Health Risks and Economic Consequences of extreme weather events in U.S.

Synopsis

Severe weather events such as storm may be a threat of public health and cause economic problems for communities and municipalities. Hence preventional or mitigation policies may come are of a great importance in rescuing human lives, severe injuries and property damages. In order to assist the development of such policies, U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage. This analysis tries to identify which types of events are most harmful with respect to U.S. population health and which types of events have the greatest economic consequences.

1. Data Processing

This analysis makes use of the NOAA storm database, which tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage. The data was supplied in the form of a comma seperated values (csv) file but compressed using the open-source bzip2 algorythm (bz2) to save space.

The data are processed and analyzed in R statistical language using RStudio as a GUI.

To repeat the analysis follow the code chunks that are given below.

Please start with the required packages.

```
# set global chunk options:
library(knitr)
library(ggplot2)
library(plyr)
library(grid)
library(gridExtra)
opts_chunk$set(cache=TRUE,cache.path = 'PA2_template_cache/', fig.path='figure/')
setInternet2(TRUE)  # for https downloads.
```

1.1 Importing the data to your local machine

Here the data are imported from the internet directly to a local machine. The import includes as missing values all the values that are either coded as NA or as empty space or as a questionmark.

```
# This section will automatically download the data to your local machine.
here  <- tempdir()
myurl <- 'https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2'
mydata  <- paste(here,'StormData.csv.bz2',sep='/')
download.file(myurl,mydata)
# Here a pre-manipulation of the data is performed in order to correctly import them in R.
mydata <- read.csv(mydata, stringsAsFactors=FALSE,strip.white = TRUE,
                   na.strings = c("NA","","?"))
```

1.2 Data Manipulation

The obtained dataset includes 37 variables with 902297 observations. Here is an overview of the data set:

```
str(mydata)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ BGN_TIME  : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE : chr  "CST" "CST" "CST" "CST" ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : chr  NA NA NA NA ...
##  $ BGN_LOCATI: chr  NA NA NA NA ...
##  $ END_DATE  : chr  NA NA NA NA ...
##  $ END_TIME  : chr  NA NA NA NA ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI   : chr  NA NA NA NA ...
##  $ END_LOCATI: chr  NA NA NA NA ...
##  $ LENGTH    : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH     : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F         : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  NA NA NA NA ...
##  $ WFO       : chr  NA NA NA NA ...
##  $ STATEOFFIC: chr  NA NA NA NA ...
##  $ ZONENAMES : chr  NA NA NA NA ...
##  $ LATITUDE  : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS   : chr  NA NA NA NA ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

For this study only a part of the variables is retained. More specifically variables that include spatial information are ignored. The used temporal scale for this study is set to year (see following section). This is due to time restrictions analyzing the data.

The variables that are used for this analysis are the following:

Variable	Description
BGN_DATE	Begin date of the event

Variable	Description
EVTYPE	Type of weather event
FATALITIES	Number of deaths recorded
INJURIES	Number of injured people
PROPDMG	Costs of property damages in dollars
PROPDMGEXP	The exponent of costs of property damages
CROPDMG	Costs of agricaltural damages in dollars
CROPDMGEXP	The exponent of costs of agricaltural damages
REFNUM	Event ID number

Subseting for these variables of interest results in the following dataset

```
# Subset for variables of interest
mydatas1<-subset(mydata,select=c(2,8,23,24,25,26,27,28,37))

# Check the number of missing values
colSums(is.na(mydatas1))
```

```
##   BGN_DATE     EVTYPE FATALITIES   INJURIES    PROPDMG PROPDMGEXP
##          0          1          0          0          0     465942
##    CROPDMG CROPDMGEXP     REFNUM
##          0     618420          0
```

Quick clean-up of the data for events that did not result in any human or economic losses.

```
# Subset for events that reported injuries, fatalities, property damages andagriculrtural damages

mydatas1<- subset(mydatas1,
                  FATALITIES != 0 | INJURIES !=0 | PROPDMG != 0 | CROPDMG != 0)

mydatas1<-subset(mydatas1,!is.na(EVTYPE))
```

1.2.1 Temporal scale

For this study the focal temporal scale is set to year. For this reason the *BGN_DATE* is reformated to year.

```
# First format BGN_DATE to POSIXTct format and reformat it to year
mydatas1$BGNYEAR<-as.numeric(format(
        strptime(mydatas1$BGN_DATE,"%m/%d/%Y %H:%M:%S"),"%Y"))
```

The events in the database start in the year 1950 and end in November 2011. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete. This is clearly illustrated in the following plot where the number of recorder events is plotted against the year.

```
ggplot(mydatas1,aes(x=BGNYEAR))+
        geom_histogram(colour = "darkgreen", fill = "white",binwidth=1)+
        theme_bw()+xlab("years")+ylab("Number of events")+
        ggtitle("Total recorded events per year")
```

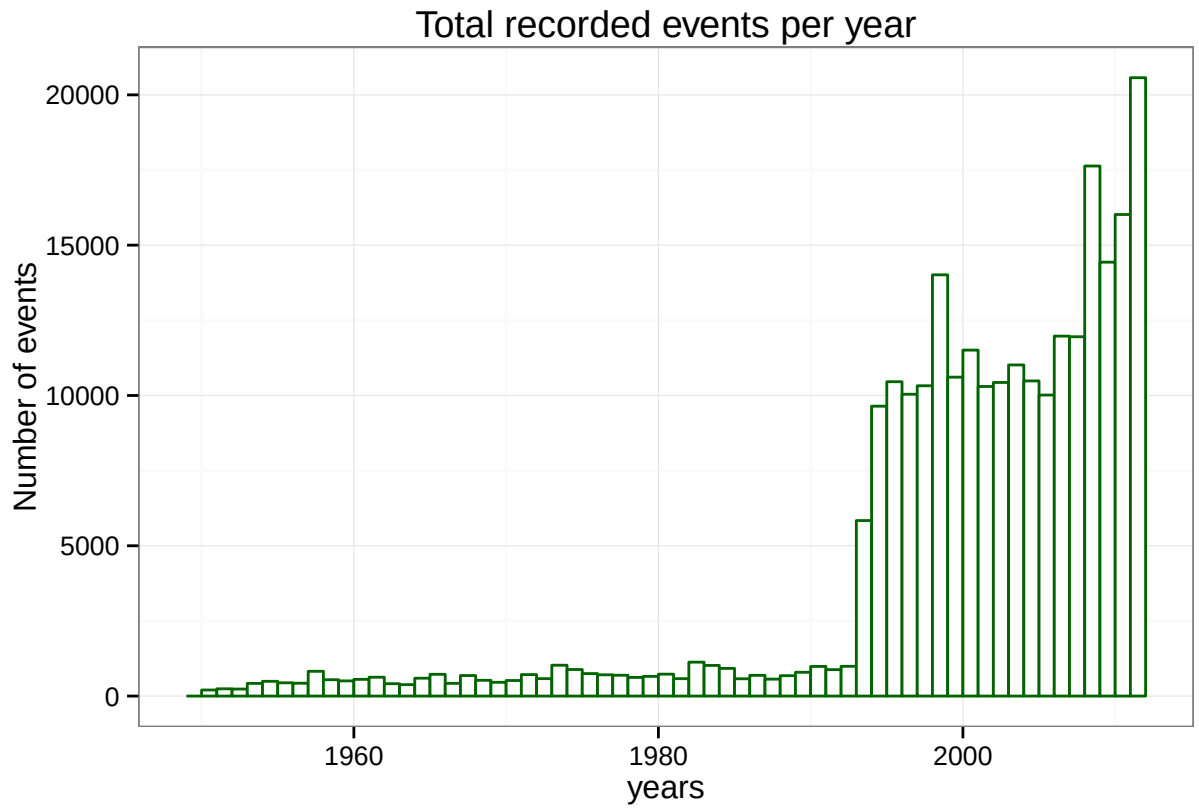

Figure 1. Year effect on the recorded events. Notice the abrupt increase of events after 1994.

Before the early nighties the recorded weather events are few. After that the number of events increases steaply. This indicates that the events before 1994 are less likely to be complete. Therefore all year before 1994 are excluded from the analysis.

```
mydatas2<-subset(mydatas1,BGNYEAR>=1994)
```

1.2.2 Human Casualties

In this dataset there are two variables indicative for human casualties either injuries or death caused by weather events. These variables are measured in number of people injured or dead. Whereas number of deaths has a clear cut, the degree of injury is disputable. There are injuries that really severe and injuries that are curable. Therefore the number of people may not be the best measurment to derive human costs.

For this reason every country issues the value of life in term both losses and injuries. For example in U.S. there are several studies by govermental and private organisations that attach to human casualties a monetary value.

For the purpose of this study the measurment of human casualties remains in number of people. However this issue can be revisited for a follow-up study.

1.2.3 Economic Costs

In NOAA dataset the economic impact of weather events is reported in 2 variables. One that refers to property damages and one that refer to agricultural damages. Both variables are numeric describing costs in dollars and are accompanied with another variable giving the exponent (the maginitude) of the costs.

```
# Here both variables are transformed to lower cases in order to merge different factor levels that hav
# For the property damages the unique exponents are:
unique(tolower(mydatas2$PROPDMGEXP))
```

```
##  [1] NA  "b" "k" "m" "+" "0" "5" "6" "4" "h" "2" "7" "3" "-"
```

```
# For the agricultural damages the unique exponents are:
unique(tolower(mydatas2$CROPDMGEXP))
```

```
## [1] NA  "m" "k" "b" "0"
```

As illustrated above the levels of the exponents have different coding. For that mater k may be equal to 3 representing thousands of dollar. For this reason the character codes are coerged to numeric. Moreover the variables that are represented with NA, - and + for this study are taken to be 0.

```
# First change the NAs
mydatas2$PROPDMGEXP[is.na(mydatas2$PROPDMGEXP)]<-0
mydatas2$CROPDMGEXP[is.na(mydatas2$CROPDMGEXP)]<-0

# Then + and - symbols

mydatas2$PROPDMGEXP[mydatas2$PROPDMGEXP == "+" | mydatas2$PROPDMGEXP == "-"]<-0
mydatas2$PROPDMGEXP[mydatas2$CROPDMGEXP == "+" | mydatas2$CROPDMGEXP == "-"]<-0

# And now transform all letters to numeric values
mydatas2$PROPDMGEXP<-tolower(mydatas2$PROPDMGEXP)
mydatas2$CROPDMGEXP<-tolower(mydatas2$CROPDMGEXP)

mydatas2$PROPDMGEXP[mydatas2$PROPDMGEXP == "h"]<-2
mydatas2$PROPDMGEXP[mydatas2$PROPDMGEXP == "k"]<-3
mydatas2$PROPDMGEXP[mydatas2$PROPDMGEXP == "m"]<-6
mydatas2$PROPDMGEXP[mydatas2$PROPDMGEXP == "b"]<-9

mydatas2$CROPDMGEXP[mydatas2$CROPDMGEXP == "h"]<-2
mydatas2$CROPDMGEXP[mydatas2$CROPDMGEXP == "k"]<-3
mydatas2$CROPDMGEXP[mydatas2$CROPDMGEXP == "m"]<-6
mydatas2$CROPDMGEXP[mydatas2$CROPDMGEXP == "b"]<-9

# Transform them to numeric variables
mydatas2$PROPDMGEXP<-as.numeric(mydatas2$PROPDMGEXP)
mydatas2$CROPDMGEXP<-as.numeric(mydatas2$CROPDMGEXP)

mydatas2$PROPCOST<-mydatas2$PROPDMG*(10^mydatas2$PROPDMGEXP)
mydatas2$CROPCOST<-mydatas2$CROPDMG*(10^mydatas2$CROPDMGEXP)
```

1.2.3 Event Type

The event type includes all kinds of weather events occured in U.S. and recorded in NOAA database. Although NOAA reports 48 events the levels of the factor *EVTYPE* are more(around 448). This is due to mistakes in reporting the events either spelling or tautology in classifications or both.

```
length(unique(mydatas2$EVTYPE))
```

```
## [1] 448
```

For the purpose of this analysis, I will attempt to reduce the number of classes for the weather events, where the aforementioned mistakes were introduced. However, because this process is time consuming and for the purpose of the coursera analysis this exceeds the time input that is available I won't go into details changing all the mistakes that exist.

```
# First try to tide up the symbols
EVTYPE<-tolower(mydatas2$EVTYPE)
EVTYPE <- gsub('\\\\|&|/|(-\\s)|\\s+|;',' ',EVTYPE)
EVTYPE <- gsub('\\s+',' ',EVTYPE)
EVTYPE <- gsub('^\\s','',EVTYPE)
EVTYPE <- gsub('\\s$|-$|\\.$','',EVTYPE)

# Some frequent appearing words
EVTYPE <- gsub('flooding|fldg|fld|floodin|floods','flood',EVTYPE)
EVTYPE <- gsub('flood flood','flood',EVTYPE)
EVTYPE <- gsub('flood flash','flash flood',EVTYPE)
EVTYPE <- gsub('tornados|torndao|tornadoes','tornado',EVTYPE)
EVTYPE <- gsub('tornado f.','tornado',EVTYPE)
EVTYPE <- gsub('(wintery|wintry)','winter',EVTYPE)
EVTYPE <- gsub('and','',EVTYPE)
EVTYPE <- gsub(' and$','',EVTYPE)
EVTYPE <- gsub(' \\(minor$','',EVTYPE)
EVTYPE <- gsub('\\s+|,',' ',EVTYPE)
EVTYPE <- gsub('(\\s|\\))$','',EVTYPE)
EVTYPE <- gsub('(heavy rains)|(heavy rainfall)','heavy rain',EVTYPE)
EVTYPE <- gsub('cstl','coastal',EVTYPE)
EVTYPE <- gsub('wnd','wind',EVTYPE)
EVTYPE <- gsub('winds$','wind',EVTYPE)
EVTYPE <- gsub('thunderstorms','thunderstorm',EVTYPE)
EVTYPE <- gsub('tstm wind|tstmw|tstm winds','tsunami winds',EVTYPE)
EVTYPE <- gsub('thunderstrom|thunerstorm|thundertsorm|
               thundertorm|thundestorm|thuderstorm','thunderstorm',EVTYPE)
EVTYPE <- gsub('tunderstorm|thunderestorm|thundeerstorm|
               thunderstorms|thundertorm','thunderstorm',EVTYPE)
EVTYPE <- gsub('thunderstorms','thunderstorm',EVTYPE)
EVTYPE <- gsub('thunderstorm w inds','thunderstorm wind',EVTYPE)
EVTYPE <- gsub('thunderstormw winds','thunderstorm wind',EVTYPE)
EVTYPE <- gsub('thunderstormwinds','thunderstorm wind',EVTYPE)
EVTYPE <- gsub('(thunderstorm wnd)|(thunderstorm wins)','thunderstorm wind',EVTYPE)
EVTYPE <- gsub('(thunderstorm damage)|(thunderstorm damage to)','thunderstorm',EVTYPE)
EVTYPE <- gsub('thunderstorm wind.+','thunderstorm wind',EVTYPE)
EVTYPE <- gsub('thunderstormw.+|thunderstormw','thunderstorm',EVTYPE)
EVTYPE <- gsub('wild fire','wildfire',EVTYPE)
```

```
EVTYPE <- gsub('(wintery|wintry)','winter',EVTYPE)
EVTYPE <- gsub('hvy','heavy',EVTYPE)
EVTYPE <- gsub('(sml stream)|(small strm)','small stream',EVTYPE)
EVTYPE <- gsub('unseasonal|unseasonable','unseasonal',EVTYPE)
EVTYPE <- gsub('coastalstorm','coastal storm',EVTYPE)
EVTYPE <- gsub('erosin','erosion',EVTYPE)
EVTYPE <- gsub('wild forest fire','wildfire',EVTYPE)

mydatas2$EVTYPE<-EVTYPE
length(unique(mydatas2$EVTYPE))
```

```
## [1] 311
```

1.2.4 Final Dataset

The final dataset used for this study is summarized per event type.

```
final<-ddply(mydatas2,.(EVTYPE), summarise,
             fatal = sum(FATALITIES),
             injuries = sum(INJURIES),
             propcosts = sum(PROPCOST),
             agricosts = sum(CROPCOST)
             )
```

2. Results

2.1 Question 1

Across the United States, which types of events (as indicated in the *EVTYPE* variable) are most harmful with respect to population health?

The top 10 extreme weather events and their costs in human lives and injuries

```
# Fatalities
finalHC10f<-head(final[order(-final$fatal),],10)
deaths<-ggplot(finalHC10f,aes(y=fatal,x=as.factor(EVTYPE)))+
        geom_bar(stat="identity",colour = "darkgreen", fill = "white")+
        theme_bw()+
        theme(axis.text.x = element_text(angle=90))+
        xlab("Weather Event")+
        ylab("Number of Fatalities")
# Injuries
finalHC10i<-head(final[order(-final$injuries),],10)
injurs<-ggplot(finalHC10i,aes(y=injuries,x=as.factor(EVTYPE)))+
        geom_bar(stat="identity",colour = "darkgreen", fill = "white")+
        theme_bw()+
        theme(axis.text.x = element_text(angle=90))+
        xlab("Weather Event")+
        ylab("Number of People Injured")
```

```
titleplot2<- textGrob("Impact on Human Health of the Top 10 of Exteme Weather Events in the U.S.", gp=g
grid.arrange(deaths, injurs,  ncol = 2, main = titleplot2)
```

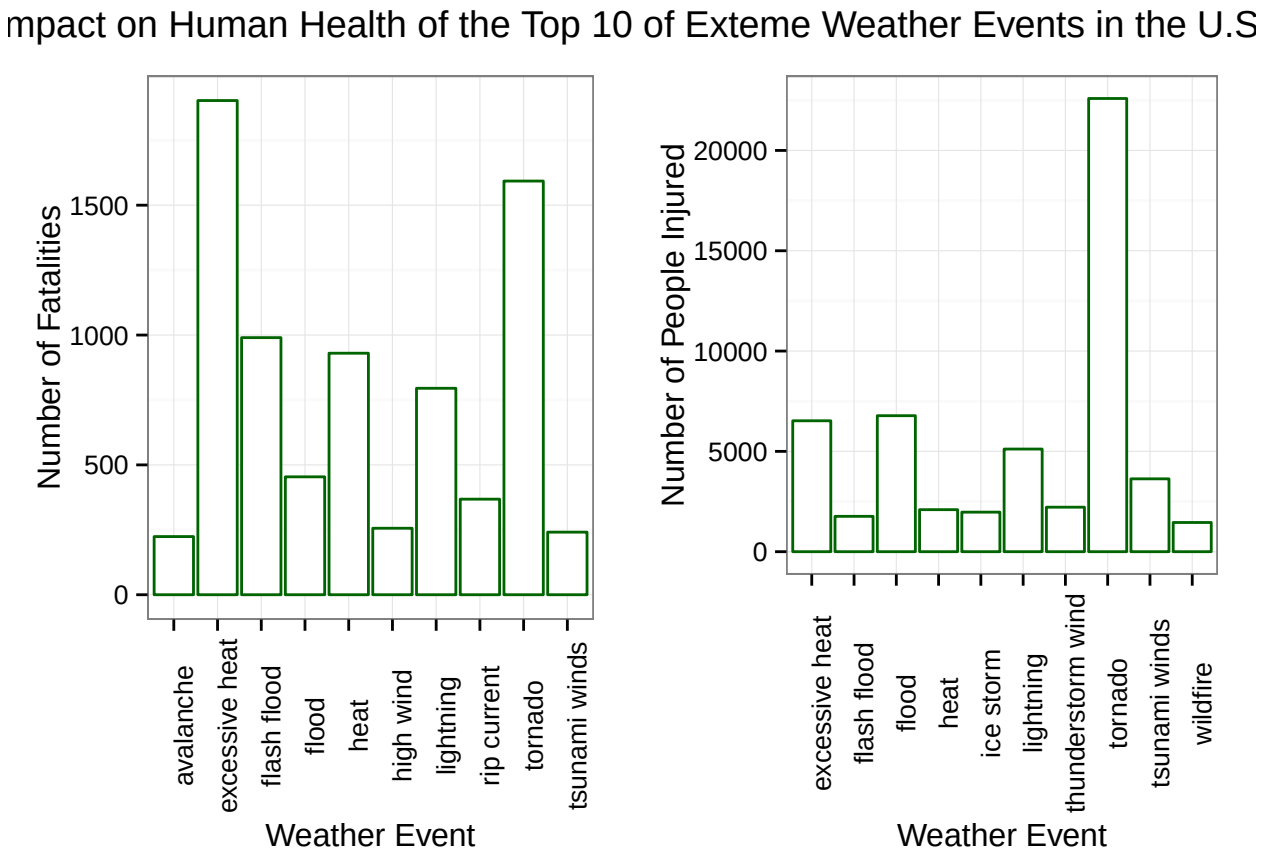

Figure 2. Top 10 of extreme weather events impact on the health of U.S. population. Please note the total number of casualties since 1994 in human lives and in number of injured people.

The above plot illustrates that the most hazardous weather event for U.S. population is the excesive heat with tornados coming to the second place. When only injuries are taken into account the tornados exhibit the most malicious weather event.

2.2 Question 2

Across the United States, which types of events have the greatest economic consequences?

The top 10 extreme weather events and their costs to U.S. economy

```
# Properties
finalE10prop<-head(final[order(-(final$propcosts)/1e9),],10)
props<-ggplot(finalE10prop,aes(y=propcosts/1e9,x=as.factor(EVTYPE)))+
        geom_bar(stat="identity",colour = "darkgreen", fill = "white")+
        theme_bw()+
        theme(axis.text.x = element_text(angle=90))+
        xlab("Weather Event")+
        ylab("Damage in ($Billion)")+
        ggtitle("Costs of Property Damages")
# Crops
```

```
finalE10crop<-head(final[order(-(final$agricosts)/1e9),],10)
crops<-ggplot(finalE10crop,aes(y=agricosts/1e9,x=as.factor(EVTYPE)))+
        geom_bar(stat="identity",colour = "darkgreen", fill = "white")+
        theme_bw()+
        theme(axis.text.x = element_text(angle=90))+
        xlab("Weather Event")+
        ylab("Damage in ($Billion)")+
         ggtitle("Costs of Agricultural Damages")


titleplot3<- textGrob("Impact on U.S. Economy of the Top 10 of Exteme Weather Events",
                      gp=gpar(fontsize=14))
grid.arrange(props, crops,  ncol = 2, main = titleplot3)
```

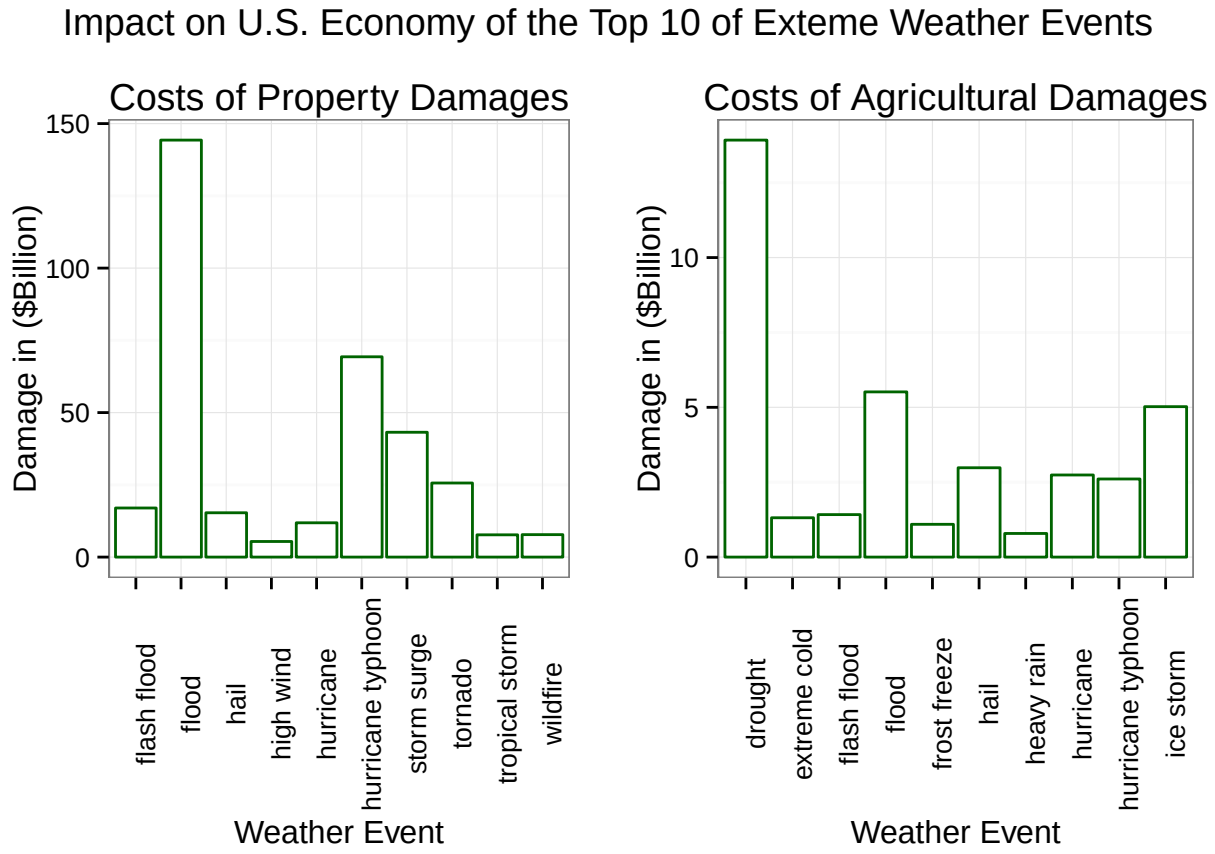

Figure 3. Top 10 of extreme weather events impact on U.S. Economy. These are the total losses in dollars of U.S. economy since 1994.

According to figure 3 U.S. economy is threaten mostly from floods and droughts for the property damages and agricultural damages respectively. It is also apparent that floods cost more than $5 billion dollars to agricultural economy.

3. Conclussions

Based on the aforementioned descriptive analysis so far for the USA the most adverse weather events for both public health and econmy are excesive heat, tornados, floods and draughts. Further investigation should

be conducted to reveal the causality of these events.