Full REU Paper Rough Draft Mark Mimms

**Abstract**

In an age of “Big Data”, data dissemination is a growing topic of interest for natural hazard researchers. As data collection grows in both quantity and quality, the need for an actively managed repository where researchers can access this data becomes clear. DesignSafe-CI is designed with researcher collaboration in mind, and allows for users to make available their own data to their peers as well as peruse other datasets that might be beneficial in their own work. This paper showcases the collaboration capabilities of DesignSafe-CI by demonstrating the creation of a graphical user interface (GUI) to publish curated tornado damage assessment data collected by Dr. David Prevatt and his team after the 2015 tornado outbreak in Dallas, TX. By showcasing the data visualization and analysis tools developed for this dataset, future users may adapt and build their own systems that similarly provide easy access of their datasets to other natural hazards researchers.

**Introduction**

Data collected in the field is a core necessity for natural hazard researchers, whether it be for wind events (e.g., Masters et al., 2010),tornados (e.g., Haan et al., 2010), or other design level phenomena. Demand for these datasets is high, however there are many disparate methods for curating and disseminating the field data such as uploading it to personal websites or project hubs (e.g. <http://fcmp.ce.ufl.edu>) which complicates the ability of researchers to share data and collaborate on research projects.

As data collection methods have changed and improved over time, more high-quality data is being collected yet ways to maximize the effectiveness of these datasets have lagged behind. DesignSafe-CI, developed by the NSF’s NHERI program (NSF solicitation #15-598), is an online resource that has been developed to help close this gap; researchers can make their data available for use in other research projects, and they can also peruse other uploaded datasets that can be used to help answer their own research questions. DesignSafe-CI provides a much needed forum for data-sharing and collaboration, yet there remains a need for a standardized, “best-practices” method for uploading data and ensuring that the data is of high-quality and organized in a logical, user-friendly manner. Work has already been started on addressing this need as seen by Gurram, et al. (2017) which described their data analysis and use of DesignSafe-CI in regards to data collected during Hurricane Matthew in October 2016.

In order to realize the full potential of the DesignSafe-CI workspace, it is necessary to develop an organized framework for the curation and dissemination of data that will contribute to answering questions regarding the best way to promote collaboration between researchers. This work will create an efficient data-sharing pipeline to satisfy the demand for high-quality data collected in the field and link together the researchers who collect it, and those who wish to use it to answer the pressing research questions in the natural hazard engineering field.

This paper will demonstrate a data curation and dissemination methodology created for large amounts of data collected during research on major tornados (EF4 & EF5) as categorized by the Enhanced Fujita Scale (<http://www.spc.noaa.gov/faq/tornado/ef-scale.html>), examples of which include the tornados in Tuscaloosa, AL & Joplin, MO. (Prevatt et al., 2011, Roueche & Prevatt, 2013)**.** Proven data quality control methods such as pruning data resulting from malfunctioning equipment and more as seen in Balderrama et al. (2012) and a novel curation methodology that logically organizes the data will be paired with a GUI (Graphical User Interface) developed in MATLAB and uploaded to DesignSafe-CI, where users will be able to access the metadata and visualize specific datasets to ascertain their usefulness for the researcher’s own projects. The dataset used involves large amounts of forensic engineering analysis and other data such as damage photos collected in the field by Dr. Prevatt in the aftermath of the tornado outbreak in winter of 2015 in Dallas, TX. Data of this type has been collected after tornados as mentioned previously in Tuscaloosa and Joplin, and has been used in empirical fragility studies and measuring the impact on residential structures (e.g. Roueche et al., 2016, Roueche & Prevatt, 2013)

**Methodology**

Data used in this project was collected in the aftermath of a tornado outbreak in Dallas, TX during December of 2015. Tornado damage assessment methodology have changed greatly over the years, and the Dallas dataset represents the most detailed and thorough collection yet. This dataset has previously been available to outside researchers only by making contact with Dr. Prevatt and working out the details of data sharing on a case-by-case basis. This paper will establish the benefits of a systematic approach to data curation and dissemination via DesignSafe-CI and demonstrate a repeatable method for reproduction with other datasets compiled in the natural hazards engineering fields.

**Brief Literature Review of Previous Methodologies**

During the literature review two general formats were observed depending on if the study was a review paper (e.g. Balderrama et al. 2011) or if the focus of the paper was more experimental (e.g. Gurram et al. 2017). Both of these papers provide excellent examples and deserved further consideration because of their contribution to this paper.

*Review of Balderrama et al. 2011*

Balderrama et al. 2011 was a review paper on the Florida Coastal Monitoring Program which has collected hurricane wind data since 1999. Its methodology section was broken into three main parts concerning the motivation, background, and infrastructure of the data collection project. This format best suits review papers that seek to provide context as to where the measured data comes from, why it was collected in the first place, and the specifics of the data collection so that limitations of the datasets can be identified and accounted for. More resources were devoted to the technical description of the project and the instruments used, and less for the uses of that data (a subject expounded on in a later section).

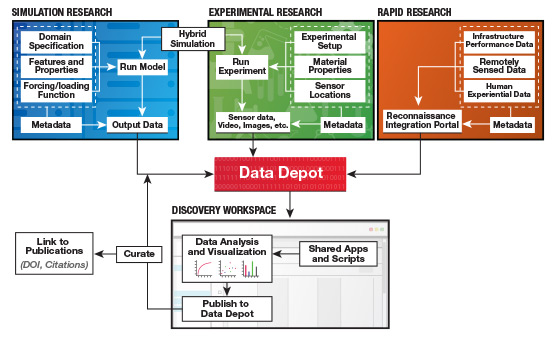
*Review of Gurram et al. 2017*

This paper dovetails with the purpose of this project, and as such is an excellent reference. Gurram et al. excels in its use of graphics to give the reader a visual indication of how the methodology was formed. Whereas Balderrama 2011 et al. was a review of the methodology used by the Florida Coastal Monitoring Program, the Gurram paper used their own unique data collection process, which was laid out in detail in their methodology section. The methodology section served to provide a chronological roadmap of how the researchers performed their work and how they arrived at their stated conclusions. The paper relies heavily on the use of graphics and tables to augment their characterization of the work performed, and to provide visual aids to help comprehension like screenshots. Examples of other aids are flow charts as used in Gurram et al. (Fig. 1).

|  |
| --- |
|  |
| **Figure 1: Example of a flow chart used in Gurram et al. 2017 to display a GUI layout** |

**DesignSafe-CI**

DesignSafe-CI is one part of the Natural Hazard Engineering Research Infrastructure (NHERI) project, which is funded by the National Science Foundation (NSF solicitation #15-598). DesignSafe-CI is online hub for the NHERI program, and is the cyberinfrastructure component of NHERI. ([www.designsafe-ci.org](http://www.designsafe-ci.org)).



**Figure 2: Flow chart provided on the DesignSafe website detailing project workflows. Image courtesy of DesignSafe-CI.org.**

This paper focuses on the process of moving from experimental research through the data depot to the workspace in DesignSafe where users will interact with the data. DesignSafe has cloud computing capabilities that allow for the storage of the data in the cloud so that it can be analyzed directly from the workspace using the tools available in the GUI’s. Rathje et al. (2017) details the creation of the cyberinfrastructure and website. The DesignSafe website offers general guidelines for data curation, including minimum metadata requirements, projected timelines for the publication of data, and data publishing, among others. In all, the DesignSafe infrastructure is a cutting edge tool designed to ease the process of collaboration amongst researchers, and is provided with the capabilities to empower researchers to accomplish that goal.

**Overview of Development Method**

*Data types that will be categorized*

In the Dallas dataset, the data collection methodology was quite rigorous and produced a large amount of data for each location surveyed. During the development on the GUI, it was determined through consultation with Dr. Prevatt that only a sub-set of these variables should be included in the program. Concentrating on only the most important information that researchers who are using the program would be interested in fits with the stated goal of providing the user a concise overview of the dataset. For the Dallas dataset, the following 13 columns were retained for use:

* Complete Address: Alphanumeric string of surveyed home’s street address.
* Year Built: Construction year of the home.
* Wind Speed (mph): Assigned wind speed based on Degree of Damage assigned.
* Degree of Damage: Damage assessment rated from 0 (no damage) to 9.
* EF Rating: Damage assessment based off of Degree of Damage rating and the assigned wind speed.
* Distance from Center (ft): Calculated distance of the home from the tornado centerline.
* Longitude: GPS measured longitude of the home.
* Latitude: GPS measured latitude of the home.
* Roof Type: Description of the home’s roof structure.
* Roof Covering: Description of the material used for the roof covering.
* Number of Stories: Number of the home’s stories or floors.
* Street Number: Column used for data sorting purposes only, refer to Complete Address.
* Street Name: Column used for data sorting purposes only, refer to Complete Address.

*Data organization*

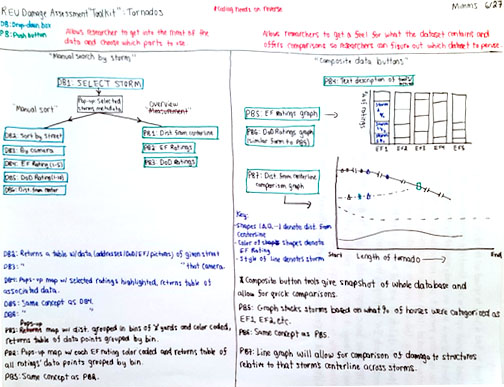
From a development standpoint, the back-end organization of the data will be in what is referred to as a “Master Table” that is used to provide the data that the user selects through the GUI. This “Master Table” is the end result of the work done to organize the given data in a method that is logical and concise. As seen in Figure 4, the data imported from the Dallas dataset was organized

|  |
| --- |
|  |
| **Figure 4: MATLAB “Master Table”, the foundation for data analysis tools in the GUI.** |

and rearranged into a table format in MATLAB, with 712 rows representing the homes surveyed in Dallas. This table is the foundation of the storm-specific GUI. Every analysis tool that was developed references the table to retrieve the appropriate data for display, based on the user input.

*Development of Data Analysis Tools & GUI Layout*

|  |
| --- |
| **Figure 5: GUI Tool Development Planning** |

From a development standpoint, the product used in DesignSafe-CI was developed in a “bottom up” method; the GUI that displays the specific parts of data the user wished to visualize was created first and then the system was built backwards from there, with the final piece being the user-friendly GUI the user sees first when operating the program. Figure 5 is an example of the design process for the MATLAB GUI analysis tools, a process that went through many iterations. On the left hand side of the paper a structure that appears after the user has selected the storm they wish to view from the splash page, which was not yet designed at that point in time. That structure, after much trial and error, developed into the main user interface seen on the final product. On the right hand side of the page are mock-ups of potential data visualization tools in the form of graphs comparing data points across storms. These features were discarded at a later time, another result of the design process.

This concept of working backwards from the end result allowed for the focus to be on providing the tools researchers need to determine if the dataset fits their need, and then moving on to creating a user-friendly pathway to get there. Two different types of data analysis tools were developed for the Dallas dataset, those focused on visual representation of the data and those which provided a user-selected subset of the data. Interactive web maps with the data layered on top thanks to the geo-tagged photos comprise of the first set of tools. These large maps were developed as overview tools so the researchers could see the dataset as a whole spatially organized on a map, with the ability to selectively filter by the amount of damage each home received.

|  |
| --- |
|  |
| **Figure 6: Listbox example** |

For the second set, sorting tools were developed so that the user could inspect in greater detail specific portions of the dataset that might be more pertinent to helping them answer their research questions. Many options were considered as the design process went through multiple iterations, with the final five chosen being the best at conveying the meaning of the data through different stratifications. The tools were created as listboxes that allow the user to select on or multiple options of a listbox, with the selected data appearing in table form. An example of one these sorting tools is provided in Figure 6.

The sorting options that were decided on are as follows:

* Sort by Street: Allows user to sort dataset by each street that data was collected for, or the entire dataset in alphabetical order.
* Sort by EF Rating: User can sort the entire database by ascending order in EF Rating, or can choose one or more individual ratings to display.
* Sort by DoD Rating: User can sort the entire database by ascending order in Degree of Damage rating, or can choose one or more individual ratings to display.
* Sort by Roof Type: User can choose to see only the data of one of the two observed roof types.
* Sort by Year Built: User can sort the entire database by ascending order of year built, or can choose one or more options, grouped into bins of 10 year periods.

*Migration of GUI to DesignSafe-CI*

Work continues with a liaison from DesignSafe-CI to safely and securely migrate not only the GUI, but also the entire Dallas dataset including geo-tagged photos to DesignSafe-CI. Once the migration is complete, any user of the DesignSafe-CI website will be able to use the program and if they find the data to be useful for their research they can export a copy of the dataset to their own research workbench. Complete with up to 100 GB of cloud storage, the research workbench provides a place for those researchers to then dive further into the entire dataset, using their own analysis tools or pulling out sub-sections of the dataset that fit their specific needs.

**Demonstration**

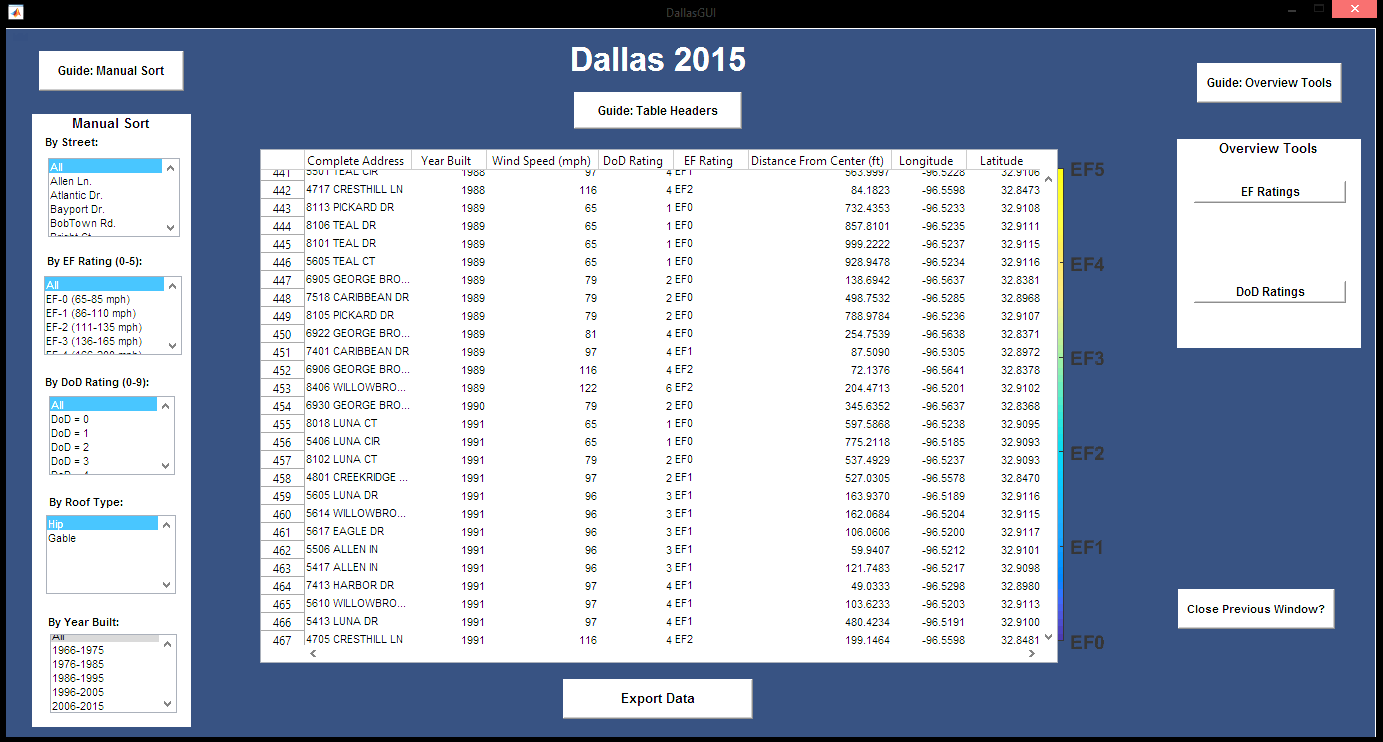
*Tornado Front End GUI*

|  |
| --- |
|  |
| **Figure 7: Tornado Front End User Interface** |

The first thing a user will see when opening up the program is the main splash page, as seen in Figure 7. This page is designed to guide the user through the use of the program and learn to get a feel for the GUI structure. In future iterations will serve as the home base for users to see which tornado datasets are available for viewing and compare across storms.

Users would begin by clicking on the Learn More button, which provides a link to more information about the datasets collected by Dr. Prevatt over the years. Since the focus of the paper is on showcasing the available tools, this demonstration will continue along the same theme and use the Dallas dataset. Many tools shown on this page are currently disabled and await future datasets to become available so that they may be put in use.

*Dallas GUI*

The Dallas GUI consists of three main sections, the sorting tools on the left, the data table in the center, and the overview tools on the right of the screen. Referring to Figure 8, each of these three sections has guide buttons built in above them to aid the user in understanding what each entails. 

**Figure 8: Dallas GUI display. Table displays all data, sorted by year built.**

As detailed earlier in this paper, the user would first use the overview tools, located on the right hand side, in order to get a visual glimpse of the dataset contained within. In this demonstration, the user has selected to view the homes that were assigned Degree of Damage ratings of seven and higher, perhaps their research deals with extreme structure failure and thus the user wishes to see the most damaged homes in the dataset. As seen in Figure 9, the user’s selections have been populated on an interactive web map.

|  |  |
| --- | --- |
|  |  |
| **Figure 9: On the left, the user selects three filters. On the right, that data is plotted.** | |

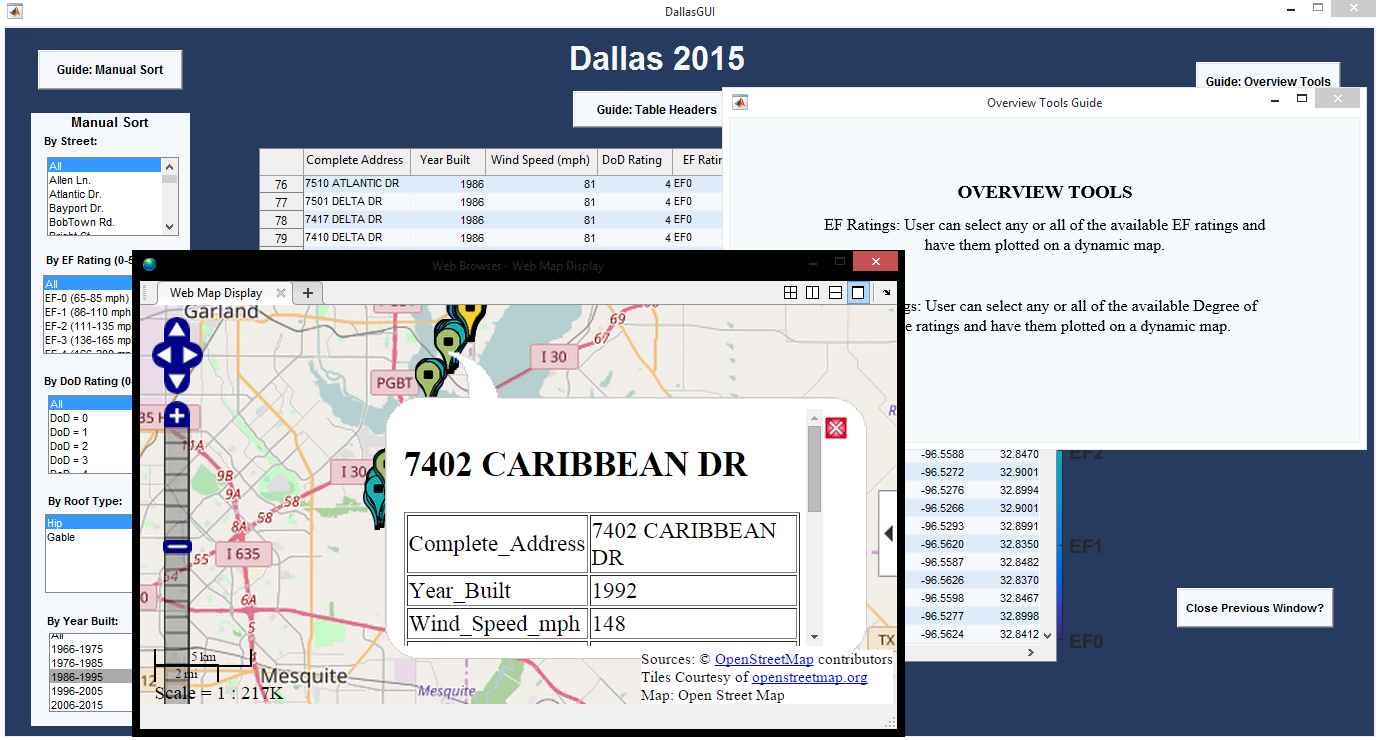
On the web map, the user has selected a data point. This brings up individual details about that home, such as its damage ratings and address. Researchers can quickly use the overview tools to hone in on the specific areas they are looking for, and in essence retrieve its metadata, albeit in an easy to read format.

After the user has clicked through the overview tools and their large-scale web maps and decided that this dataset might be useful in their own work, they are able to return to the Dallas GUI and delve deeper into the data via the use of the sorting tools on the left hand side. An example of the in-depth analysis possible is demonstrated in Figure 10, where the user has selected the sorting option containing homes built between 1986 – 1995 in the Year Built listbox.

In this example, the user had also previously clicked on the guide button for the overview tools, which remains displayed here even after they have used those tools as seen above. The data that the user requested is displayed in the table in the middle, and a web map showing just those data points is created. This smaller sub-set of data is now available for the researcher to analyze both visually through the specifically generated web map and the data displayed in tabular form on the main page of the GUI.

**Figure 10: Display demonstrating the pop-up features of the GUI, such as tool descriptions and a web map with data points plotted that match the sort criteria chosen.**

This functionality is available for all of the sorting options, and could be especially useful when a user wanted to sort by multiple streets. A user selecting to see the data of three streets would have those three web maps appear in the same pop-up window, and the associated data from all three streets would be passed into the data table in the GUI for the user to see.



With this demonstration of the ability to sort and visualize the Dallas dataset, it is possible to see how this concept can be applied not only to other tornado damage assessment datasets, but also those of other natural hazard researchers, such as datasets of hurricanes and earthquakes. Allowing the user to see both the big picture and dig into the specifics of the data available from the damage assessment in a quick and easy fashion is crucial to advancing the collaboration between researchers.

**Discussion**

Due to the volume of data that researchers collect, an actively managed repository such as DesignSafe-CI is needed to provide a foundation for data sharing amongst natural hazards researchers. Issues prohibiting the expanded use of data sharing in today’s research environment is the disparate methods to gain access to such data, each researcher has a different way of choosing how much of their data to share and with whom, and the difficulty of understanding the metadata that describes the data when they are able to access it. DesignSafe-CI is designed for the first issue commonly seen, and graphical user interfaces (GUIs) alleviate the second issue by visually explaining what data was collected and how.

As adoption of these data-sharing models grows, the barrier to entry will be lowered and collaboration will become more commonplace. Researchers will be able to share what they choose and how to do so through creation of their own software, and it will all be hosted online, saving space and computing power for their current projects.

The next step for this particular GUI is to coordinate with DesignSafe-CI to migrate the product onto their site and ensure full functionality of the data analysis tools and the ability to import the selected data into a researchers workbench, should they find it useful for their own studies. Future iterations of this GUI will enhance the user experience and improve the functionality of the tools currently available, such as allowing for cross-filtering across multiple sorting options. As the number of collaborators and the quality of the data dissemination grows, the pace at which natural hazard research progresses will be quickened and the more researchers will be able to give back and help protect the communities across the world most harmed by natural disasters.

**Acknowledgements**

Appreciation for and acknowledgement of the NHERI REU Program, sponsored by the National Science Foundation, is warranted for the funding and organizational support of this paper. Special thanks is needed for faculty mentors Dr. Kurt Gurley & Dr. David Prevatt, along with Dr. Karina Vielma, for their valuable guidance throughout the research process. Finally, thanks must be given to the MathWorks webinars and forums community andwhich helped with technical aspects of the program and to Florida International University for hosting the REU program for the presentations.

**References**

1. Balderrama, J. A., Masters, F. J., Gurley, K. R., Prevatt, D. O., Aponte-Bermúdez, L. D., Reinhold, T. A., ... & Chowdhury, A. G. (2011). The Florida coastal monitoring program (FCMP): A review. *Journal of Wind Engineering and Industrial Aerodynamics*, *99*(9), 979-995
2. Balderrama, J. A., Masters, F. J., & Gurley, K. R. (2012). Peak factor estimation in hurricane surface winds. *Journal of Wind Engineering and Industrial Aerodynamics*, *102*, 1-13.
3. Gurram, Harika, Subramanian, Chelakara, Pinelli, J.P., Basu, Rituparna (2017). Processing Data From Hurricane Matthew In DesignSafe-ci. *The 13th Americas Conference on Wind Engineering, May 24th, 2017*.
4. Haan, F.L., Jr., Balaramudu,V.K., and Sakar,P.P. (2010). “Tornado induced wind loads on a low-rise building.” J. Struct. Eng., 136(1), 106–116.
5. Masters, F. J., Tieleman, H. W., & Balderrama, J. A. (2010). Surface wind measurements in three Gulf Coast hurricanes of 2005. *Journal of Wind Engineering and Industrial Aerodynamics*, *98*(10), 533-547.
6. Prevatt, D. O. (2013). Joplin, Missouri, Tornado of May 22, 2011. American Society of Civil Engineers,
7. Prevatt, D. O., Roueche, D. B., Van De Lindt, J. W., Pei, S., Dao, T., Coulbourne, W., & Grau, D. (2012). Building damage observations and EF classifications from the Tuscaloosa, AL, and Joplin, MO, Tornadoes. In *Structures Congress 2012* (pp. 999-1010).
8. Rathje, E., Dawson, C. Padgett, J.E., Pinelli, J.-P., Stanzione, D., Adair, A., Arduino, P., Brandenberg, S.J., Cockerill, T., Dey, C., Esteva, M., Haan, Jr., F.L., Hanlon, M., Kareem, A., Lowes, L., Mock, S., and Mosqueda, G. 2017. “DesignSafe: A New Cyberinfrastructure for Natural Hazards Engineering,” ASCE Natural Hazards Review, doi: 10.1061/(ASCE)NH.1527-6996.0000246
9. Roueche, David B., and David O. Prevatt. "Residential damage patterns following the 2011 Tuscaloosa, AL and Joplin, MO tornadoes." *Journal of Disaster Research* 8.6 (2013): 1061-1067.
10. Roueche, D. B., Lombardo, F. T., & Prevatt, D. O. Development of Empirically-Based Fragilities of Residential Damage in the 2011 Joplin, Missouri, Tornado. In *Geotechnical and Structural Engineering Congress 2016* (pp. 1283-1294).