Examination of the public health and economic consequences of storm events across the United States

Ozan Aygun

November 10th, 2016

Synopsis

Here I present the detailed analysis of the data provided by the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database to determine the storm events that cause highest economic and public health consequences in the United States. The results presented here suggests that tornadoes are the most harmful weather event that is associated with over 5000 death and 75000 injuries across the U.S between 1950 and 2011. On the other hand, flood is determined as the leading severe weather event that is linked to the highest property damage, costing approximately 150 billion U.S dollars in the same period. In terms of U.S. agriculture, drought has a bigger impact on the crop damage, resulting in a loss of nearly 15 billion U.S dollars. Finally, I also present the geospatial distribution of these severe weather events to demonstrate their relative impact on public health and economy across different states. These observations might facilitate the decision making of government authorities and allow prioritization of resources to prepare for different types of severe weather events.

Data Processing

Data

The data used in this analysis is obtained from the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, which records the severe weather and storm events between 1950 and 2011, along with public health and economic consequences associated with them. Detailed information of the dataset can be obtained in the National Weather Service Storm Data Documentation, and the data set can be downloaded here.

Reading and cleaning the data set

Downloading and reading the data:

```r
fileURL = "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
if(!exists("./storm_data.csv")){
download.file(fileURL,destfile = "./storm_data.csv")}
storm_data <- read.csv("./storm_data.csv", stringsAsFactors = FALSE)
```

When inspecting the **EVTYPE** variable,which represents the types of events recorded, I noticed that there are several redundancies related to inconsistent data entry.

```r
head(levels(factor(storm_data$EVTYPE)),15)
```

```
##  [1] "   HIGH SURF ADVISORY" " COASTAL FLOOD"
##  [3] " FLASH FLOOD"          " LIGHTNING"
##  [5] " TSTM WIND"            " TSTM WIND (G45)"
##  [7] " WATERSPOUT"           " WIND"
##  [9] "?"                     "ABNORMAL WARMTH"
## [11] "ABNORMALLY DRY"        "ABNORMALLY WET"
## [13] "ACCUMULATED SNOWFALL"  "AGRICULTURAL FREEZE"
## [15] "APACHE COUNTY"
```

```
tail(levels(factor(storm_data$EVTYPE)),15)
```

```
##  [1] "WINTER MIX"              "WINTER STORM"
##  [3] "WINTER STORM HIGH WINDS" "WINTER STORM/HIGH WIND"
##  [5] "WINTER STORM/HIGH WINDS" "WINTER STORMS"
##  [7] "Winter Weather"          "WINTER WEATHER"
##  [9] "WINTER WEATHER MIX"      "WINTER WEATHER/MIX"
## [11] "WINTERY MIX"             "Wintry mix"
## [13] "Wintry Mix"              "WINTRY MIX"
## [15] "WND"
```

Therefore, I clean the data for this variable.

```
storm_data$EVTYPE <- trimws(storm_data$EVTYPE, which = c("both"))
storm_data$EVTYPE <- toupper(storm_data$EVTYPE)

events <- length(levels(as.factor(storm_data$EVTYPE)))
unknownevents <- sum(which(storm_data$EVTYPE == "?"))
```

I noticed that there are 890 event type combinations remained after cleaning this variable. One of these categories is “?”, representing unspecified severe weather events. There are 246124 unspecified events recorded in the dataset, which will be included in this analysis.

Results

In the this analysis, I sought to find answers to the following important questions:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

Below I present the results of the detailed analysis along with the associated code, which is aimed to answer these interesting questions.

Tornadoes are the most harmful severe weather events across the United States

Public health impact is represented by FATALITIES and INJURIES variables in the data set. Death is the most important public health outcome, therefore it takes the highest priority.

In order to answer this question, I first explored the relationship between the frequency of different types of events and the total number of deaths. Afterwards, I investigated the events that are most frequent, and those associated with the highest number of death and injuries. **Figure 1** summarizes these observations.

```
library(dplyr)
storm_events_health <- data.frame(storm_data %>% group_by(EVTYPE) %>% summarise(count= n(),
                Total_fatalities=sum(FATALITIES),
                Total_injuries = sum(INJURIES)))
# Top 10 most frequent storm events:

library(ggplot2)
storm_events_health$EVTYPE = factor(storm_events_health$EVTYPE)

A <- ggplot(storm_events_health, aes(x= log10(count), y= log10(Total_fatalities)))+
        geom_point()+
        labs(y ="Total Fatalities across the U.S. (Log10)",
             x = "The number of events observed (Log10)",
             title = "(A) Fatalities increase with the frequency of events ")+
        geom_smooth(method = "lm")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9))


counts <- storm_events_health %>% arrange(desc(count))
B<- ggplot(data = head(counts,10),aes(x=log10(count),
                                  y=reorder(EVTYPE,count)))+
        geom_point(size = 3) +
        labs(y ="",
             x = "The number of events observed (Log10)",
             title = "(B) 10 most frequent types of storm
events observed across the U.S. ")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9),
              panel.grid.major.x = element_blank(),
              panel.grid.minor.x = element_blank(),
              panel.grid.major.y = element_line(color='grey60', linetype='dashed'))


fatalities <- storm_events_health %>% arrange(desc(Total_fatalities))
C <- ggplot(data = head(fatalities,10),aes(x=Total_fatalities,
                                               y=reorder(EVTYPE,Total_fatalities)))+
        geom_point(size = 3, color = "red") +
        labs(y ="",
             x = "Total number of fatalities",
             title = "(C) 10 most deadly storm events
             across the U.S. ")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9),
              panel.grid.major.x = element_blank(),
              panel.grid.minor.x = element_blank(),
              panel.grid.major.y = element_line(color='grey60', linetype='dashed'))

injuries <- storm_events_health %>% arrange(desc(Total_injuries))
D<- ggplot(data = head(injuries,10),aes(x=Total_injuries,
                                               y=reorder(EVTYPE,Total_injuries)))+
        geom_point(size = 3, color= "purple") +
        labs(y ="",
             x = "Total number of injuries",
             title = "(D) Storm events that are causing the
highest number of injuries across the U.S. ")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9),
              panel.grid.major.x = element_blank(),
              panel.grid.minor.x = element_blank(),
              panel.grid.major.y = element_line(color='grey60', linetype='dashed'))
```

```r
# To combine the ggplots, use the multiplot function
# from : http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
        library(grid)

        # Make a list from the ... arguments and plotlist
        plots <- c(list(...), plotlist)

        numPlots = length(plots)

        # If layout is NULL, then use 'cols' to determine layout
        if (is.null(layout)) {
                # Make the panel
                # ncol: Number of columns of plots
                # nrow: Number of rows needed, calculated from # of cols
                layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                                 ncol = cols, nrow = ceiling(numPlots/cols))
        }

        if (numPlots==1) {
                print(plots[[1]])

        } else {
                # Set up the page
                grid.newpage()
                pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

                # Make each plot, in the correct location
                for (i in 1:numPlots) {
                        # Get the i,j matrix positions of the regions that contain this subplot
                        matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

                        print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                                        layout.pos.col = matchidx$col))
                }
        }
}


multiplot(A,B,C,D, cols = 2)
```

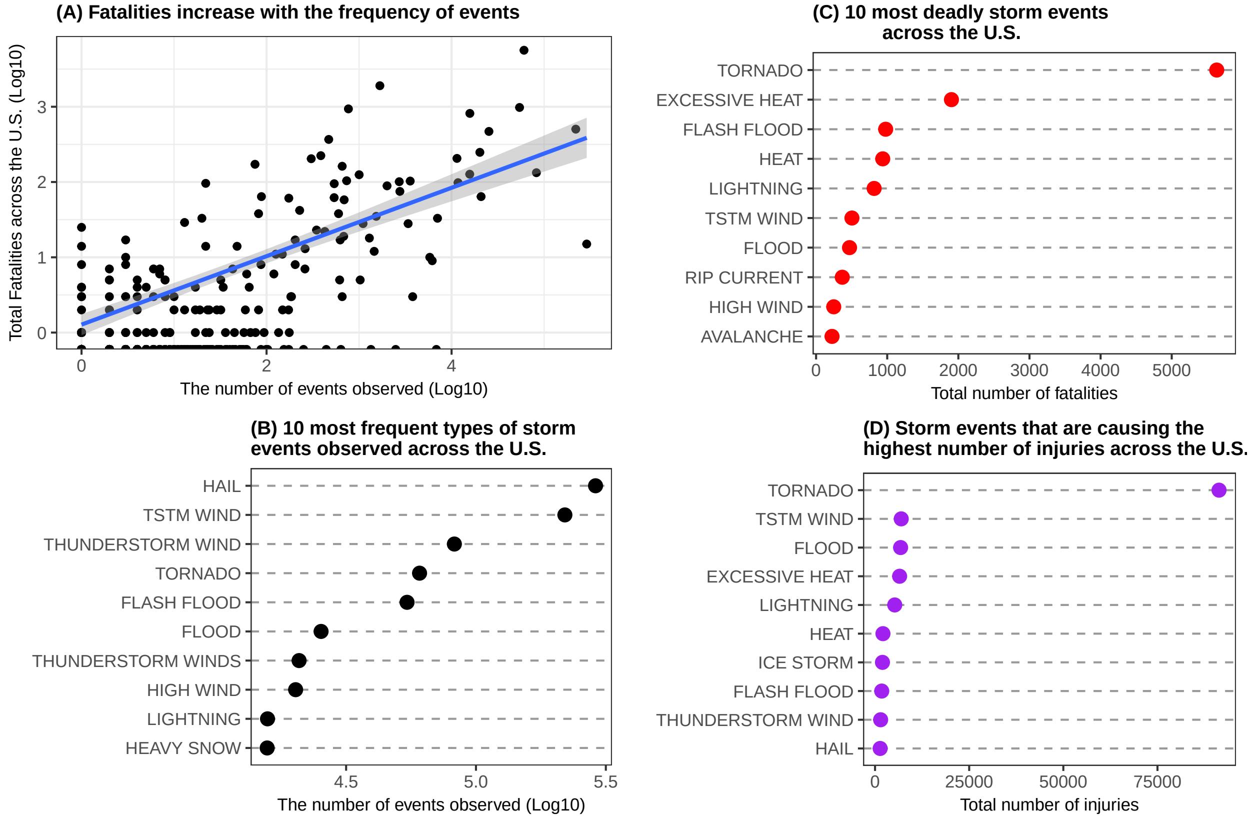

FIGURE 1

As illustrated in **FIGURE 1A**, the number of fatalities increases with the number of severe events observed, suggesting that the great portion of deaths are not associated with rare but devastating events. In other words, the severe events that are associated with the highest number of fatalities are those have been recorded the most frequently in the U.S.

When the most frequently occurring events are explored, hail appears as the leading event, followed by thunderstorms and tornadoes **(FIGURE 1B)**. Importantly, tornadoes are the most harmful weather event that is associated with over 5000 death and 75000 injuries across the U.S between 1950 and 2011 **(FIGURE 1C and 1D)**.

Therefore, these observations collectively indicate that tornadoes are the leading severe weather events that impact on the public health in the U.S.

Flood and Drought have the greatest economic consequences for the United States

The economic consequences are reflected in the data set by two variables:

1. PROPDMG
2. CROPDMG

The related PROPDMGEXP and CROPDMGEXP variables carry the abbreviations that note the magnitude of the monetary loss (K: thousands, M: millions, B: billions)

Both of these variables contain entries that are not defined by the data codebook given in the National Weather Service Storm Data Documentation:

```
table(storm_data$PROPDMGEXP)
```

```
## 
##             -      ?      +      0      1      2      3      4      5 
## 465934      1      8      5    216     25     13      4      4     28 
##      6      7      8      B      h      H      K      m      M 
##      4      5      1     40      1      6 424665      7  11330 
```

```r
table(storm_data$CROPDMGEXP)
```

```
## 
##             ?      0      2      B      k      K      m      M 
## 618413      7     19      1      9     21 281832      1   1994 
```

Note that the entries that contain no abbreviations have no monetary loss associated with them.

```r
library(dplyr)
storm_data %>%
        filter(PROPDMGEXP == " ") %>%
        summarize(property_damage = sum(PROPDMG,na.rm = TRUE))
```

```
##   property_damage
## 1               0
```

```r
storm_data %>%
        filter(CROPDMGEXP == " ") %>%
        summarize(crop_damage = sum(CROPDMG,na.rm = TRUE))
```

```
##   crop_damage
## 1           0
```

Therefore, I further cleaned up the data, so that it retains only K, M, B abbreviations that are approved by the agency. These entries represent the majority of the monetary loss that is associated with the severe weather events recorded.

In order to answer this question, I first explored the relationship between the frequency of different types of events and the economic consequences. Afterwards, I investigated the events that are associated with the highest crop and property damage. **Figure 2** summarizes these observations.

```r
# Property damage:
library(dplyr)
storm_events_propdmg <- storm_data %>%
        filter(PROPDMGEXP == "K" |PROPDMGEXP == "M" | PROPDMGEXP == "B")

K <- which(storm_events_propdmg$PROPDMGEXP == "K")
M <- which(storm_events_propdmg$PROPDMGEXP == "M")
B <- which(storm_events_propdmg$PROPDMGEXP == "B")

#Calculating the exact property damage:
storm_events_propdmg$PROPD[K] <- storm_events_propdmg$PROPDMG[K]*(10^3)
storm_events_propdmg$PROPD[M] <- storm_events_propdmg$PROPDMG[M]*(10^6)
storm_events_propdmg$PROPD[B] <- storm_events_propdmg$PROPDMG[B]*(10^9)

# Summarizing the total property damage caused by each storm event
storm_events_propdmg_summary <- storm_events_propdmg %>%
        group_by(EVTYPE) %>%
        summarise(count = n(),Total_PROPD = sum(PROPD, na.rm= TRUE)) %>%
        arrange(desc(Total_PROPD))

storm_events_propdmg_summary$millon_damage <- (storm_events_propdmg_summary$Total_PROPD) /10^6
```

```
# Crop damage:
storm_events_cropdmg <- storm_data %>%
        filter(CROPDMGEXP == "K" |CROPDMGEXP == "M" | CROPDMGEXP == "B")

K <- which(storm_events_cropdmg$CROPDMGEXP == "K")
M <- which(storm_events_cropdmg$CROPDMGEXP == "M")
B <- which(storm_events_cropdmg$CROPDMGEXP == "B")

#Calculating the exact crop damage:
storm_events_cropdmg$CROPD[K] <- storm_events_cropdmg$CROPDMG[K]*(10^3)
storm_events_cropdmg$CROPD[M] <- storm_events_cropdmg$CROPDMG[M]*(10^6)
storm_events_cropdmg$CROPD[B] <- storm_events_cropdmg$CROPDMG[B]*(10^9)

# Summarizing the total property damage caused by each storm event
storm_events_cropdmg_summary <- storm_events_cropdmg %>%
        group_by(EVTYPE) %>% summarise(count = n(),
                Total_CROPD = sum(CROPD, na.rm = TRUE)) %>%
                arrange(desc(Total_CROPD))

storm_events_cropdmg_summary$millon_damage <- (storm_events_cropdmg_summary$Total_CROPD) /10^6

A <- ggplot(storm_events_propdmg_summary, aes(x= log10(count), y= log10(Total_PROPD)))+
        geom_point()+
        labs(y ="Total property damage across the U.S.($) (Log10)",
             x = "The number of events observed (Log10)",
             title = "(A) Property damage increases with the frequency of events ")+
        geom_smooth(method = "lm")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9))

B <- ggplot(storm_events_cropdmg_summary, aes(x= log10(count), y= log10(Total_CROPD)))+
        geom_point()+
        labs(y ="Total crop damage across the U.S.($) (Log10)",
             x = "The number of events observed (Log10)",
             title = "(B) Crop damage increases with the frequency of events ")+
        geom_smooth(method = "lm")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9))


C<- ggplot(data = head(storm_events_propdmg_summary,10),aes(x= millon_damage,
                                    y=reorder(EVTYPE,millon_damage)))+
        geom_point(size = 3, color= "blue") +
        labs(y ="",
             x = "Total property damage across the U.S.(million $)",
             title = "(C) Storm events that are causing the
             highest property damage across the U.S. ")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9),
              axis.text.x = element_text(hjust = TRUE, angle = 45),
              panel.grid.major.x = element_blank(),
              panel.grid.minor.x = element_blank(),
              panel.grid.major.y = element_line(color='grey60', linetype='dashed'))

D<- ggplot(data = head(storm_events_cropdmg_summary,10),aes(x=millon_damage, y=reorder(EVTYPE,millon_da
        geom_point(size = 3, color= "blue") +
        labs(y ="",
             x = "Total crop damage across the U.S.(million $)",
             title = "(D) Storm events that are causing the
             highest crop damage across the U.S. ")+
        theme_bw()+
        theme(plot.title = element_text(size = 10,face = "bold"),
              axis.title = element_text(size = 9),
              axis.text.x = element_text(hjust = TRUE, angle = 45),
              panel.grid.major.x = element_blank(),
              panel.grid.minor.x = element_blank(),
              panel.grid.major.y = element_line(color='grey60', linetype='dashed'))

multiplot(A,B,C,D, cols = 2)
```

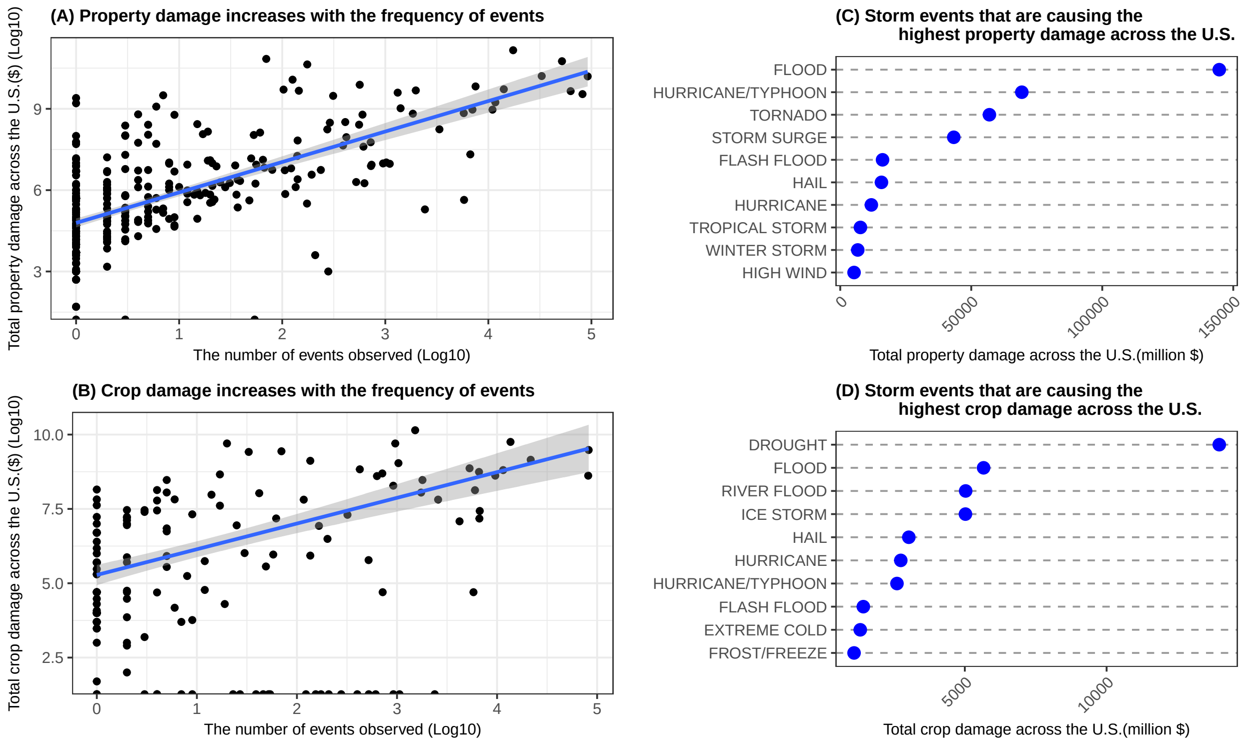

FIGURE 2

As illustrated in **FIGURE 2A and FIGURE 2B**, both property and crop damage increases with the frequency of the events that are recorded. This relationship suggests that similar to the public health consequences,the greatest amount of the economic damage is not explained by rarely occurring severe events. Although there has been two rare events that occurred only once and caused over 1 billion $ property damage **(FIGURE 2A)**,there are far more frequent events that caused more severe economic consequences **(FIGURES 2C and 2D)**.

My analyses illustrate that flood is associated with the highest property damage in the United States, resulting in over 150 billion $ monetary loss between 1950 and 2011 **(FIGURE 2C)**. There have been over 17 thousand flood events recorded in the database. Although flood is the leading factor for the property damage, drought appears as the most harmful severe weather event for U.S. agriculture, resulting in a crop damage that equals to the loss of approximately 15 billion $ **(FIGURE 2C)**.

Therefore, these observations collectively indicate that flood and drought are the leading severe weather events that impact on the U.S. economy, leading highest property and crop damage, respectively.

Geospatial distribution of leading severe events reveals patterns of economic and public health consequences

The analyses presented so far highlight that tornadoes, flood and drought are the most harmful severe weather events for the public health and U.S. economy. In order to facilitate the allocation of resources to get better prepared for these events, it is also important to understand the relative distribution of these events across the country. **FIGURE 3** summarizes the geospatial distribution of these severe weather events.

```
library(openintro)
library(dplyr)
library(ggplot2)
library(maps)

us <- map_data("state") # convert state spatial data into a ggplot map_data
# note that this contains 49 states, the matching data above has to be filtered to present only these 4


storm_data$statename <- tolower(abbr2state(storm_data$STATE)) # convert state abbreviations to full nam

tornado_states_summary <- storm_data %>%
filter(EVTYPE == "TORNADO" & statename %in% levels(factor(us$region))) %>%
group_by(statename) %>% summarise(count = n(),
        Total_injuries = sum(INJURIES, na.rm= T),
        Total_fatalities = sum (FATALITIES, na.rm= T))

storm_events_propdmg$statename <-tolower(abbr2state(storm_events_propdmg$STATE))

flood_states_summary <- storm_events_propdmg %>%
        filter(EVTYPE == "FLOOD" & statename %in% levels(factor(us$region))) %>%
        group_by(statename) %>% summarise(count = n(),
        Total_property_damage_million_USD = (sum(PROPD, na.rm= T))/10^6)


storm_events_cropdmg$statename <-tolower(abbr2state(storm_events_cropdmg$STATE))

drought_states_summary <- storm_events_cropdmg %>%
filter(EVTYPE == "DROUGHT" & statename %in% levels(factor(us$region))) %>%
group_by(statename) %>% summarise(count = n(),
Total_crop_damage_million_USD = (sum(CROPD, na.rm= T))/10^6)

statenames <- data.frame(statename=levels(factor(us$region)))
drought_states_summary <- merge(drought_states_summary,statenames,
                                by="statename", all = T)
```

```r
gg <- ggplot() # make an empty plot

# first add the us map_data as the plain U.S. map frame
template <- gg + geom_map(data = us,map = us, aes(x = long, y = lat,
map_id = region ),fill = "#ffffff", color = "#ffffff", size = 0.15)

# then add the actual data and provide a nicely contrasted scale_fill_continuous

# Tornado-related death across the U.S
A <- template + geom_map (data = tornado_states_summary, map = us,
                     aes(fill = Total_fatalities, map_id = statename),
                     color="#ffffff", size = 0.3) +
        scale_fill_continuous(low='thistle2', high='darkred',
                              guide='colorbar', name = "Number of\nfatalities")

# The rest is to polish the map:
A <- A + labs(x=NULL, y=NULL,
title = "(A) Distribution of fatalities caused by tornadoes
across the United States")+
 coord_map("albers", lat0 = 39, lat1 = 45)+
 theme(panel.border = element_blank())+
 theme(panel.background = element_blank())+
 theme(axis.ticks = element_blank())+
 theme(axis.text = element_blank())+
 theme(title = element_text(face = "bold",size = 8))

# The number of recorded tornadoe events across the U.S
B <- template + geom_map (data = tornado_states_summary, map = us,
                          aes(fill = count, map_id = statename),
                          color="#ffffff", size = 0.3) +
        scale_fill_continuous(low='thistle2', high='purple4',
                              guide='colorbar',
                              name = "Number of\ntornadoes recorded")

# The rest is to polish the map:
B <- B + labs(x=NULL, y=NULL,
title = "(B) Distribution of the number of tornadoes recorded
across the United States")+
        coord_map("albers", lat0 = 39, lat1 = 45)+
        theme(panel.border = element_blank())+
        theme(panel.background = element_blank())+
        theme(axis.ticks = element_blank())+
        theme(axis.text = element_blank())+
        theme(title = element_text(face = "bold",size = 8))

# Flood-related property damage across the U.S:
C <- template + geom_map (data = flood_states_summary, map = us,
                          aes(fill = log10(Total_property_damage_million_USD), map_id = statename),
                          color="#ffffff", size = 0.3) +
        scale_fill_continuous(low='lightsteelblue2', high='navy',
                              guide='colorbar', name = "Property damage\nMillion $ (Log10)")

# The rest is to polish the map:
```

```
C <- C + labs(x=NULL, y=NULL,
              title = "(C) Distribution of flood-related property damage\nacross the United States")+
        coord_map("albers", lat0 = 39, lat1 = 45)+
        theme(panel.border = element_blank())+
        theme(panel.background = element_blank())+
        theme(axis.ticks = element_blank())+
        theme(axis.text = element_blank())+
        theme(title = element_text(face = "bold",size = 8))

# Drought-related crop damage across the U.S:
D <- template + geom_map (data = drought_states_summary, map = us,
                          aes(fill = log10(Total_crop_damage_million_USD), map_id = statename),color="#
        scale_fill_continuous(low='peachpuff', high='darkorange2',
                              guide='colorbar', name = "Crop damage\nMillion $")

# The rest is to polish the map:
D <- D + labs(x="Grey: no drought-related crop damage recorded", y=NULL,
title = "(D) Distribution of drought-related crop damage
across the United States")+
        coord_map("albers", lat0 = 39, lat1 = 45)+
        theme(panel.border = element_blank())+
        theme(panel.background = element_blank())+
        theme(axis.ticks = element_blank())+
        theme(axis.text = element_blank())+
        theme(title = element_text(face = "bold",size = 8))

multiplot(A,C,B,D, cols = 2)
```

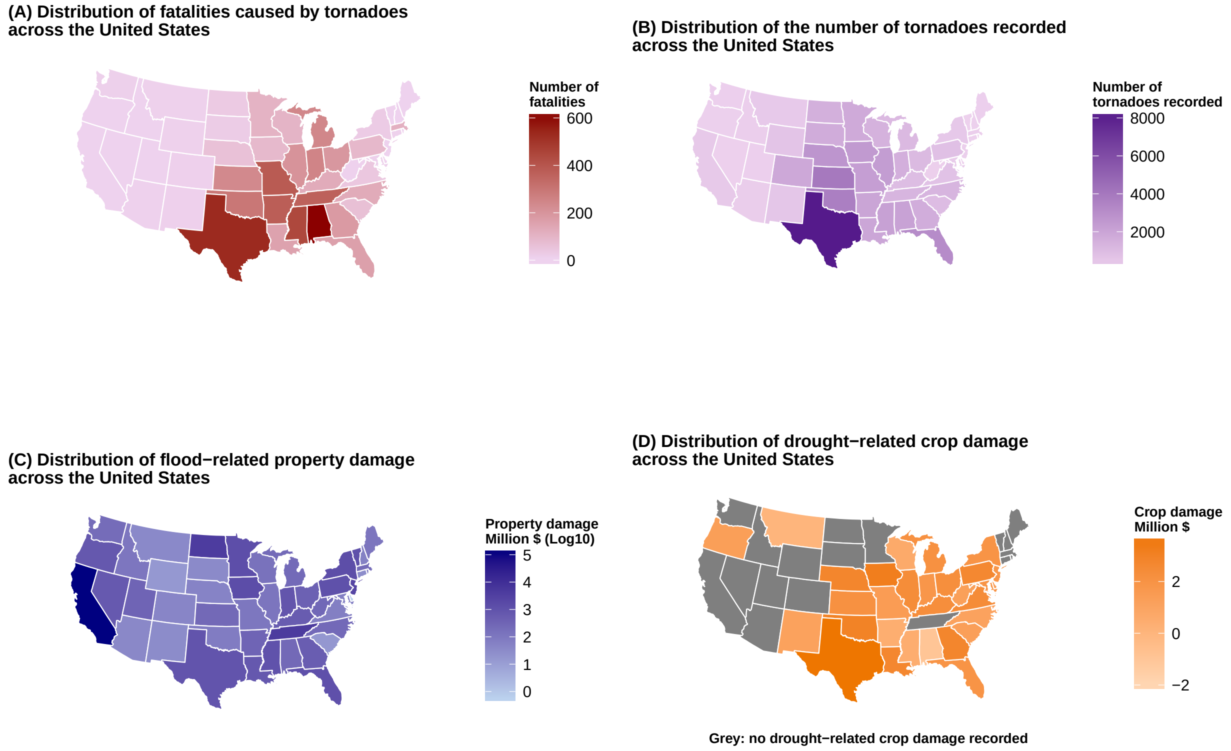
(A) Distribution of fatalities caused by tornadoes across the United States
Number of fatalities
600
400
200
0
(B) Distribution of the number of tornadoes recorded across the United States
Number of tornadoes recorded
8000
6000
4000
2000
(C) Distribution of flood−related property damage across the United States
Property damage Million $ (Log10)
5
4
3
2
1
0
(D) Distribution of drought−related crop damage across the United States
Crop damage Million $
2
0
−2
Grey: no drought−related crop damage recorded

FIGURE 3 Tornado-related deaths are quite prominent in southern states, particularly in Texas and Alabama **(FIGURE 3A)**. When compared with the distribution of the number of tornadoes **(FIGURE 3B)**, it is anticipated that Texas would suffer from the most damage, given that most of the recorded tornadoes occurred in this state. Surprisingly however, Alabama has similar number of tornado-related deaths, even though it hasn't observed as many tornadoes as Texas **(FIGURES 3A and 3B)**. Furthermore, although Florida and Kansas experienced more tornadoes than Alabama, they had less number of tornado related deaths **(FIGURES 3A and 3B)**. These observations suggest that Alabama is a particularly vulnerable state against tornadoes. It is tempting to speculate that Alabama might have experienced less frequent but more stronger tornadoes that resulted in more deaths. Alternatively, the society or infrastructure in Alabama might be less prepared to tornadoes compared to Florida and Kansas. The exact reasons behind this phenomenon requires further investigation and additional data. In any case Alabama appears as a state that requires particular attention with respect to the impact of tornadoes on public health.

Economic consequences of severe weather events also reveal interesting patterns across the country **(FIGURES 3C and 3D)**. California has suffered the greatest flood-associated property damage, which is more than 10 billion $ **(FIGURE 3C)**. On the other hand, Texas, Georgia and Iowa experienced the highest amount of crop damage that is associated with droughts **(FIGURE 3D)**.

Conclusions

The analyses illustrated here provides an important resource for further investigation of the weather events that are harmful for public health and have severe economic consequences. I hope this would enable government officials to allocate and prioritize the required resources for the most fragile states in order to get better prepared for these natural events.