NOAA Storm DB Analysis

nuno r

January 28, 2020

Instructions

Document Layout Language: Your document should be written in English. Title: Your document should have a title that briefly summarizes your data analysis Synopsis: Immediately after the title, there should be a synopsis which describes and summarizes your analysis in at most 10 complete sentences.

There should be a section titled Data Processing which describes (in words and code) how the data were loaded into R and processed for analysis. In particular, your analysis must start from the raw CSV file containing the data. You cannot do any preprocessing outside the document. If preprocessing is time-consuming you may consider using the cache = TRUE option for certain code chunks.

There should be a section titled Results in which your results are presented. You may have other sections in your analysis, but Data Processing and Results are required. The analysis document must have at least one figure containing a plot.

Your analysis must have no more than three figures. Figures may have multiple plots in them (i.e. panel plots), but there cannot be more than three figures total. You must show all your code for the work in your analysis document. This may make the document a bit verbose, but that is okay. In general, you should ensure that echo = TRUE for every code chunk (this is the default setting in knitr).

Your data analysis must address the following questions:

Q1) Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

Q2) Across the United States, which types of events have the greatest economic consequences?

Consider writing your report as if it were to be read by a government or municipal manager who might be responsible for preparing for severe weather events and will need to prioritize resources for different types of events. However, there is no need to make any specific recommendations in your report.

Introduction

This report analizes data from NOAA Storm DB and identifies the impacts on health and damages to properties and crops. We will demonstrate the top 10 events with the most impact on health and an aggregate of all damages to properties and crops. We expect anyone reading this document will be able to analyze the data and process it to reach similar results.

Data Processing

Let's load and prepare the data for analysis.

Load dplyr to manipulate data

```
# Additional information can be found here:
# https://dplyr.tidyverse.org/
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Load ggplot2 to create the charts

```
# Additinal information can be found here:
# https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

Load the NOAA DB file directly into the workspace with no external manipulations

```
# NOTE: The original NOAA DB file can be found here:
# https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2
noaa_db_url = "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
download.file(noaa_db_url, dest = "StormData.csv.bz2")
```

Read the Storm Data DB extract into a table

```
# NOTE: additional info on read.table:
# https://stat.ethz.ch/R-manual/R-devel/library/utils/html/read.table.html
noaa_table <- read.table("StormData.csv.bz2",
                         sep = ",",
                         header = TRUE,
                         quote = "\"'",
                         dec = ".",
                         numerals = c("allow.loss",
                                      "warn.loss",
                                      "no.loss"))

# Find the names of the columns in the dataset
names(noaa_table)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

We don't need all columns. Pick the necessary columns for further analysis

```
accidents <- noaa_table[,c(8,23:24)] # include EVTYPE, FATALITIES, INJURIES
property_damages <- noaa_table[,c(8,25:28)] # include EVTYPE, PROPFMG, PROPDMGEXP, CROPDMG, CROPDMGEXP
```

Used only for debugging, too much information otherwise

```
#NOTE: used only for debugging, too much information otherwise
#print(accidents)
#print(property_damages)
```

Identify the highest incidence of events, sorted by fatalities and injuries that did not result in a fatality

```
top_display_events = 10
top.accidents <- aggregate(cbind(INJURIES,FATALITIES) ~ EVTYPE,
                           data = accidents,
                           sum,
                           na.rm=TRUE)

top.accidents <- arrange(top.accidents,
                         desc(INJURIES + FATALITIES))

# pick the top event types
top.accidents <- top.accidents[1:top_display_events,]
top.accidents
```

```
##               EVTYPE INJURIES FATALITIES
## 1            TORNADO    91346       5633
## 2     EXCESSIVE HEAT     6525       1903
## 3          TSTM WIND     6957        504
## 4              FLOOD     6789        470
## 5          LIGHTNING     5230        816
## 6               HEAT     2100        937
## 7        FLASH FLOOD     1777        978
## 8          ICE STORM     1975         89
## 9  THUNDERSTORM WIND     1488        133
## 10      WINTER STORM     1321        206
```

Results

Q1) Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

Prepare the fatalities data to be visualized

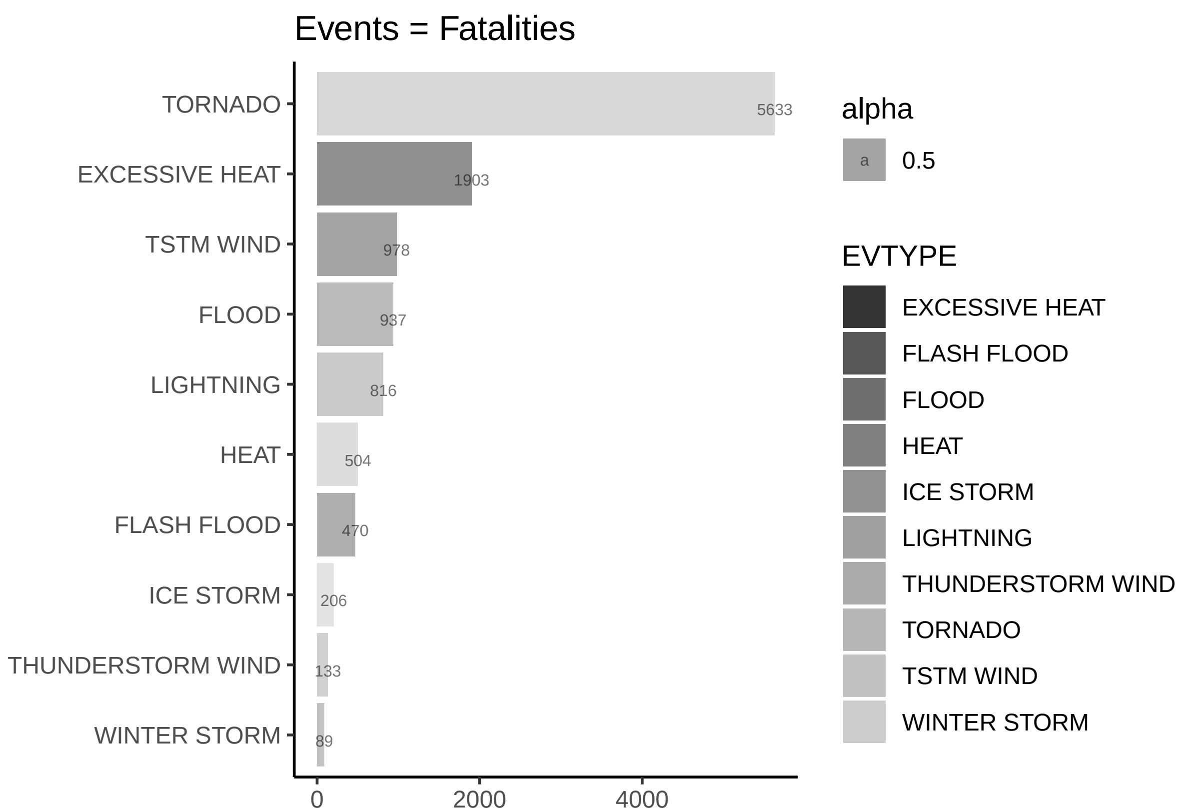
```
fatalities_chart = ddply(top.accidents,
                         .(EVTYPE),
                         summarize,
                         sum_fatalities = sum(FATALITIES,na.rm=TRUE))

fatalities_chart = fatalities_chart[order(fatalities_chart$sum_fatalities, decreasing = TRUE), ]
head(fatalities_chart,top_display_events)
```

```
##               EVTYPE sum_fatalities
## 8            TORNADO           5633
## 1     EXCESSIVE HEAT           1903
## 2        FLASH FLOOD            978
## 4               HEAT            937
## 6          LIGHTNING            816
```

```
## 9          TSTM WIND            504
## 3              FLOOD            470
## 10      WINTER STORM            206
## 7  THUNDERSTORM WIND            133
## 5          ICE STORM             89
```

Events that resulted in fatalities

```r
ggplot(fatalities_chart[1:top_display_events, ],
       aes(EVTYPE,
           x = reorder(top.accidents$EVTYPE, sum_fatalities),
           y = sum_fatalities,
           fill=EVTYPE,
           alpha=0.5
          )
) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = fatalities_chart$sum_fatalities),
            size = 2,
            hjust = 0.5,
            vjust = 1,
            position = "stack") +
  ggtitle("Events = Fatalities") +
  guides(color = "none") +
  coord_flip() +
  xlab("") +
  ylab("") +
  scale_fill_grey() +
  theme_classic()
```




```r
ggsave("fatalities-1.png")
```

```
## Saving 6.5 x 4.5 in image
```

```r
dev.off()
```

```
## null device
##           1
```

Prepare the injuries data to be visualized

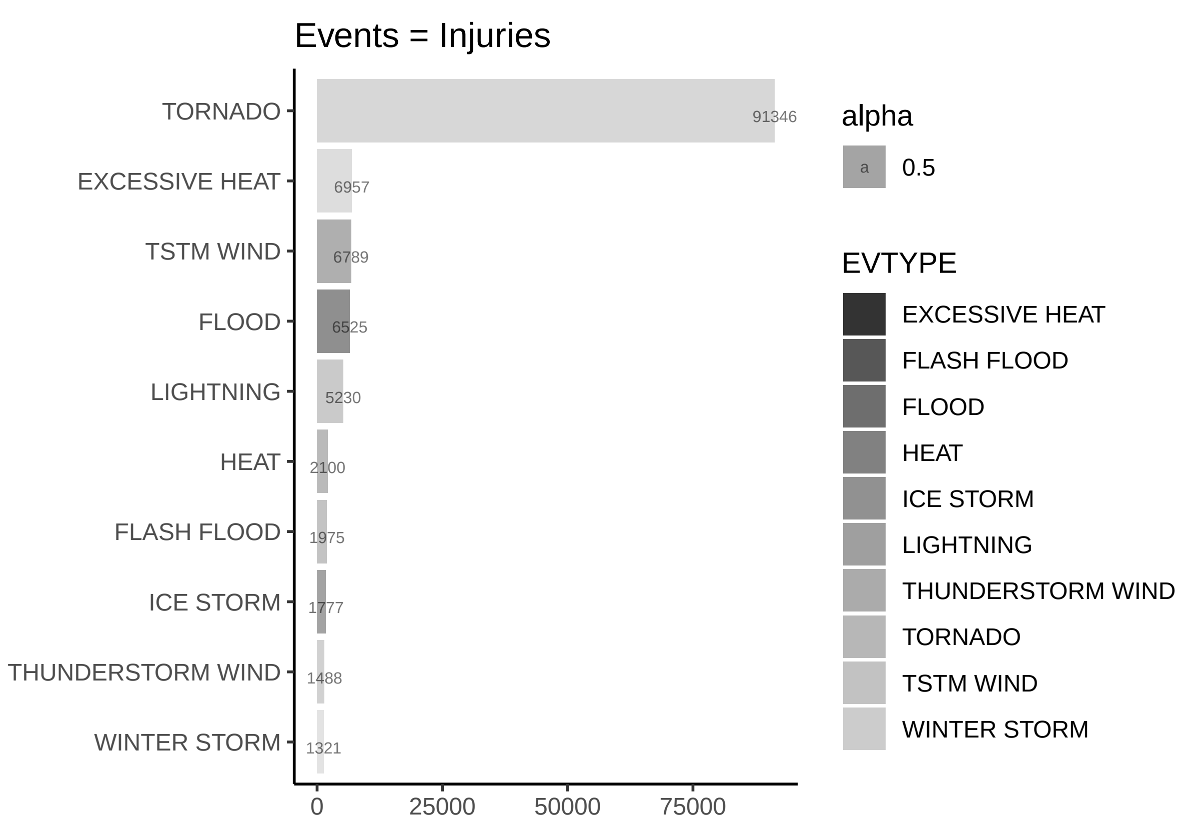
```r
injuries_chart = ddply(top.accidents,
                       .(EVTYPE),
                       summarize,
                       sum_injuries = sum(INJURIES,na.rm=TRUE))
injuries_chart = injuries_chart[order(injuries_chart$sum_injuries, decreasing = TRUE), ]
head(injuries_chart,top_display_events)
```

```
##              EVTYPE sum_injuries
## 8           TORNADO        91346
## 9         TSTM WIND         6957
## 3             FLOOD         6789
## 1    EXCESSIVE HEAT         6525
## 6         LIGHTNING         5230
## 4              HEAT         2100
## 5         ICE STORM         1975
## 2       FLASH FLOOD         1777
## 7 THUNDERSTORM WIND         1488
```

```
## 10       WINTER STORM          1321
```

Events that resulted in injuries

```r
ggplot(injuries_chart[1:top_display_events, ],
       aes(EVTYPE,
           x = reorder(top.accidents$EVTYPE, sum_injuries),
           y = sum_injuries,
           fill=EVTYPE,
           alpha=0.5,
           label = sum_injuries
       )
) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = injuries_chart$sum_injuries),
            size = 2,
            hjust = 0.5,
            vjust = 1,
            position = "stack") +
  ggtitle("Events = Injuries") +
  guides(color = "none") +
  coord_flip()+
  xlab("") +
  ylab("") +
  scale_fill_grey() +
  theme_classic()
```



```r
ggsave("injuries-1.png")
```

```
## Saving 6.5 x 4.5 in image
```

```r
dev.off()
```

```
## null device
##           1
```

Q2) Across the United States, which types of events have the greatest economic consequences? Find out property damage and crop damages

Prepare the damages data to be visualized. This data includes property damages and crop damages

```r
top.damages <- aggregate(cbind(PROPDMG,CROPDMG) ~ EVTYPE,
                          data = property_damages,
                          sum,
                          na.rm=TRUE)

top.damages <- arrange(top.damages,
                       desc(PROPDMG + CROPDMG))

top.damages <- top.damages[1:top_display_events,]
top.damages
```

```
##                EVTYPE   PROPDMG   CROPDMG
## 1             TORNADO 3212258.2 100018.52
## 2         FLASH FLOOD 1420124.6 179200.46
## 3           TSTM WIND 1335965.6 109202.60
## 4                HAIL  688693.4 579596.28
## 5               FLOOD  899938.5 168037.88
## 6   THUNDERSTORM WIND  876844.2  66791.45
## 7           LIGHTNING  603351.8   3580.61
## 8  THUNDERSTORM WINDS  446293.2  18684.93
## 9           HIGH WIND  324731.6  17283.21
## 10       WINTER STORM  132720.6   1978.99
```

Find out how to apply the multiplier for PROPDMGEXP and CROPDMGEXP

```r
table(property_damages$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(property_damages$CROPDMGEXP)
```

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

Property multipliers

```r
property_damages$PROPDMGCALC [property_damages$PROPDMG==0] <- 0
property_damages$CROPDMGCALC [property_damages$CROPDMG==0] <- 0
```

```
property_damages$PROPDMGCALC [property_damages$PROPDMGEXP=="H" |
                                property_damages$PROPDMGEXP=="h"] <- property_damages$PROPDMG[property_
                                                                                           propert
property_damages$PROPDMGCALC [property_damages$PROPDMGEXP=="K" |
                                property_damages$PROPDMGEXP=="k"] <- property_damages$PROPDMG[property_
                                                                                           propert

property_damages$PROPDMGCALC [property_damages$PROPDMGEXP=="M" |
                                property_damages$PROPDMGEXP=="m"] <- property_damages$PROPDMG[property_
                                                                                           propert

property_damages$PROPDMGCALC [property_damages$PROPDMGEXP=="B" |
                                property_damages$PROPDMGEXP=="b"] <- property_damages$PROPDMG[property_
                                                                                           propert
```

CROP multipliers

```
property_damages$CROPDMGCALC [property_damages$CROPDMGEXP=="H" |
                                property_damages$CROPDMGEXP=="h"] <- property_damages$CROPDMG[property_
                                                                                           propert
property_damages$CROPDMGCALC [property_damages$CROPDMGEXP=="K" |
                                property_damages$CROPDMGEXP=="k"] <- property_damages$CROPDMG[property_
                                                                                           propert

property_damages$CROPDMGCALC [property_damages$CROPDMGEXP=="M" |
                                property_damages$CROPDMGEXP=="m"] <- property_damages$CROPDMG[property_
                                                                                           propert
property_damages$CROPDMGCALC [property_damages$CROPDMGEXP=="B"|
                                property_damages$CROPDMGEXP=="b"] <- property_damages$CROPDMG[property_
                                                                                           propert
```

Find total damage

```
total_damages <- aggregate(cbind(PROPDMGCALC,CROPDMGCALC) ~ EVTYPE,
                         data = property_damages,
                         sum,
                         na.rm=TRUE)

total_damages <- arrange(total_damages,
                       desc(PROPDMGCALC + CROPDMGCALC))

total_damages <- total_damages[1:top_display_events,]
total_damages
```

```
##              EVTYPE  PROPDMGCALC CROPDMGCALC
## 1             FLOOD 144657709800  5661968450
## 2  HURRICANE/TYPHOON  69305840000  2607872800
## 3           TORNADO  56936990480   364950110
## 4       STORM SURGE  43323536000        5000
## 5              HAIL  15732262220  3000949450
## 6       FLASH FLOOD  16140811510  1420717100
## 7           DROUGHT   1046106000 13972566000
## 8         HURRICANE  11868319010  2741910000
## 9       RIVER FLOOD   5118945500  5029459000
## 10        ICE STORM   3944927810  5022110000
```

```r
total_damages_sum = total_damages$PROPDMGCALC + total_damages$CROPDMGCALC
total_damages_sum
```

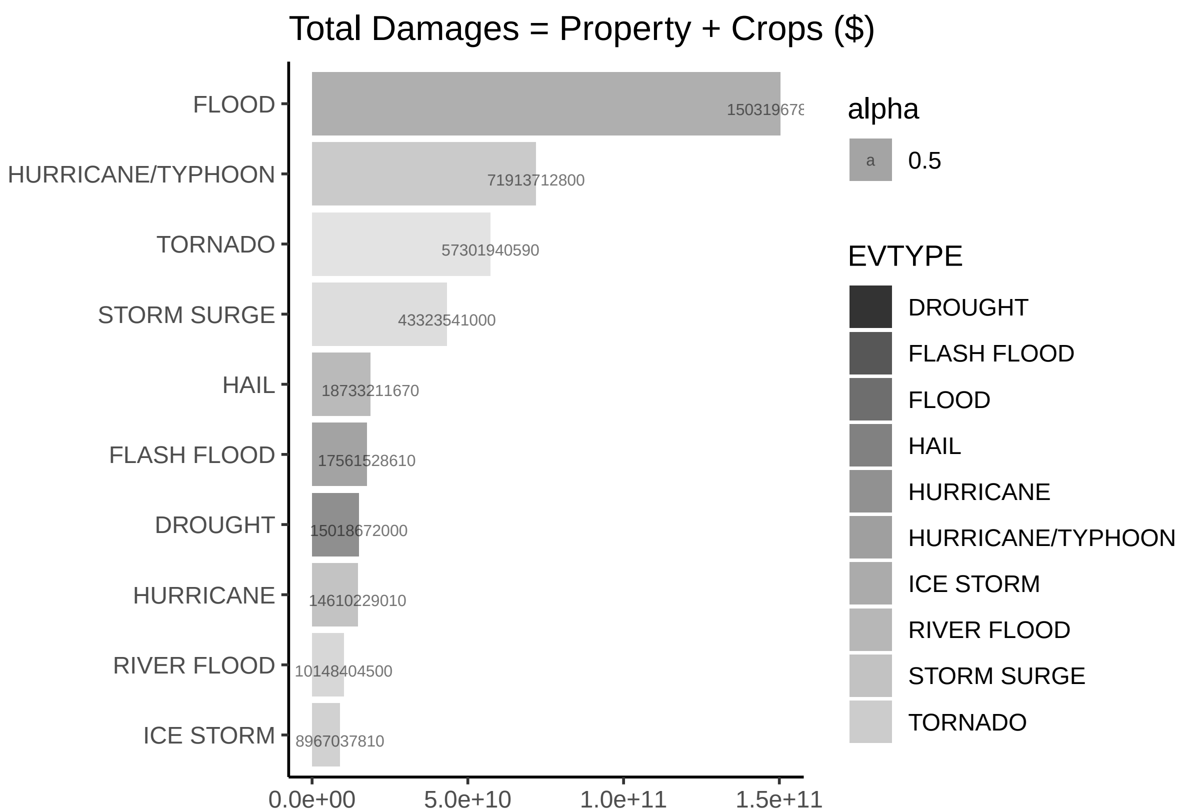
```
##  [1] 150319678250  71913712800  57301940590  43323541000  18733211670
##  [6]  17561528610  15018672000  14610229010  10148404500   8967037810
```

```r
total_damages_table = as.data.frame(cbind(total_damages, total_damages_sum))
total_damages_table
```

```
##               EVTYPE  PROPDMGCALC CROPDMGCALC total_damages_sum
## 1              FLOOD 144657709800  5661968450      150319678250
## 2  HURRICANE/TYPHOON  69305840000  2607872800       71913712800
## 3            TORNADO  56936990480   364950110       57301940590
## 4        STORM SURGE  43323536000        5000       43323541000
## 5               HAIL  15732262220  3000949450       18733211670
## 6        FLASH FLOOD  16140811510  1420717100       17561528610
## 7            DROUGHT   1046106000 13972566000       15018672000
## 8          HURRICANE  11868319010  2741910000       14610229010
## 9        RIVER FLOOD   5118945500  5029459000       10148404500
## 10         ICE STORM   3944927810  5022110000        8967037810
```

Total $ amount of damages to properties and crops resulting from weather-related events

```r
ggplot(total_damages[1:top_display_events, ],
       aes(EVTYPE,
           x = reorder(total_damages$EVTYPE, total_damages_sum),
           y = total_damages_sum,
           fill=EVTYPE,
           alpha=0.5,
           label = total_damages_sum
       )
) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = total_damages_sum),
            size = 2,
            hjust = 0.5,
            vjust = 1,
            position = "stack") +
  ggtitle("Total Damages = Property + Crops ($)") +
  guides(color = "none") +
  coord_flip()+
  xlab("") +
  ylab("") +
  scale_fill_grey() +
  theme_classic()
```




```r
ggsave("damages_costs-1.png")
```

```
## Saving 6.5 x 4.5 in image
```

```r
dev.off()
```

```
## null device
##           1
```

Conclusion and analysis of results

After exploring the data extensively we were able to identify FLOODS as the most impactful from a an economy point of view. TORNADOS are the most devastating weather-related event, capable of causing highest number of injuries and fatalities.