Opportunities and challenges for the use of NOAA Storm Events database for societal impacts research

Therese Kondash   Andrea Schumacher   Brooke Anderson

Abstract

Severe weather events currently cause tens of thousands of deaths per year in the United States and cost trillions of dollars in damaged property and economic loss. These weather events and their impacts are tracked and recorded by several organizations such as the National Weather Service (NWS), the National Oceanic and Atmospheric Administration (NOAA), and the Federal Emergency Management Agency (FEMA). This disaster data is used by scientists in several fields of study (economics, epidemiology, atmospheric science, agriculture, etc.) to examine how weather events impact human life and how we can reduce or avoid these impacts. However, storing disaster data faces unique challenges because of the quickly evolving nature of recording processes and weather technology. These challenges pose concerns regarding biases and errors in disaster data. This paper investigates the presence of mechanisms that could lead to biases in the NOAA Storm Events database. This database contains information on storms, significant weather phenomena, rare or unusual weather, and other significant meteorological events across the United States. We use the noaastormevents package in R to examine evidence of bias within the Storm Events weather data and propose ways to avoid errors while using this data in future studies. This examination reveals five major categories of bias present in the Storm Events data set that are common in weather data. These five categories include hazard bias, temporal bias, threshold bias, accounting bias, and geographic bias. We suggest that these biases result from various factors including structural changes over time, reporting errors, inherent bias, and others. Outlining these factors and understanding their impacts will help scientists use large amounts of data appropriately and portray new findings accurately.

```r
library(noaastormevents)
library(tidyverse)
library(viridis)
library(lubridate)
library(dendextend)
library(kableExtra)
events_2019 <- create_storm_data(date_range = c("2019-01-01", "2019-12-31"))
```

Introduction

Severe weather events currently cost the US trillions of dollars, damage property and crops, interrupt commerce, and cause extensive human injury and fatality. Since 1980, the US has experienced 273 weather and climate disasters that each cost over $1 billion (Environmental Information, n.d.). Additively, these events cost the US over $1.790 trillion and caused 14,223 deaths (Environmental Information, n.d.). Scientists expect these outcomes to worsen over time as climate change causes weather events to occur more frequently and with more severity. This is demonstrated by increasing concentrations of greenhouse gases in the atmosphere that have shown to lead to increasing annual and extreme temperatures. These impacts and others will increase over time as climate change intensifies extreme temperatures and other dangerous weather events. Other weather events will also continue to impact human health and cause more damages in the US as a result of climate change.

Several organizations currently collect data in attempts to compile, categorize, and quantify weather events and their impacts. Some of these organizations include the National Oceanic and Atmospheric Administration

(NOAA), the U.S. Geological Survey (USGS), and the Federal Emergency Management Agency (FEMA). Data collection by these organizations improves scientists' ability to predict weather events and outline monetary losses, property damage, etc. These predictions are commonly accomplished with interdisciplinary research involving fields of study such as epidemiology, atmospheric science, agriculture, economics, etc. Continued interdisciplinary research is critical to help avoid such large fallout and loss from future weather events.

As this research continues, it is incumbent that scientists understand the biases and limitations in weather datasets to use the large amount of data appropriately and portray new findings accurately. Biases and limitations arise in these datasets because storing disaster data is a challenging endeavor. Challenges arise as the process of recording disaster data and the technology we use to do it are constantly evolving.[would a sentence here about what this entails be helpful?] Gall et al. recognized this issue and wrote a paper covering several bias types associated with major disaster databases used in the US (Gall, Borden, and Cutter 2009). These biases include hazard bias, temporal bias, threshold bias, accounting bias, geography bias, and systemic bias (Gall, Borden, and Cutter 2009).

One such weather database in which these biases arise is the National Oceanic and Atmospheric Administration's Storm Events database. This database was created in 1950 and is now maintained by branches of NOAA including the National Weather Service (NWS) and the National Center for Environmental Information (NCEI). It originally only contained information regarding tornado events, but has expanded to include data on 55 different storms, significant weather phenomena, rare/unusual weather, and other significant meteorological events across the United States. This data is saved in files categorized by year and published annually in the NOAA Storm Events Publication [link]. This expansive dataset is an important resource used by atmospheric scientists to analyze fatalities and other losses associated with weather events to complete scientific research [cite examples- using for other reasons]. The American Meteorological Society has published several papers that use this research to characterize specific weather events and identify unique vulnerabilities or behaviors that place humans in harm's way (Terti et al. 2017, Ashley and Mote (2005), Ashley and Black (2008), Gensini and Ashley (2010)). These studies help provide justification for improvements to weather warning systems or future risk assessments.

In this paper, we will examine the NOAA Storm Events database and investigate the presence of mechanisms that lead to biases in the data. We outline these biases in five major categories of hazard bias, temporal bias, threshold bias, accounting bias, and geographic bias. The goal of this investigation is to improve the ease and accuracy of research using this data to study harmful impacts of weather events in the US.

Examples of Storm Events being used in literature

The American Meteorological Society has published various papers analyzing fatalities associated with weather events (Terti et al. 2017, Ashley and Mote (2005), Ashley and Black (2008), Gensini and Ashley (2010)). Ashley and Black examined the cause, spatial distributions, and fatalities caused by nonconvective high-wind events in the US from 1980 to 2005. This study found that fatalities associated with nonconvective high-wind events often occur in boats or vehicles. These fatalities are most likely to occur on the West Coast or Northeast as a result of extratropical cyclones. These areas of the US have large forests and bodies of water with high population densities that are opportune to being highly affected by high winds (Ashley and Black 2008). Keeping these vulnerabilities in mind, the study outlines possible improvements to high wind warning systems in the areas. These improvements aim to protect people most likely to be injured or killed by nonconvective high-wind events.

Ashley and Mote conducted a similar study examining derecho events in the US from 1986 to 2003 (Ashley and Mote 2005). Derechoes are often overlooked when examining impactful weather events. However, this study revealed how they actually produce damage comparable to tornadoes and hurricanes. Derechoes caused 153 fatalities and over 2,600 injuries throughout the 18 year span of this study. They were also responsible for as much or more monetary loss than some hurricane and tornado events over the time span. Their study helps to draw attention to this type of weather event in hopes to advance derecho risk assessments.

Terminology and how the database works

Throughout this paper, we will be using specific terminology when describing the structure and function of the NOAA Storm Events database. We use the term hazard or weather event to refer to all of the possible event types that can be entered in the NOAA Storm Events database. A hazard can be defined as something that has the potential to cause damage or a threat that can severely disrupt society's workings. A hazard's potential to cause harm is compounded by human factors like culture, gender, race, socioeconomic status and political structure and by human activity like land development, urbanization, emergency preparedness, and response plans [cutter1996responses]. These factors can influence the actual potential for harm, the perceived potential for harm, and the ability to respond if a disaster does occur.

The terms risk and disaster are two other concepts that tie into the definition of a hazard. Risk is the probability of a hazard occurring and actually causing harm. A disaster is the occurrence of a hazard that creates a large impact on a group of people or peoples. For example, a shark swimming in the ocean is a hazard that has potential to cause damage. However, the risk of this hazard causing a person harm varies depending on the person's location relative to the shark. This hazard would become a disaster if the shark attacked the swimmer.

We use the term disaster to refer to a reported occurrence of a specific hazard. NOAA Storm Events currently records 55 different event types that are listed in the NWS instruction manual (NWS, n.d.). Most of these event types correlate to one type of hazard but some of them correlate to different intensities of the same type of hazard. Some examples of this include the event type "Strong Wind" versus "High Wind" or "Cold/Wind Chill" versus "Extreme Cold/Wind Chill". In these instances, the hazard would be wind or wind chill while the intensities are strong, high, cold, or extreme cold.

For 2019, the database reported 51 total events of 55 different event types. The most frequent type of events reported were Thunderstorm Wind, which had over twice as many reports as the second most frequent, Hail. Hail, in turn, had almost twice as many events reported as the next most common few—Flood, Flash Flood, and Winter Weather. The least common reported events included Volcanic Ashfall, Sleet, Dense Smoke, Sneakerwave, Seiche, and Marine Tropical Depression.

[TK: include plot of event types in 2019]

| Event type	Number of events in 2019
Thunderstorm Wind	18,635
Hail	9,015
Flood	4,951
Flash Flood	4,093
Winter Weather	3,808
High Wind	3,790
Winter Storm	3,354
Heavy Snow	2,886
Marine Thunderstorm Wind	2,502
Tornado	1,732
Strong Wind	1,591
Heavy Rain	1,441
Heat	1,301
Extreme Cold/Wind Chill	1,065
Drought	1,007
Excessive Heat	889
Blizzard	852
Frost/Freeze	654
Dense Fog	652
Cold/Wind Chill	470
High Surf	468
Funnel Cloud	348
Lightning	343
Coastal Flood	240
Ice Storm	200
Wildfire	197
Debris Flow	185
Waterspout	183
Lakeshore Flood	146
Tropical Storm	143
Lake-Effect Snow	78
Marine High Wind	77
Rip Current	74
Avalanche	44
Dust Storm	40
Marine Tropical Storm	37
Marine Hail	32
Tropical Depression	26
Storm Surge/Tide	25
Marine Strong Wind	17
Astronomical Low Tide	16
Hurricane	10
Dust Devil	8
Freezing Fog	6
Marine Hurricane/Typhoon	6
Marine Tropical Depression	4
Seiche	4
Sneakerwave	3
Dense Smoke	2
Sleet	1
Volcanic Ashfall	1

NOAA further characterizes its data by categorizing large storms as episodes which contain several individual events. NOAA assigns the disaster with one episode ID and its associated events with event IDs. Further details of the episode and events are recorded under the event narrative and episode narrative. The NWS instruction describes an episode as a swath of events occurring along a path from Point A to Point B during Time C to Time D (NWS, n.d.). Events belonging to the same episode cannot be more than five calendar days apart to ensure they all resulted from one meteorological episode. This categorization helps to ensure that a single storm is not interpreted as a series of unrelated events. An example of this would be a hurricane given one episode ID, and the rain, wind, floods, etc. associated with that hurricane given event IDs that fall under that episode.

To showcase this idea, for the episodes with the most events in 2019, the following graph shows the number of events reported for the episode. This figure demonstrates how one large weather episode can include several other events types. For example, one episode in this figure corresponds to an atmospheric river and cold front near Mount Saint Helena and indicates several reports of flood, debris flow, strong winds, high wind, and hail. All of these different event types occurred as a result of one large storm labelled episode 133801. This unique episode ID helps related events from getting double counted as separate entities.

[TK: include plot and add in names of episode in paragraph rather than numbers]

`summarise()` ungrouping output (override with `.groups` argument)

`summarise()` regrouping output by 'EPISODE_ID' (override with `.groups` argument)
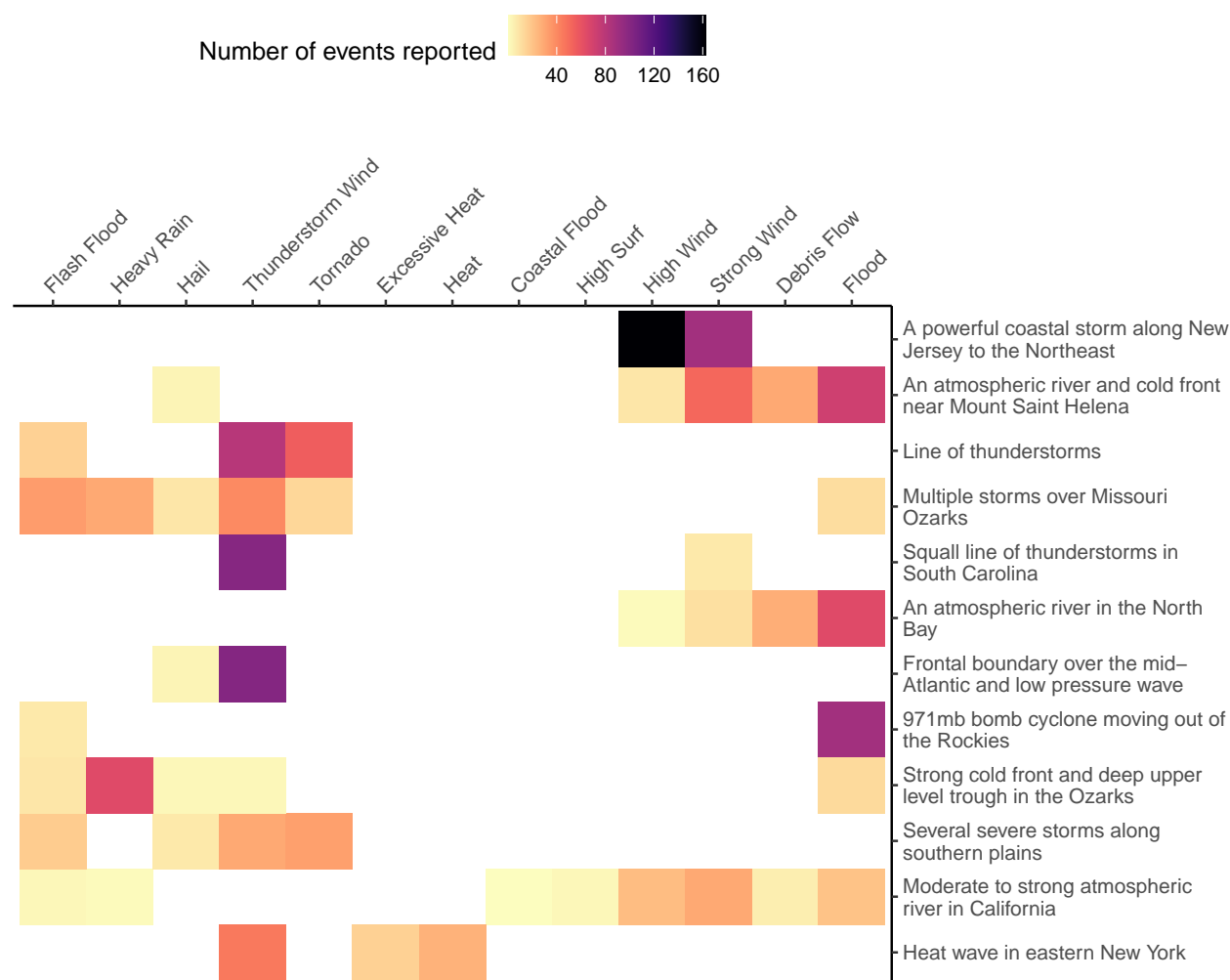
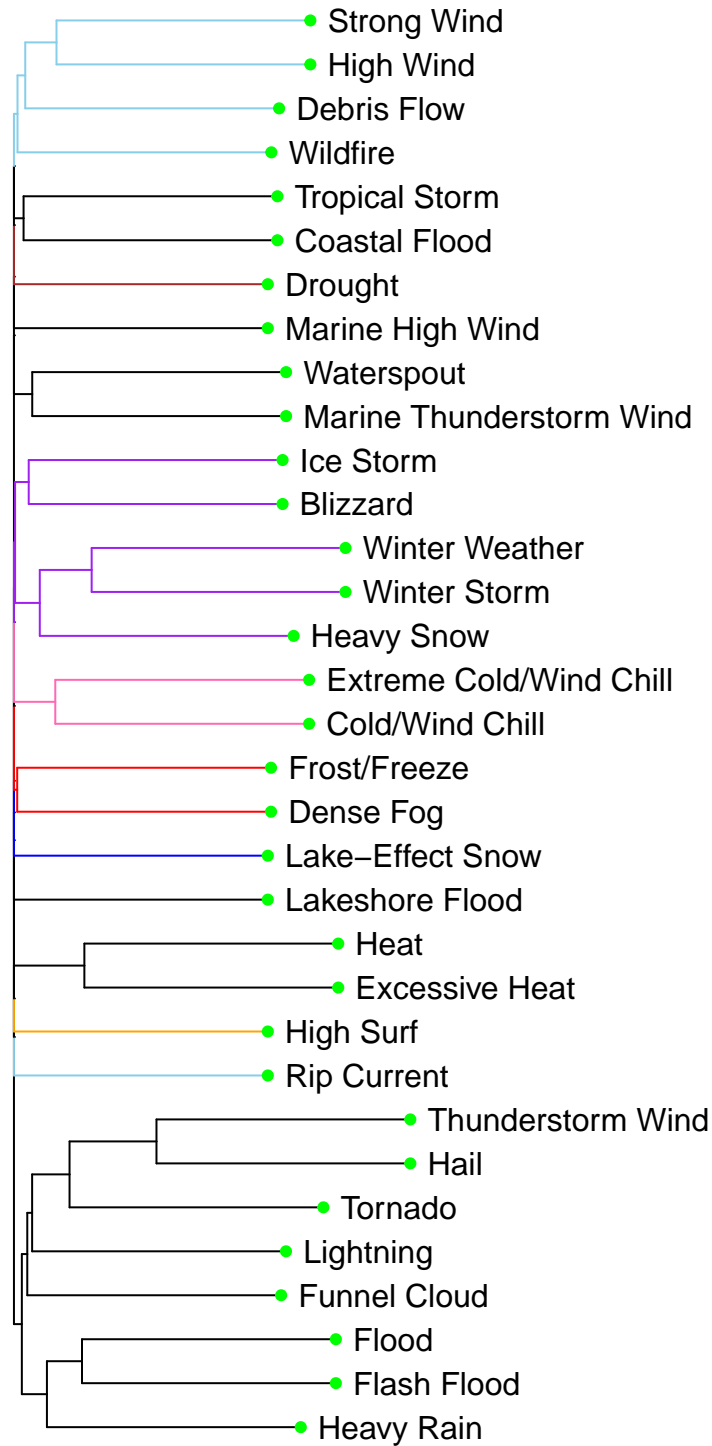In some cases, events falling under one episode are the same type of hazard just at different intensities, as

described above. In these circumstances, the collection of events may be recorded in only one county/zone but will showcase the spatial range and and varying intensity of the event. This is likely why we see the events with different intensities of the same hazard clustered together in one county/zone in the database, like heat/excessive heat and cold/extreme cold. In other cases, a larger synoptic weather system might bring a variety of different hazards leading to several different recorded event types (e.g., thunderstorm wind and hail, which cluster together). In these cases, there may be several counties/zones in the affected area that report more than one event for the episode.

[TK: potentially add in map of events over different counties vs one county]

The following plot shows a cluster analysis to group events that are more likely to (be recorded as occurring) occur together within an episode. We removed event types with less that 50 listings in 2019. The figure shows that there are certain event types which are very commonly reported together such as high wind and strong wind or heat and excessive heat. These types of events are likely to be reported together because they are likely to occur at the same time. It is important to keep these trends in mind as there may be overlap or uncertainty in how to categorize what happened during an episode.

```
## `summarise()` regrouping output by 'EPISODE_ID' (override with `.groups` argument)
```

Event Type Cluster Analysis

- Strong Wind
- High Wind
- Debris Flow
- Wildfire
- Tropical Storm
- Coastal Flood
- Drought
- Marine High Wind
- Waterspout
- Marine Thunderstorm Wind
- Ice Storm
- Blizzard
- Winter Weather
- Winter Storm
- Heavy Snow
- Extreme Cold/Wind Chill
- Cold/Wind Chill
- Frost/Freeze
- Dense Fog
- Lake–Effect Snow
- Lakeshore Flood
- Heat
- Excessive Heat
- High Surf
- Rip Current
- Thunderstorm Wind
- Hail
- Tornado
- Lightning
- Funnel Cloud
- Flood
- Flash Flood
- Heavy Rain

Hazard Bias

In this section, we will outline how hazard bias may be created when using the NOAA Storm Events database for societal impacts research. Hazard bias can present itself as the probability of an event being recorded varying according to event type or the quality/quantity of information recorded varying according to event type. This can manifest in certain event types being under or over-reported in the dataset or having much more reliable and helpful descriptions of certain event types. This occurs because of variety of factors relating to how the database functions and how weather monitoring occurs.

Hazard bias by standards for when different events are recorded

[TK: this is just one mechanism for why this is going on]

One mechanism that creates the opportunity for hazard bias arises out of the standards that the NOAA Storm Events database uses in deciding whether or not to include a disaster report in its dataset. The NWS Documetation record all standards used in making this decision here [link]. It appears that these standards vary based on event type. In general, the documentation states that the Storm Events database seeks to record rare, unusual, recording breaking, or highly impactful weather events. This implies that the database generally does not seek to record every event that is occuring, but rather focuses on events that are unusual (snow in Florida, record-breaking events) and/or things that impact humans (loss of life, property damage). However, the documentation goes on to state that "significant events" like Tornado events should always be recorded, regardless of if they had any impact like injury, fatality, or damage (NWS, n.d.). Another example occurs in that the NOAA documentation states that avalanches should only be recorded if human fatality or injury occurs. However, there are many types of avalanches (slab, loose snow, icefall, cornice fall, etc.) that can occur naturally due to a variety of environmental conditions which may not cause any human fatality or injury. These variations have the potential to create hazard bias in the database by impacting the probability of an event being recorded given that it actually happened. This probability will be high for tornado events, because the database seeks to record all tornado events regardless of their impact. This probability will be low for avalanches because the database only seeks to record impactful occurences.

To avoid this bias in a study using data from Storm Events, it is important to match the intent of a research study with the intent of the databbase for the event type at hand. Your results could vary based on if you want to examine all occurences of a certain hazard type in a meteorological sense versus if you just want to examine their human impacts. If you are trying to catch every physical instace of a hazard type, this may not even be the goal of the database itself. We have organized the hazard types into general categories based on the NWS's standard for recording.

[TK: include table that lists all events and if they recorded based on impact, or on intensity at the end of this section]

[TK: other examples: Drought hazards exemplify this issue; they are notoriously under-reported because there is a lack of physical damage and it is hard to quantify spatial and monetary losses (Gall, Borden, and Cutter 2009).

```
event_reporting_areas <- tribble(
  ~ event_type, ~ reporting_area,
  "Astronomical Low Tide", "Forecast Zone",
  "Avalanche", "Forecast Zone",
  "Blizzard", "Forecast Zone",
  "Coastal Flood", "Forecast Zone",
  "Cold/Wind Chill", "Forecast Zone",
  "Debris Flow", "County",
  "Dense Fog", "Forecast Zone",
  "Dense Smoke", "Forecast Zone",
  "Drought", "Forecast Zone",
  "Dust Devil", "County",
  "Dust Storm", "Forecast Zone",
```

```
    "Excessive Heat", "Forecast Zone",
    "Extreme Cold/Wind Chill", "Forecast Zone",
    "Flash Flood", "County",
    "Flood", "County",
    "Freezing Fog", "Forecast Zone",
    "Frost/Freeze", "Forecast Zone",
    "Funnel Cloud", "County",
    "Hail", "County",
    "Heat", "Forecast Zone",
    "Heavy Rain", "County",
    "Heavy Snow", "Forecast Zone",
    "High Surf", "Forecast Zone",
    "High Wind", "Forecast Zone",
    "Hurricane/Typhoon", "Forecast Zone",
    "Ice Storm", "Forecast Zone",
    "Lakeshore Flood", "Forecast Zone",
    "Lake-Effect Snow", "Forecast Zone",
    "Lightning", "County",
    "Marine Dense Fog", "Marine Zone",
    "Marine Hail", "Marine Zone",
    "Marine Heavy Freezing Spray", "Marine Zone",
    "Marine High Wind", "Marine Zone",
    "Marine Hurricane/Typhoon", "Marine Zone",
    "Marine Lightning", "Marine Zone",
    "Marine Strong Wind", "Marine Zone",
    "Marine Thunderstorm Wind", "Marine Zone",
    "Marine Tropical Depression", "Marine Zone",
    "Marine Tropical Storm", "Marine Zone",
    "Rip Current", "Forecast Zone",
    "Seiche", "Forecast Zone",
    "Sleet", "Forecast Zone",
    "Sneaker Wave", "Forecast Zone",
    "Storm Surge/Tide", "Forecast Zone",
    "Strong Wind", "Forecast Zone",
    "Thunderstorm Wind", "County",
    "Tornado", "County",
    "Tropical Depression", "Forecast Zone",
    "Tropical Storm", "Forecast Zone",
    "Tsunami", "Forecast Zone",
    "Volcanic Ash", "Forecast Zone",
    "Waterspout", "Marine Zone",
    "Wildfire", "Forecast Zone",
    "Winter Storm", "Forecast Zone",
    "Winter Weather", "Forecast Zone"
)


event_type_2019 <-events_2019 %>%
  mutate(EVENT_TYPE = case_when(
    EVENT_TYPE == "Sneakerwave" ~ "Sneaker Wave",
    EVENT_TYPE == "Hurricane" ~ "Hurricane/Typhoon",
    EVENT_TYPE == "Volcanic Ashfall" ~ "Volcanic Ash",
```

```r
    TRUE ~ EVENT_TYPE
  )) %>%
  rename(event_type = EVENT_TYPE) %>%
  group_by(event_type) %>%
  summarize(N = n()) %>%
  arrange(desc(N))



reporting_categories <- tribble(
  ~ event_type, ~ reporting_standard,
  "Astronomical low tide", "impact, noteworthiness, public interest",
  "Avalanche" , "impact, noteworthiness, public interest",
  "Coastal Flood ", "impact, noteworthiness, public interest" ,
  "  Dense Fog ", "impact, noteworthiness, public interest" ,
  "Dense Smoke ", "impact, noteworthiness, public interest" ,
  "Dust Devil ", "impact, noteworthiness, public interest" ,
  "Dust Storm ","impact, noteworthiness, public interest",
  "Flash Flood", "impact, noteworthiness, public interest",
  "Flood", "impact, noteworthiness, public interest" ,
  "Freezing Fog", "impact, noteworthiness, public interest" ,
  "Funnel Cloud" , "impact, noteworthiness, public interest" ,
  "Freezing Fog" , "impact, noteworthiness, public interest" ,
  "Funnel Cloud" , "impact, noteworthiness, public interest" ,
  "Heavy Rain" , "impact, noteworthiness, public interest" ,
  "High Surf" , "impact, noteworthiness, public interest" ,
  "Lakeshore Flood" , "impact, noteworthiness, public interest" ,
  "Lightning" , "impact, noteworthiness, public interest" ,
  "Marine Dense Fog" , "impact, noteworthiness, public interest" ,
  "Marine Hail" , "impact, noteworthiness, public interest" ,
  "Marine Heavy Freezing Spray" , "impact, noteworthiness, public interest" ,
  "Marine High Wind" , "impact, noteworthiness, public interest" ,
  "Marine Hurricane/Typhoon","impact, noteworthiness, public interest" ,
  "Marine Lightning" , "impact, noteworthiness, public interest" ,
  "Marine Strong Wind" , "impact, noteworthiness, public interest" ,
  "Marine Thunderstorm Wind" , "impact, noteworthiness, public interest" ,
  "Marine Tropical Depression" , "impact, noteworthiness, public interest" ,
  "Marine Tropical Storm" , "impact, noteworthiness, public interest" ,
  "Rip Currents" , "impact, noteworthiness, public interest" ,
  "Seiche" , "impact, noteworthiness, public interest" ,
  "Sneaker Wave" , "impact, noteworthiness, public interest" ,
  "Storm Surge/Time" , "impact, noteworthiness, public interest",
  "Strong Wind ", "impact, noteworthiness, public interest" ,
  "Thunderstorm Wind" , "impact, noteworthiness, public interest" ,
  "Tsunami" , "impact, noteworthiness, public interest" ,
  "Volcanic Ash" , "impact, noteworthiness, public interest" ,
  "Wildfire" , "impact, noteworthiness, public interest" ,
  "Winter Weather" , "impact, noteworthiness, public interest",
  "Blizzard", "intensity",
  "Cold/Wind Chill ", "intensity",
  "Debris Flow", "intensity",
  "Excessive Heat", "intensity",
  "Extreme Cold/Wind Chill", "intensity",
```

```
  "Frost", "intensity",
  "Hail", "intensity",
  "Heat", "intensity",
  "Heavy Snow", "intensity",
  "High Wind ", "intensity",
  "Hurricnae/Typhoon", "intensity",
  "Ice Storm", "intensity",
  "Lake Effect Snow", "intensity",
  "Sleet", "intensity",
  "Tornado", "intensity",
  "Tropical Depression", "intensity",
  "Tropical Storm", "intensity",
  "Waterspout", "intensity",
  "Winter Storm", "intensity"
)



event_type_2019 %>%
  full_join(event_reporting_areas) %>%
  full_join(reporting_categories) %>%
  arrange(desc(N)) %>%
  mutate(N = case_when(
    is.na(N) ~ as.integer(0),
    TRUE ~ N
  )) %>%
  mutate(N = prettyNum(N, big.mark = ",")) %>%
  knitr::kable(col.names = c("Event type", "Number of events in 2019", "Reporting Area",
                             "Reporting Standard"))
```

Event type	Number of events in 2019	Reporting Area	Reporting Standard
Thunderstorm Wind	18,635	County	impact, noteworthiness, public interest
Hail	9,015	County	intensity
Flood	4,951	County	impact, noteworthiness, public interest
Flash Flood	4,093	County	impact, noteworthiness, public interest
Winter Weather	3,808	Forecast Zone	impact, noteworthiness, public interest
High Wind	3,790	Forecast Zone	NA
Winter Storm	3,354	Forecast Zone	intensity
Heavy Snow	2,886	Forecast Zone	intensity
Marine Thunderstorm Wind	2,502	Marine Zone	impact, noteworthiness, public interest
Tornado	1,732	County	intensity
Strong Wind	1,591	Forecast Zone	NA
Heavy Rain	1,441	County	impact, noteworthiness, public interest
Heat	1,301	Forecast Zone	intensity
Extreme Cold/Wind Chill	1,065	Forecast Zone	intensity
Drought	1,007	Forecast Zone	NA
Excessive Heat	889	Forecast Zone	intensity
Blizzard	852	Forecast Zone	intensity
Frost/Freeze	654	Forecast Zone	NA
Dense Fog	652	Forecast Zone	NA
Cold/Wind Chill	470	Forecast Zone	NA
High Surf	468	Forecast Zone	impact, noteworthiness, public interest
Funnel Cloud	348	County	impact, noteworthiness, public interest
Funnel Cloud	348	County	impact, noteworthiness, public interest
Lightning	343	County	impact, noteworthiness, public interest
Coastal Flood	240	Forecast Zone	NA
Ice Storm	200	Forecast Zone	intensity
Wildfire	197	Forecast Zone	impact, noteworthiness, public interest
Debris Flow	185	County	intensity
Waterspout	183	Marine Zone	intensity
Lakeshore Flood	146	Forecast Zone	impact, noteworthiness, public interest
Tropical Storm	143	Forecast Zone	intensity
Lake-Effect Snow	78	Forecast Zone	NA
Marine High Wind	77	Marine Zone	impact, noteworthiness, public interest
Rip Current	74	Forecast Zone	NA
Avalanche	44	Forecast Zone	impact, noteworthiness, public interest
Dust Storm	40	Forecast Zone	NA
Marine Tropical Storm	37	Marine Zone	impact, noteworthiness, public interest
Marine Hail	32	Marine Zone	impact, noteworthiness, public interest
Tropical Depression	26	Forecast Zone	intensity
Storm Surge/Tide	25	Forecast Zone	NA
Marine Strong Wind	17	Marine Zone	impact, noteworthiness, public interest
Astronomical Low Tide	16	Forecast Zone	NA
Hurricane/Typhoon	10	Forecast Zone	NA
Dust Devil	8	County	NA
Freezing Fog	6	Forecast Zone	impact, noteworthiness, public interest
Freezing Fog	6	Forecast Zone	impact, noteworthiness, public interest
Marine Hurricane/Typhoon	6	Marine Zone	impact, noteworthiness, public interest
Marine Tropical Depression	4	Marine Zone	impact, noteworthiness, public interest
Seiche	4	Forecast Zone	impact, noteworthiness, public interest
Sneaker Wave	3	Forecast Zone	impact, noteworthiness, public interest
Dense Smoke	2	Forecast Zone	NA
Sleet	1	Forecast Zone	intensity
Volcanic Ash	1	Forecast Zone	impact, noteworthiness, public interest
Marine Dense Fog	0	Marine Zone	impact, noteworthiness, public interest
Marine Heavy Freezing Spray	0	Marine Zone	impact, noteworthiness, public interest
Marine Lightning	0	Marine Zone	impact, noteworthiness, public interest
Tsunami	0	Forecast Zone	impact, noteworthiness, public interest

By Varying Amounts of Information Provided

There are also instances where different amounts of information are available for different types of hazards listed in the database. The Storm data dataset has around 51 columns of information per entry that include information like the event ID, time, state, county, damage, textual descriptions, etc. Not every one of these columns is filled with each entry in the database.

Narratives

This variation in the amount of information shows up in the textual descriptions provided per entry. For some types of reporting, the *episode* gets the narrative, but not separate narratives for each event, while for some events (like tornadoes), each *event* might get a narrative. The NWS Documentation covers this in section 2.9 Textual Description of Storm [link?]. The basic requirement is that each disaster [TK: questionable terminology here] entry within the Storm Data dataset recieves a short episode narrative. In addition to this, an event narrative should be entered if the event caused some sort of injury or damage, or something else significant. The documentation also states that an event narrative is requried for all Tornado events, all Thunderstorm Wind events, and all Lightning events, whether over land or marine zones. For these hazard types, the event may not have caused significant damage, but an event narrative will be still be present. Other than this, if there is nothing noteworthy about the event itself, no event narrative is required. The documentation provides Hail events as an example because as a single phenomenon, Hail events do not need narratives unless they are part of a more complex weather episode or cause fatality/injury or property/crop damage (NWS, n.d.). This is important to keep in mind because the presence of an event narrative does not necessarily indicate significance if it is for the hazard types where an event narrative is required.

Spatial Data

Similar variation occurs in the spatial data provided per hazard type. The NWS states that a hydrometereorological event should be referenced to the nearest hudredth of a mile, to the geographical center of a particular village, city, airport, inland lake, or location provided that the reference point is in the Storm Data software location database (NWS, n.d.). The Storm Data preparer is also "strongly encouraged" to enter latitude/longitude to describe the location.

Additionally, all event types are categorized into a county/parish (`CZTYPE` of "C"), a forecast zone (`CZTYPE` of "Z"), or a marine zone (`CZTYPE` of "M") (see table below)** [TK: referencing table with all event types, could also create a table exmaining an event that crossed several counties or forecast zones]. Forecast zones are created as subsets of counties for more specific location data. Specific types of events are typically either always reported for a county or always reported for a forecast zone. Events typically reported by county include floods, tornado-like events, and a few other events often related to thunderstorms. Events typically reported by forecast zone include severe winter weather, extreme heat, events related to the water or coast, and a few others ("High Wind", "Dense Fog", "Strong Wind", "Wildfire", "Dust Storm", "Dense Smoke").

This demonstrates the difficulty of narrowing down where exactly certain events occurred. Some events might be very localized (for example, a flash flood, or even storm surge that's only near the coast of the county), but the county-level reporting of the event means that the whole county gets classified as experiencing or not experiencing the event. For these localized events, the researcher may need to pay closer attention to the reference point or latitude/logitude data.

For all tornado events, the database also includes path length (in miles and tenths of miles) and width (in yards). This information is given in the header strip of the Storm Event publication per episode/event and it indicates the length of one segment in one county. A user of the Storm Data database can determine the entrie length of a multi-segmented tornado by adding the lengths of each segment or using latitude/longitude information. The documentation includes more stipulations on this topic in that the given path length will exclude sections without "surface damage/disturbance, unless other evidence of the touchdown (e.g., a trained spotter report, videotape of the tornado over a plowed field) is available" (NWS, n.d.). This excluded section cannot exceed 2 continuous miles or 4 consecutive minutes of travel time or else it will be categorized as separate tornado events within the episode.

Once again it appears that tornado events get a higher level of detailed information than other event types.

This variation should be kept in mind when using this database to examine multiple hazard types.

[TK: documentation information from section 2.4 and 47.7. I can't find which events get latitude/lognitude data, I think it just can be entered if the preparer has that info?]

Estimates of Damage (Monetary and Injury/Fatality)

Another mechanism that creates the opportunity for hazard bias has to do with how monetary damage estimates are created for the database. The NWS documentation has this description for how damages should be recorded:

> 2.7 Damage. Property damage estimates should be entered as actual dollar amounts, if a reasonably accurate estimate from an insurance company or other qualified individual is available. If this estimate is not available, then the preparer has two choices: either check the "no information available" box, or make an estimate. The lone exception is for flood events. The Storm Data preparer enters monetary damage amounts for flood events, even if it is an estimate. The U.S. Army Corps of Engineers requires the NWS to provide monetary damage amounts (property and/or crop) resulting from any flood event."

In a general sense, this could lead to bias if a researcher was trying to examine overall monetary damage across all event types because there is large variation in how this number is obtained. More specifically, estimates from insurance companies are probably more likely to be available for events that are highly impactful in larger cities. For example, hurricanes or tornadoes that pass through large cities would attract the attention of insurance companies that insure high concentrations of buildings and properties there. This may produce clearer records of damages than a smaller weather event that occurs in rural area. Additionally, the documentation points out that estimates for flood damage are required. This places higher importance on the amount of information recorded for one specific event type. Once again, this should be kept in mind if using this database for analysis of weather event related damages.

[TK: ask Andrea about the above claim? ask James Done?]

[TK: I don't know if the following information belongs in the hazard bias section but I think it is important to point out in terms of varying levels of detail in event vs episode nanrratives]

For injury and fatality information, the database includes information on direct, indirect, and delayed injuries/fatalities. [TK: I could explain how the database distinguishes between these here, or just reference and/or link the documentation section] When a direct injury or fatality occurs from a weather event, it gets entered into the Storm Data's "fatality" and "injury" fields. When an indirect fatality or injury occurs, this information is included in the event narrative. Delayed fatalities are included in the original weather event entry with the actual date of fatality in the fatality enntry field. The documentation further states that Aan explanation can be given in the *episode* narrative for *zone-based* events, or in the *event* narrative for *county-based* or *marine zone-based* events.

Hazard bias by who reported event

Hazard bias may also arise from the variation in source type per hazard type. There is a wide range of people and detection devices that collect data for this dataset which may impact the probabilty of an event being recorded and the quality of information that is provided for a reported event. We have categorized the types of event sources for this database into four main categories: automated observation systems, officials or emergency response/agency, trained volunteers, and company or general public (see table below). This brings up several potential challenges.

The likelihood of catching an event varies based on source because of different levels of detection technology. We may have a better process for ID-ing cases of one hazard type compared to another in that some are recorded with detection technology and some are reported by eye-witnesses. Events that are typically recorded by automated observation systems may be more likely to be reported than ones that are typically reported by the public. For example, an automated system would probably "see" any event that happens where the monitor is located, while an event that's usually reported by the public would miss anything that happens

when a person isn't around or can't see it (e.g., at night, in a sparsely populated area). This can be seen in that sneakerwaves (also seiche, sleet) are rarely reported in the dataset and are only reported by sources like broadcast media, fire department/rescue, and newspaper. [TK: do comparison of # of entries for an event type vs its source)

Even within one type of source, that there may be controversy in the agreement across reports for a hazard entry. There could be variation amongst the monitoring equipment for automated observation systems due to random error or inconsistent calibration. There could also be variation amongst people's opinions across volunteer reports especially if event criterias are unclear. Along these same lines, the quality of information provided in these accounts can vary based on if there is an established scale or not. For example, the Enhanced Fujita scale is the accepted way to rank and characterize tornado events. This scale provides a standard for explaing the intensity and impact of tornado events, whereas less common events do not have a widely accepted scale for description. These factors will impact how cohesive the information is for individual hazard entries and for collective hazard types.

`summarise()` ungrouping output (override with `.groups` argument)

Source of event report	Number of events reported in 2019
Trained Spotter	9,742
Public	8,378
Emergency Manager	6,061
Mesonet	5,717
Law Enforcement	4,061
ASOS	3,169
COOP Observer	2,758
911 Call Center	2,615
Broadcast Media	2,358
Social Media	2,065
AWOS	2,029
Department of Highways	1,917
River/Stream Gage	1,761
NWS Storm Survey	1,749
State Official	1,748
Official NWS Observations	1,424
CoCoRaHS	1,330
Fire Department/Rescue	1,187
Amateur Radio	1,176
Drought Monitor	958
NWS Employee	901
Buoy	590
SNOTEL	484
Other Federal Agency	479
Storm Chaser	447
County Official	443
RAWS	402
C-MAN Station	396
Utility Company	393
Newspaper	274
Park/Forest Service	244
Unknown	194
Local Official	71
AWSS	29
Mariner	22
Lifeguard	21
Post Office	20
Tribal Official	14
Coast Guard	9
Airplane Pilot	7
Insurance Company	5
WLON	3
SHAVE Project	1

[TK: include figure with # of events reported by event source in 2019, figure showing which sources typically report which event types, maybe table that lists sources into these four categories][BA: One may be that the (this is a similar idea to how changes in technology over time might lead to temporal bias). I wonder if some of our rare events (Sneakerwave, Seiche, Sleet) might be examples of spots where we don't have great technology yet to detect them (or not a large enough network of monitors even if we do have the technology)? Another point on this idea of different amount/quality of information on different hazards might be that some hazards are recorded based on clear ways to measure intensity (for example, using a well-defined scale, as is done for tornadoes with the Fujita scale), while others don't. The NOAA guidebook might have some

details on this in terms of telling people to use specific scales to report some types of events.]

[From Andrea: One note in terms of time-variability of tornado reporting - it's thought that that detection of weaker tornadoes (EF-1) has improved significantly in the last few decades (due to increases in population and better observation networks). For that reason, researchers sometimes remove EF-1's when analyzing tornado trends to try to reduce this type of bias.]

implications and discussion

[TK: I have woven the answer to some of these questions in the above parapraphs. Do you have ideas for what else we should expand on here?]

For example, would it only be an issue when comparing frequencies of different types of events? Could it lead to confounding in a single-hazard study that's looking at disaster impacts? It would certainly result in undercounts of certain types of hazards, if you're trying to characterize how frequently a type of hazard tends to happen. Would there be implications from information bias, particularly measurement error? (We could save this paragraph to draft for after we've drafted the rest of this section, and just put in a placeholder for now.)

[TK: These were health implications of weather events that I omitted from the introduction] In specific, scientists expect the number of naturally occurring wildfires to increase (Crimmins and L. Ziska 2016). Wildfires emit particulate matter and ozone precursors which decrease air quality. Low air quality harms the human respiratory and cardiovascular systems and decreases overall well being and productivity.

Climate change is also projected to increase the number and severity of extreme natural disasters. Natural disasters threaten physical and mental health resulting from "damage to property, destruction of assets, loss of infrastructure and public services, social and economic impacts, environmental degradation, and other factors" (Crimmins and L. Ziska 2016).

References

Ashley, Walker S, and Alan W Black. 2008. "Fatalities Associated with Nonconvective High-Wind Events in the United States." *Journal of Applied Meteorology and Climatology* 47 (2): 717–25.

Ashley, Walker S, and Thomas L Mote. 2005. "Derecho Hazards in the United States." *Bulletin of the American Meteorological Society* 86 (11). American Meteorological Society: 1577–92.

Crimmins, J. Balbus, A., and Eds. L. Ziska. 2016. "The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment." *USGCRP*. U.S. Global Change Research Program, 312.

Environmental Information, NOAA National Centers for. n.d. "Billion-Dollar Weather and Climate Disasters: Overview."

Gall, Melanie, Kevin A Borden, and Susan L Cutter. 2009. "When Do Losses Count? Six Fallacies of Natural Hazards Loss Data." *Bulletin of the American Meteorological Society* 90 (6). American Meteorological Society: 799–810.

Gensini, Victor A, and Walker S Ashley. 2010. "An Examination of Rip Current Fatalities in the United States." *Natural Hazards* 54 (1). Springer: 159–75.

NWS, NOAA. n.d. "NATIONAL Weather Service Instruction."

Terti, Galateia, Isabelle Ruin, Sandrine Anquetin, and Jonathan J Gourley. 2017. "A Situation-Based Analysis of Flash Flood Fatalities in the United States." *Bulletin of the American Meteorological Society* 98 (2): 333–45.