The fatal impact of tornadoes and economic effects of floods

1- Synopsis

This report illustrates through an analysis about the impacts of diversed weather events on the US Population Health & Economics.

This report downloads data from NOAA Storm Database and performs a statistical analysis on the impact of physical events to population health and economy.

Examining the event types, we observe that most of the physical phenomena cause injuries to people, which sometimes are fatal. By far, Tornadoes are the most dangerous events, caused the most number of injuries on the last 60 years.

When analysing the event types by the impact on the economy, we observe that floods and hails caused the most massive damages in the last few decades, mostly on properties.

2 - Analysis Question

A). Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?
B). Across the United States, which types of events have the greatest economic consequences?

3 - Data Processing

```
library(ggplot2)  # plot
```

```
## Warning: package 'ggplot2' was built under R version 3.1.2
```

```
# attributes
ff <- file.path(getwd(), "repdata_data_StormData.csv.bz2")
data <- read.csv(ff, stringsAsFactors = FALSE, sep=",", header=T)
```

This is followed by exploring the raw data, to have a brief understanding on the data.

```
summary(data)
```

```
##     STATE__        BGN_DATE           BGN_TIME          TIME_ZONE
##  Min.   : 1.0   Length:902297      Length:902297      Length:902297
##  1st Qu.:19.0   Class :character   Class :character   Class :character
##  Median :30.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :31.2
##  3rd Qu.:45.0
##  Max.   :95.0
##
##      COUNTY     COUNTYNAME           STATE              EVTYPE
##  Min.   :  0   Length:902297      Length:902297      Length:902297
##  1st Qu.: 31   Class :character   Class :character   Class :character
##  Median : 75   Mode  :character   Mode  :character   Mode  :character
##  Mean   :101
##  3rd Qu.:131
##  Max.   :873
##
##    BGN_RANGE      BGN_AZI           BGN_LOCATI          END_DATE
##  Min.   :   0   Length:902297      Length:902297      Length:902297
##  1st Qu.:   0   Class :character   Class :character   Class :character
##  Median :   0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :   1
##  3rd Qu.:   1
##  Max.   :3749
##
##    END_TIME           COUNTY_END COUNTYENDN       END_RANGE
##  Length:902297      Min.   :0    Mode:logical   Min.   :  0
##  Class :character   1st Qu.:0    NA's:902297    1st Qu.:  0
##  Mode  :character   Median :0                   Median :  0
##                     Mean   :0                   Mean   :  1
##                     3rd Qu.:0                   3rd Qu.:  0
##                     Max.   :0                   Max.   :925
##
##    END_AZI           END_LOCATI            LENGTH           WIDTH
##  Length:902297      Length:902297      Min.   :   0.0   Min.   :   0
##  Class :character   Class :character   1st Qu.:   0.0   1st Qu.:   0
##  Mode  :character   Mode  :character   Median :   0.0   Median :   0
##                                        Mean   :   0.2   Mean   :   8
##                                        3rd Qu.:   0.0   3rd Qu.:   0
##                                        Max.   :2315.0   Max.   :4400
##
##        F               MAG          FATALITIES     INJURIES
##  Min.   :0        Min.   :    0   Min.   :  0   Min.   :   0.0
##  1st Qu.:0        1st Qu.:    0   1st Qu.:  0   1st Qu.:   0.0
##  Median :1        Median :   50   Median :  0   Median :   0.0
##  Mean   :1        Mean   :   47   Mean   :  0   Mean   :   0.2
##  3rd Qu.:1        3rd Qu.:   75   3rd Qu.:  0   3rd Qu.:   0.0
##  Max.   :5        Max.   :22000   Max.   :583   Max.   :1700.0
##  NA's   :843563
```

```
##     PROPDMG      PROPDMGEXP           CROPDMG       CROPDMGEXP
##  Min.   :   0   Length:902297      Min.   :  0.0   Length:902297
##  1st Qu.:   0   Class :character   1st Qu.:  0.0   Class :character
##  Median :   0   Mode  :character   Median :  0.0   Mode  :character
##  Mean   :  12                      Mean   :  1.5
##  3rd Qu.:   0                      3rd Qu.:  0.0
##  Max.   :5000                      Max.   :990.0
##
##      WFO             STATEOFFIC         ZONENAMES            LATITUDE
##  Length:902297      Length:902297      Length:902297      Min.   :   0
##  Class :character   Class :character   Class :character   1st Qu.:2802
##  Mode  :character   Mode  :character   Mode  :character   Median :3540
##                                                           Mean   :2875
##                                                           3rd Qu.:4019
##                                                           Max.   :9706
##                                                           NA's   :47
##    LONGITUDE        LATITUDE_E     LONGITUDE_       REMARKS
##  Min.   :-14451   Min.   :   0   Min.   :-14455   Length:902297
##  1st Qu.:  7247   1st Qu.:   0   1st Qu.:     0   Class :character
##  Median :  8707   Median :   0   Median :     0   Mode  :character
##  Mean   :  6940   Mean   :1452   Mean   :  3509
##  3rd Qu.:  9605   3rd Qu.:3549   3rd Qu.:  8735
##  Max.   : 17124   Max.   :9706   Max.   :106220
##                   NA's   :40
##      REFNUM
##  Min.   :     1
##  1st Qu.:225575
##  Median :451149
##  Mean   :451149
##  3rd Qu.:676723
##  Max.   :902297
##
```

```
str(data)
```

```
## '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 0:00:0
0" ...
##  $ 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  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ 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  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ 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  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ 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  "" "" "" "" ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

Since the weather data was collected over a period of 60 years, with a more complete data in the later years, letâs first try to understand the frequency of the data being collected over the years from 1950 to 2011.

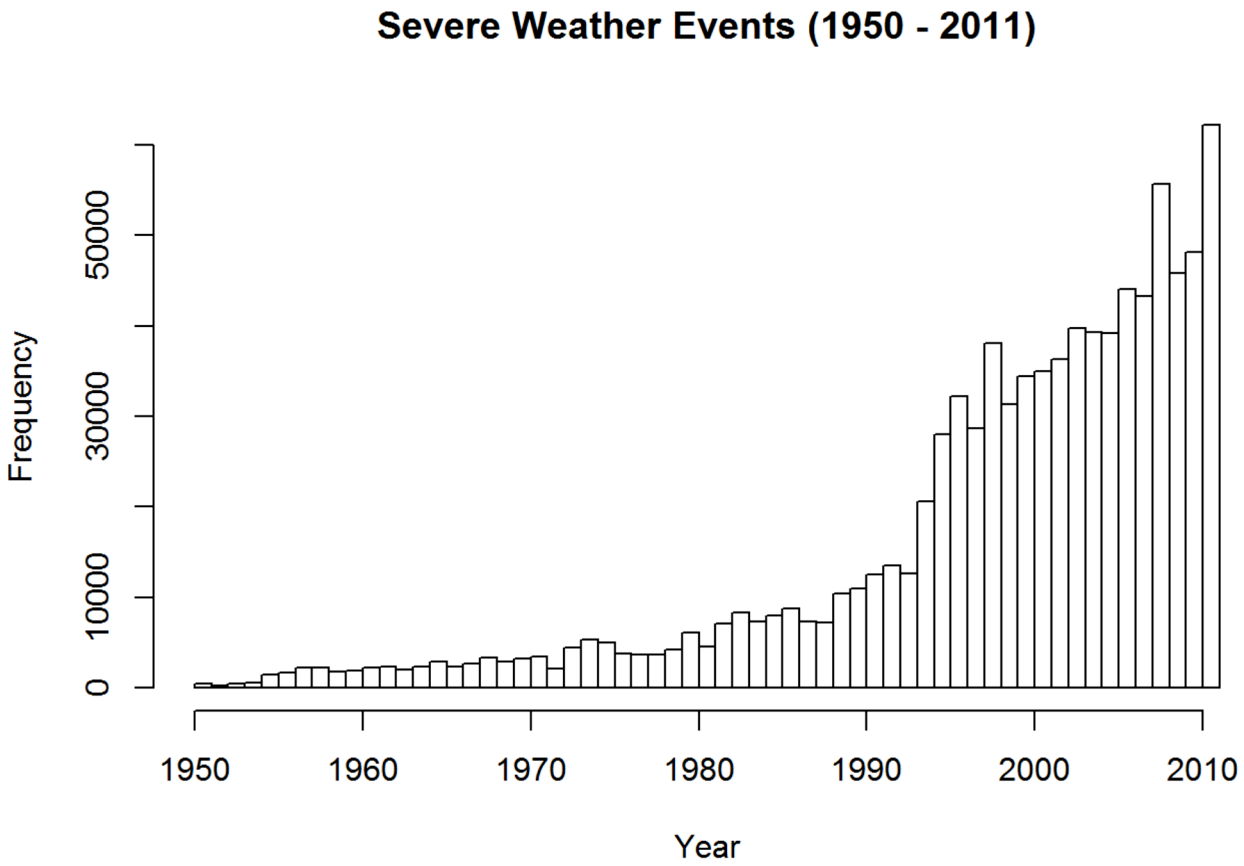
We will first trim the time format from the BGN_DATE variable.

```
data$DATE <- gsub(" 0:00:00", "", data$BGN_DATE)
data$DATE <- strptime(data$DATE, "%m/%d/%Y")
```

```
library("lubridate")
```

```
## Warning: package 'lubridate' was built under R version 3.1.2
```

```
# Check to see how the data is collected in the years
hist(year(data$DATE), main="Severe Weather Events (1950 - 2011)", xlab="Year", breaks=61)
```



From the above histogram, there are more completed records collected in recent years. The data collected from 1950 - 1970 were sparse, hence potentially unreliable. Therefore for this analysis, only the data from 1970 to 2011 will be used.

```
# Clean data
dataFocused <- data[year(data$DATE) >= 1970,]
```

There are also too many different event types in the data and will require to do some cleaning to tidy them for our data plotting later.

```
{r}# Check EVTYPE and find the general event type unique(dataFocused$EVTYPE) # Remove data with "summary"
```

By looking at the top 20 event types which have the most number of fatalities and injuries, it is needed to check whether the names of these event types are suitable.

```
# check the event type
dataEvent <- aggregate(FATALITIES~ EVTYPE, data=dataFocused, sum)
# Understand the top 10 natural disasters
head(dataEvent[order(dataEvent$FATALITIES, decreasing=TRUE),], 20)
```

```
##                     EVTYPE FATALITIES
## 834                TORNADO       3272
## 130         EXCESSIVE HEAT       1903
## 153            FLASH FLOOD        978
## 275                   HEAT        937
## 464              LIGHTNING        816
## 856              TSTM WIND        504
## 170                  FLOOD        470
## 585            RIP CURRENT        368
## 359              HIGH WIND        248
## 19               AVALANCHE        224
## 972            WINTER STORM        206
## 586            RIP CURRENTS        204
## 278              HEAT WAVE        172
## 140            EXTREME COLD        160
## 760       THUNDERSTORM WIND        133
## 310              HEAVY SNOW        127
## 141 EXTREME COLD/WIND CHILL        125
## 676             STRONG WIND        103
## 30                BLIZZARD        101
## 350              HIGH SURF        101
```

```
dataEvent <- aggregate(INJURIES~ EVTYPE, data=dataFocused, sum)
# Understand the top 10 natural disasters
head(dataEvent[order(dataEvent$INJURIES, decreasing=TRUE),], 20)
```

```
##                 EVTYPE INJURIES
## 834            TORNADO    59611
## 856          TSTM WIND     6957
## 170              FLOOD     6789
## 130     EXCESSIVE HEAT     6525
## 464          LIGHTNING     5230
## 275               HEAT     2100
## 427          ICE STORM     1975
## 153        FLASH FLOOD     1777
## 760  THUNDERSTORM WIND     1488
## 244               HAIL     1361
## 972       WINTER STORM     1321
## 411  HURRICANE/TYPHOON     1275
## 359          HIGH WIND     1137
## 310         HEAVY SNOW     1021
## 957           WILDFIRE      911
## 786 THUNDERSTORM WINDS      908
## 30            BLIZZARD      805
## 188                 FOG      734
## 955   WILD/FOREST FIRE      545
## 117         DUST STORM      440
```

Some event types are identified and grouping accordingly.

```
dataFocused$EVTYPE[grep("HEAT|WARM", dataFocused$EVTYPE)] <- "HEAT"
dataFocused$EVTYPE[grep("TORNADO", dataFocused$EVTYPE)] <- "TORNADO"
dataFocused$EVTYPE[grep("HURRICANE|TYPHOON", dataFocused$EVTYPE)] <- "HURRICANE"
dataFocused$EVTYPE[grep("FLOOD|FLD", dataFocused$EVTYPE)] <- "FLOOD"
dataFocused$EVTYPE[grep("WIND", dataFocused$EVTYPE)] <- "WIND"
dataFocused$EVTYPE[grep("AVALANC", dataFocused$EVTYPE)] <- "AVALANCHE"
dataFocused$EVTYPE[grep("SNOW", dataFocused$EVTYPE)] <- "SNOW"
dataFocused$EVTYPE[grep("STORM", dataFocused$EVTYPE)] <- "STORM"
dataFocused$EVTYPE[grep("FIRE", dataFocused$EVTYPE)] <- "FIRE"
dataFocused$EVTYPE[grep("HAIL", dataFocused$EVTYPE)] <- "HAIL"
# Iterate {exploreEvent} code to check eventType which have not catered in the above category group
ing
dataFocused$EVTYPE[grep("CURRENT|SURF|WAVE|SEA|MARINE", dataFocused$EVTYPE)] <- "COAST CONDITIONS"
dataFocused$EVTYPE[grep("COLD|WINTER|GLAZE|HYPOTHERMIA|LOW|WINTRY", dataFocused$EVTYPE)] <- "COLD"
dataFocused$EVTYPE[grep("LAND", dataFocused$EVTYPE)] <- "LANDSLIDE"
dataFocused$EVTYPE[grep("FOG", dataFocused$EVTYPE)] <- "FOG"
dataFocused$EVTYPE[grep("RAIN", dataFocused$EVTYPE)] <- "RAIN"
```

4 - Results

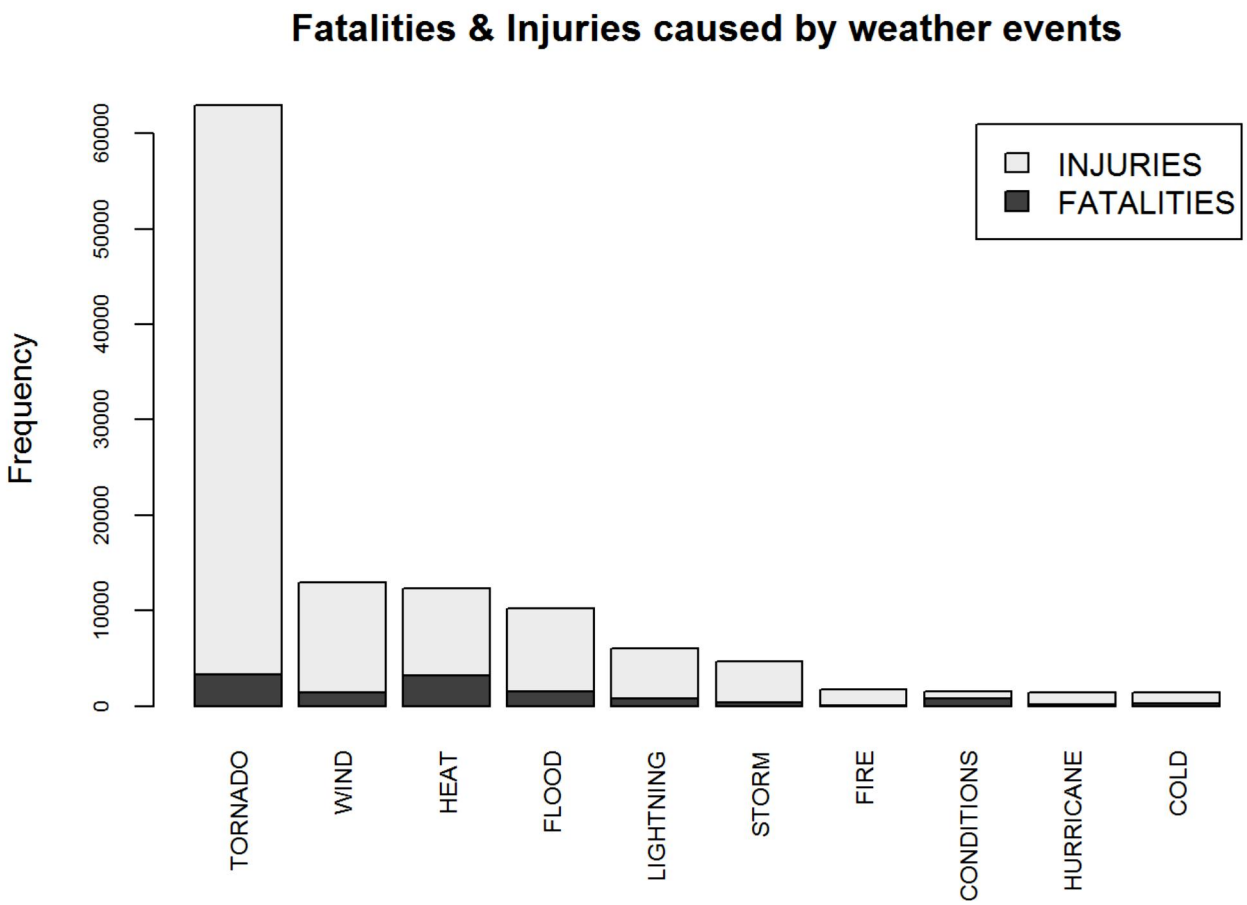
Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

All the records will be merged to include the fatalities and injuries into a dataset.

The plot figure belows show the 10 most harmful weather events with respect to population health.

```
dataEvent <- aggregate(FATALITIES~ EVTYPE, data=dataFocused, sum)
dataEvent <- merge(dataEvent, aggregate(INJURIES~ EVTYPE, data=dataFocused, sum))
dataEvent$TOTALCOUNT <- dataEvent$FATALITIES + dataEvent$INJURIES
dataEvent <- dataEvent[order(dataEvent$TOTALCOUNT, decreasing=TRUE),]
dataQ1 <- subset(dataEvent, select=c(FATALITIES, INJURIES))
dataQ1 <- t(as.matrix(dataQ1[1:10,]))

par(mfrow=c(1,1))
barplot(dataQ1, names.arg=dataEvent$EVTYPE[1:10], horiz = FALSE, las = 3, cex.names = 0.75, cex.axi
s = 0.65, offset = 0, main="Fatalities & Injuries caused by weather events", ylab="Frequency", lege
nd = rownames(dataQ1))
```

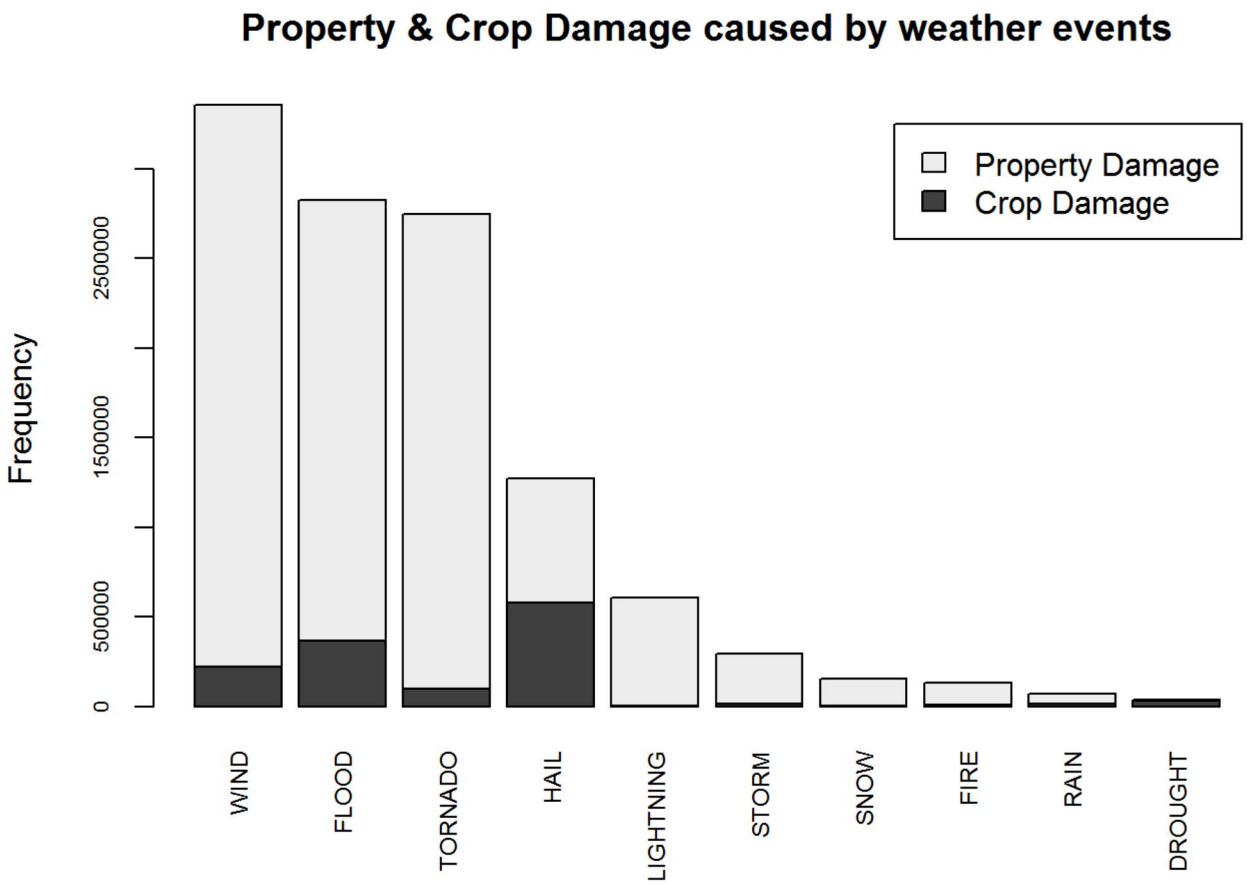


Across the United States, which types of events have the greatest economic consequences?

The plot figure belows show the top 10 weather events which have the most impact in the economic consequences.

```
dataEcon <- aggregate(CROPDMG ~ EVTYPE, data=dataFocused, sum)
dataEcon <- merge(dataEcon, aggregate(PROPDMG ~ EVTYPE, data=dataFocused, sum))
dataEcon$TOTALCOUNT <- dataEcon$PROPDMG + dataEcon$CROPDMG
dataEcon <- dataEcon[order(dataEcon$TOTALCOUNT, decreasing=TRUE),]
dataQ2 <- subset(dataEcon, select=c(CROPDMG, PROPDMG))
dataQ2 <- t(as.matrix(dataQ2[1:10,]))
rownames(dataQ2) <- c("Crop Damage", "Property Damage")

par(mfrow=c(1,1))
barplot(dataQ2, names.arg=dataEcon$EVTYPE[1:10], horiz = FALSE, las = 3, cex.names = 0.75, cex.axis
 = 0.65, offset = 0, main="Property & Crop Damage caused by weather events", ylab="Frequency", lege
nd = rownames(dataQ2))
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.