361   3.4049
#
# Coefficients:
#    Estimate Std. Error z value Pr(>|z|)
# (Intercept)   -5.980e+00  3.956e-01 -15.114  < 2e-16 ***
#    PrcpSpP2       1.574e-04  3.307e-05   4.760 1.94e-06 ***
#    TminSpP3       3.437e-02  3.000e-03  11.456  < 2e-16 ***
#    FH_A_FLD_ZONE  2.871e-03  7.225e-04   3.973 7.10e-05 ***
#    dPrcpFllP13    1.867e-03  3.668e-04   5.089 3.60e-07 ***
#    dTmaxSmP12    -2.516e-02  6.945e-03  -3.623 0.000291 ***
#    dTminSmP12     2.393e-02  6.487e-03   3.689 0.000225 ***
#    ---
#    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
# Null deviance: 3490.3  on 3072  degrees of freedom
# Residual deviance: 3278.1  on 3066  degrees of freedom
# AIC: 3292.1

model <- modelaic

class <- model$y
score <- model$fitted.values
logit_scores <- prediction(predictions=score, labels=class)
logit_auc <- performance(logit_scores, "auc")
as.numeric(logit_auc@y.values) ##AUC Value 0.675588

logit_perf <- performance(logit_scores, "tpr", "fpr")
plot(logit_perf,col = "darkblue",lwd=2,xaxs="i",yaxs="i",tck=NA,
main="MO - ROC Curve AUC = 0.676")
box()
abline(0,1, lty = 300, col = "green")
grid(col="aquamarine")
# [1] 0.675588
```