Responses to RC2

Enclosure:

Response letter to the reviewers' comments

- ~~Red~~ texts are removed from the manuscript
- Green texts are added to the manuscript

1. **The manuscript is difficult to follow. There are pieces of information on the same thing throughout the manuscript rather than organized in the same section. As a result, I had to read back and forth to get a clear view. For example, depression extraction is described in Sections 2.1.1 and 4.1 storm event clustering is described in Sections 2.1.4 and 4.2. It would be easier for the reader to follow the methodology if the information were presented concisely and organized.**

To address these concerns, we've revised the manuscript to put all of the method details in section 2 and leave section 4 for results. Specifically, the methodology is explained in Sections 2.1.1 and 2.1.4, while the results are presented in Sections 4.1 and 4.2. Sentences that repeat methodology are removed from Sections 4.1 and 4.2. Lines 367, 372, 375, and 378 are updated.

2. **Section 2.1.1: It is vague how depth extraction was used to find flooding that can cause travel disruption. Was the depth of depressions at L2-1 and L2-2 considered to determine if there was any impact? The depth of flooding at the different levels of depressions is not mentioned in the manuscript.**

We included only the depressions, as well as the depression levels, that were close to or larger than the scale of a road and could affect traffic flow. Depth of depressions is not used to find flooding. Depressions that could cause flooding are identified by visually investigating depressions' alignments on the road surface, micro topographic features and their area, as well as whether a cluster of Waze alerts is assigned to them. In the example presented in Figure 2, L2-2 is not chosen as roadway flooding because it does not overlay the road surface. Since the depth and level threshold criteria have not been explained in the manuscript before line 122, the sentence shown in red below is removed to resolve the confusion.

"Due to the complexity of urban terrain, the spatial scale of depressions at each hierarchy level is quite variable, and depressions at the same level can be as large as a neighborhood or as small as a pothole. ~~Therefore, we did not set an automated stopping criterion in terms of depression level or depth for the depression filling process. Instead,~~ Initially, depressions at all hierarchical levels are extracted, and the level that has depressions that are at the scale of, and best align with urban features, including roadway curbs and gutters, was manually selected.

Flood-prone depressions are then identified by examining overlays of the depressions and Waze flood reports, as well as the depression area and road surfaces that the depression covers. The procedure for using Waze reports to identify depressions is presented in detail in Section 2.1.5."

3. **Crowdsourced flood reports like Waze often have multiple reports surrounding a flooded area for the same event. There was no mention of removing duplicates in the manuscript. Not removing duplicates could falsely increase the probability of flooding on a depression.**

Reporting of depression flooding is considered as a binary presence/absence variable and encoded as 1 if there is one or more flood alert and 0 if there is no flood alert. This process removes duplicates. To clarify this, the following statement is added to the manuscript.

Line 201:

Pluvial flooding on any given surface depression can be modeled as a Bernoulli trial of flood failure (i.e., non-flooded) or success (i.e., flooded). If a depression has one or more Waze flood alerts linked to it, the depression is labeled as flooded (success). Assuming that the probability of being flooded is smaller than the non-flooded situation and that the likelihood of flooding in a particular storm event for each depression is independent of the probability of flooding for other depressions, a random variable $y_{i,j}$ will define the count of successes (flooding) out of the N trials (N storm events of cluster j) on depression i.

4. **Line 427-431: The comparison is not clear. Equation 11 used "yi,j is the predicted number of floodings on depression i and storm type of j". Isn't the predicted number of flooding derived using historical flood reports? In that case, the likelihood of flooding should be higher when there were Waze reports. If the purpose is to evaluate model performance in predicting flood probability, some performance metric should be used.**

As the reviewer notes, the predicted number of floodings is derived from historical flood reports, which means that the likelihood of flooding should be higher when there were Waze alerts posted. Figure 15 is intended only to demonstrate that the predicted likelihoods are plausible, but we cannot evaluate model performance explicitly since the data do not contain true negatives (i.e., flooding may be occurring in locations where there are no Waze reports).

Note that in responding to this comment, the authors realized that the description of Figure 14 could easily be made just with Figure 15. Therefore Figure 14 is removed from the revised manuscript.

5. **Line 437-440 and Figure 16: This part needs further clarification. The probability of flooding and jam could be shown of two maps, if they are overlapped. Jam level 3 is**

not clear on the map. Please consider changing the color. The jam during the storm event should be compared with the jam during the same time and same days to conclude it happed due to the flooding.

The map layout and colors are changed to make the levels clearer, along with zooming in to the example intersection. Visual comparison between the traffic jams at the same time and day of the week for the following and preceding weeks are added to the map. The manuscript is updated as follows:

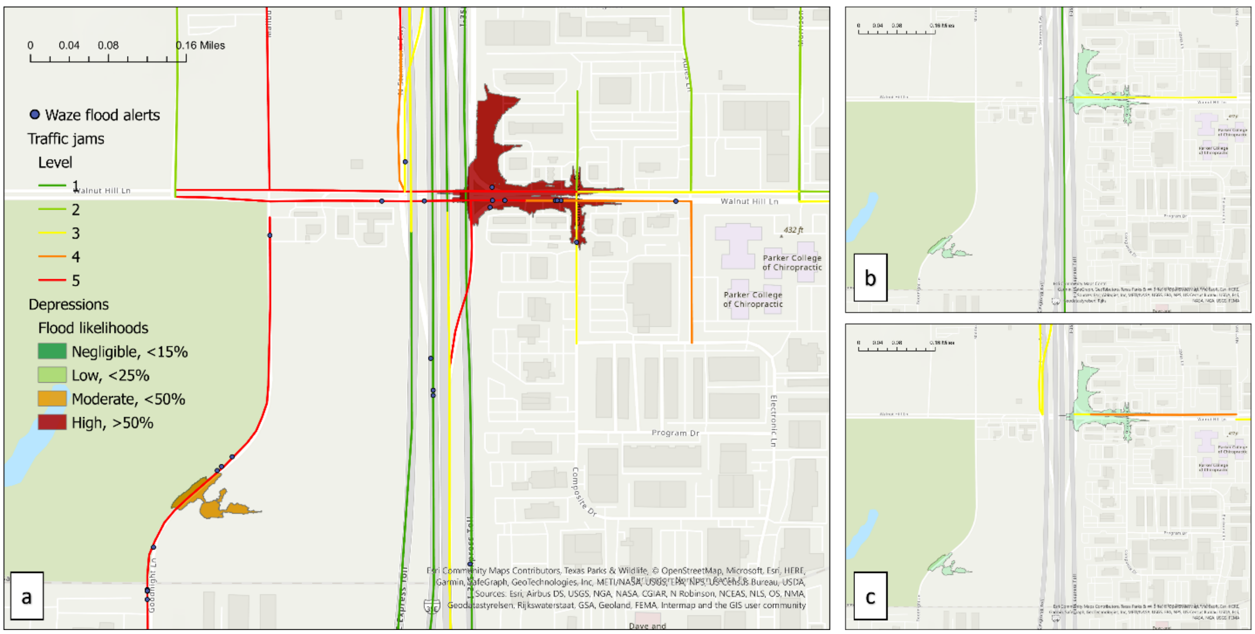

Figure 16. Severe storm PFF probability map versus flood alerts and traffic jams on a. Friday, September 22nd (the date of a severe storm), b. Friday, September 29th, 2018. and c. Friday, September 15th, 2018

Line 434:

Figure 16-a shows an example of a flood probability map for severe storms, along with historical flood-related alerts and traffic jams reported by Waze during one particular severe storm that occurred on September 22nd, 2018. Figures 16-b and 16-c show the same information during the same time and day of the week for the following and preceding weeks. Waze traffic jam reports include severity and congestion levels ranging from 1 (lowest) to 5 (highest), which denote the level of traffic slow down or complete shutdown. Negligible, low, moderate, and high flood probabilities are defined as less than %10, less than %30, less than %50, and higher than %50, respectively. In Figure 16-a, high traffic levels (Waze jam levels of 5) can be seen near a depression with high PFF probability (more than %50). Figure 16 indicates that traffic jams during severe storm are noticeably higher than at similar time intervals before and after the storm. These maps suggest that the traffic jam on the storm date, which agrees with the flood likelihood, is likely to be an anomaly relative to typical traffic conditions at this intersection. This finding is consistent with the flood alerts and predictions of severe flooding at this location during the storm.

Minor Comments:

1. **Line 22-23: This sentence should go to the last paragraph of Introduction. The first paragraph in Introduction usually provides a general background of the problem rather than specifying the goal of the study at the very first sentence.**

The authors believe that providing the goal of the study at the start of the paper helps to focus the reader on the relevant portions of the background and motivation. We would prefer to leave the sentence where it is, but if the editor prefers that we make this change then we will do so.

2. **Table 2. There is no definition of the storm clusters until Section 4, line 376.**

To address, line 170 is updated as follows:

Depending on storm events' severity and terrain characteristics, storms can produce similar patterns of depression PFF. To capture this phenomenon, storms are clustered into classes based on their severity (light, moderate and severe) using storm characteristics. For storm clustering, agglomerative hierarchical clustering is applied using a bottom-up approach that forms a single cluster for each storm event and successively merges clusters with the smallest distances between features. The benefit of using agglomerative clustering is that this algorithm is less sensitive to outliers (Edelbrock, 1979).