Response to Reviewers' Comments for TRC_2019_1394

We would like to thank the three reviewers for their constructive comments, which significantly improved our manuscript. In this document, we detail how we have addressed all the comments made by the reviewers. For the reviewers' convenience, we start our response to each reviewer in a new page. We reproduce a reviewer's comment in a grey box, then provide our response in **bold**. Furthermore, we highlight the changes in the manuscript in **blue**, with the exception of minor edits on typos and language issues.

Response to Reviewer 1's Comments

> The paper reads well and this study explores the association between truck crashes and safety-critical events using crash reports and naturalistic driving data. The topic is interesting. However, the reviewer has some concerns.

Thank you for your careful review of our paper. We appreciate your comments and suggestions. Below, we provide our point-by-point response to your comments.

> 1. Justification of the research gap is weak, especially the introduction. Besides, what are the safety-critical events (SCEs)? Could you provide a definition?

To address your comment, we now integrate the literature review (previously Section 2) with the introduction in a new subsection titled "1.1. Research gaps" To clarify the research gap, we have specifically: (a) added three columns into Table 1 highlighting the data used in the literature, study size, and crash surrogates (and observed effect direction), and (b) included a paragraph explicitly highlighting four main gaps in the literature based on Table 1.

As for the safety critical events (SCEs), we now:

- **Clarify that SCEs are routinely captured in naturalistic driving studies (NDSs):**

  In view of these limitations, naturalistic driving studies often use surrogate events in place of actual crashes. Such surrogates are usually referred to as safety-critical events (SCEs) and are selected to intuitively represent (more numerous) "near misses", i.e., special types of accident precursors that have all features of accidents, except that potentially catastrophic outcomes were avoided by last-second evasive maneuvers (Dingus et al. 2011, Saleh et al. 2013).

- **State the sensor-based monitoring system used for collecting the SCEs:**

  At the time of data collection, the company's entire fleet utilized the Bendix® Wingman® Advanced™ monitoring system. Both the data management and system maintenance operations were performed by the company. Note that this system is also widely used by other U.S. based trucking operators.

- **Define the SCEs in Section 1.2:**

  The sensor-based monitoring system used by the company captures four different kinematic events:

  - Headway, which signals an instance of tailgating for at least 118 seconds at an unsafe gap time (a measure of distance between leading and trailing vehicles) of 2.8 seconds or less (Grove et al. 2015).

- Hard brakes, which are defined as instances of a deceleration rate of 9.5 miles per hour per second or more.
- Activation of the rolling stability system, which intervenes by applying brake pressure (in addition to potentially applying trailer pressure) assisting the driver in aligning the vehicle when the system's critical thresholds are approached (Bendix® 2007).
- Activation of the forward collision mitigation system.

2. The introduction is not good and some sentences make them confused. The authors would be suggested to improve it largely.

We have completely rewritten the Introduction. It is currently divided into three main parts: (a) an overview of the importance of trucking safety, how sensor-based technologies can overcome some of the limitations in traditional reporting and the motivation for using SCEs as surrogates for crashes; (b) a subsection detailing the research gaps; and (c) a subsection detailing the study focus, goals and contributions. Based on these substantial changes, we believe that this version is more clear and convincing.

3. Authors also kept silent regarding data quality & integrity.

What is the percentage of drivers excluded?

How about the accuracy of the GPS data? As we know, sometimes, the coordination of the GPS may be far away from the actual location, even the other side of the road.

How to choose the thresholds? Any references?

The quality of crash data in this manuscript should be reported.

There are several aspects that we would like to highlight in how we addressed these comments:

- **Data section: We now include a "Data description" section where we provide some background on the company, summarize the drivers' characteristics, detail the data acquired from the sensor-based monitoring system, and explain our data aggregation process.**

- **Overall data quality considerations: By explicitly stating the name of the sensor-based monitoring system (Bendix® Wingman® Advanced™ monitoring system), we provide researchers with a detailed ability to replicate/test our data collection procedure.**

- **Percentage of excluded drivers: We have added the following description of total number of drivers and the exclusion criteria that we used.**

  The original dataset provided by the company included 34,348 drivers. We have excluded 2,520 drivers (i.e., 7.4% of the original dataset) from our analysis if any of the following criteria is met: (a) driver inactivity, where we required the driver to have at least 100 GPS pings in the data to be included; (b) the unique identification code for the driver is not found in the provided demographics table; and/or

(c) the number of SCEs reported were identified as obvious outliers (we only removed drivers who had an unrealistically large number of SCEs). Hereafter, all reported data will correspond to only those generated by the remaining 31,828 drivers, whose characteristics are summarized in Table 2.

- **Accuracy of GPS data:** **The GPS information was provided by our industrial partner. The GPS data provided by the Company included a GPS quality indicator, which showed that 98.7% of the data to be of "good quality" for the purposes of vehicle tracking and routing. We do acknowledge that general GPS devices used in motor vehicles are typically accurate within the 3.0 − 4.9 meter range (depending on the device used). However, for our purposes even such general GPS devices are adequate since: (a) GPS data is solely used in estimating the distance traveled in a trip; (b) trips are made up of a large number of pings (79.8 on average); and (c) distance is estimated based on sequence of pings from the beginning to the end of the trip. It is important to note we do not consider typical variables that can be affected significantly with the GPS accuracy, such as (number of lanes, traffic flow, etc.), given that the primary focus of this paper is examining the association between the SCEs and crashes.**

- **Thresholds:** **We now provide the thresholds for both the headway and hard brake SCEs. We do not provide thresholds for the other two SCEs given that they constitute proprietary information from Bendix®; however, by providing the name of the sensor-based system (and a reference to its user manual) we present all the information needed to capture our data capturing mechanism.**

- **Quality of crash data:** **The crash data were captured by a leading trucking company. Per the collected data, the company seems to well exceed the reportortable crash guidelines set by the federal government which defines "a reportable crash is one in which a vehicle was towed from the scene, or an injury or fatality occurred." Hence, we consider the crash data to be of high quality. That being said, we also note that a possible limitation of our work is that crashes may be under-reported in Section 5.3.**

4. In 4.1, why not report the driving experience (e.g., driving years)? Why not report the statistical summary of the variables?

We now include tables of the summary statistics for the main predictor variables and the covariates. In Table 2, we provide a summary of driver characteristics including their average age ± SD, number of drivers per gender, business unit and driver types (with their % in parentheses). Furthermore, we provide a summary of the computed response, predictor variables and covariate in Table 4. While the company did provide some data pertaining to "driving years", we elected not to include this in the table of driver demographics since the missing rate of driving years, in the demographics table provided by the company, is more than 30%. From a closer investigation of the column, it seemed that the company started to collect this information recently; drivers who have been employed by the company for over three years typically did not have any information for that variable. Hence, we do not include that variable in our manuscript.

5. Were all variables included in the models?

   So the authors did not consider the multicollinearity?

   And the tables made the reviewer confused.

We used different combinations of predictors and covariates in the different models. To make this point clearer, we have: (a) added a new table (Table 5), where we show how the data are sampled and modeled to examine the three research questions; (b) used different mathematical symbols in equation 2 to distinguish between the main predictors, i.e. the SCEs, and the covariates – we explicitly stated that the number of predictors and covariates will vary; and (c) organized the tables such that they follow the model description process in Table 5.

Yes, we have checked for both the pairwise correlation and multicolinearity in the data. In this revision, we now include the following:

- **In Section 3.1**:

  The final step in preparing the data for modeling was to examine whether the number of predictors and/or covariates can be reduced based on a correlation analysis. Note that the variables included both continuous (primary predictors, age and mean ping speed) and polytomous (gender, business unit, and driver type) variables. To account for the mixed variable types, we adopted the approach of Revelle et al. (2010, 2016) to compute the Pearson, polychoric and polyserial correlation coefficients for the pairwise evaluation of continuous, polytomous and mixed variables, respectively.

  The results of this analysis were presented in Figure 3, which showed that the pairwise correlation among the predictors and covariates is generally small.

- **In Section 3.3.3:**

  Since four different SCEs and/or multiple covariates were included in the models at the same time, it is important to check for the presence of multicollinearity. Recall that in Section 3.1, we examined the correlations among pairs of predictors and found that the pairwise correlations are small. Here, we attempt to investigate whether a linear dependence exists among three or more of our variables through computing the variance inflation factors. If the regressors are uncorrelated, the variance inflation factor obtains its minimum value of 1. In statistical practice, a inflation factor less than 4 requires no additional investigation of linear dependence among the regressors and values greater than 10 indicate serious multicolinearity requiring model corrections.

- **The results from the variance inflation factor computations are now included in Table 9 in Section 4.4.**

  The results capture all three investigated outcomes (crashes, fatalities and injuries) as well as the stratification of the drivers' data set by business unit and driver type. From the table, all variance inflation factors are less than or equal to 1.3. Hence, we can conclude that the variation of each of the regression coefficients are not inflated and that there is no evidence of any multicollinearity issues (Vatcheva et al. 2016).

As for the comment about the confusion from the tables, we believe that the addition of Table 5 and the organization of the results (and Tables 6–8) according to our research questions should clear up any confusion.

6. "In the two models using the number of fatalities as the outcome variable (column 4 and 5), all 95% CIs of IRRs included one and the CIs were very wide", it means that the variables are not statistically significant?

Yes, your observation is correct. To clarify the interpretation of credible intervals in Bayesian inference, we added the following in Section 3.3.2.

In the Bayesian setting, parameters are considered as random variables that have probabilistic distributions instead of unknown fixed values, so no p-values can be reported here. The posterior mean and 95% credible intervals (CIs) of the incidence rate ratios ($\exp(\beta)$) are reported instead. The interpretation of the incidence rate ratios in this Bayesian negative binomial model is as follows: as the number of SCEs per 10,000 miles increases by one unit, the number of crashes per mile is multiplied by $\exp(\beta)$. A 95% credible interval is the interval such that is the posterior probability of the parameter of interest falling within that range given the data is 95% (McElreath 2020) If the 95% CI of the incidence rate ratio includes one, then one is a plausible value for the true incidence rate ratio in this case, and the parameter of the variable will be deemed statistically insignificant. If the 95% credible interval excludes one, then the parameter will be considered as statistically significant.

7. The analysis on the model results, such as the association between four different types of SCEs and crashes, the relationship between the SCEs (e.g., Headways) of the variables, etc., are simple and weak. The authors would be suggested to add the deeper analysis. Otherwise, the contributions of this manuscript would be limited.

In the revision, we have clarified that we ran 17 different Bayesian Negative Binomial models (capturing three different outcome variables, different predictor combinations, and subgroup analyses stratified by different business units and driver types) to address our three research questions of interest. By reorganizing the results section and regression tables according to our three research questions, we allow the reader to map our models to our research questions which address several gaps in the literature as highlighted in Table 1 and Section 1.1. Based on the obtained results, we have shown that our work makes the following contributions to the research and practitioner communities.

Main contributions of the work We've added the following paragraphs:

This work provides statistically strong and robust evidence that SCEs are positively associated with crashes and injuries among commercial truck drivers. Furthermore, this study demonstrates that the "severity" of the SCE is associated with the crash rate, where the two automated maneuvers (involving the initiation of the forward-collision mitigation and rolling stability systems) were shown to have a statistically significant larger effects on crash rates when compared to hard brakes and head way (which can be seen as less severe maneuvers/alerts).

The current study contributes to the existing literature in three respects. First, this paper overcomes the small sample size issues in previous crashes and crash surrogates papers, which typically includes 300 or fewer drivers or vehicles and fewer than 100 crashes (Guo et al. 2010, Gitelman et al. 2018). Our study involves 1,000 times as many driving hours and miles and includes more than 30,000 commercial truck drivers and 30,000 crashes; this allows us to investigate the association between four different types of SCEs and crashes, as well as stratified analyses across business units and driver types. Second, the evidence of the association between crashes and crash surrogates among truck drivers has been scarce. Our study

gives insights to this less studied field using a nationwide large-scale sample. Third, this paper explores the association between SCEs and human injuries and fatalities, which has not been investigated in previous papers but represents important research questions that require detailed study given that they constitute an important component of truck routing models used in practice (Hu et al. 2020).

Practical relevance to trucking operators: We emphasize the practicality of our data collection mechanism, where we use routinely collected data, and discuss why this data is typically untapped by trucking operators. Then, we provide the following three recommendations for how the knowledge of the association between the SCEs and crashes:

First, recent statistics indicate that more than 90% of traffic crashes are influenced by driver behavior (Federal Highway Administration 2019). While our naturalistic driving data does not include video images, it explicitly captures important behavioral factors such as driving speed, aggressive driving through headway alerts, and potential distraction/drowsiness with an increased rate of the three other SCEs (at least when compared to drivers on similar routes). Thus, trucking operators can use our estimated model coefficients, e.g., 50.4% for the initiation of the rolling stability system, in driver training/education. Second, operators can provide incentives to their drivers to reduce their recorded number of SCEs through behavioral based safety programs (Jun et al. 2007). Third, by examining the operator's historical record of SCE data, operators can develop scheduling and routing policies that attempt to minimize the number of recorded SCEs (Mehdizadeh et al. 2020, Hu et al. 2020).

8. The discussion and conclusions would be also suggested to be improved substantially.

We now combine our discussion and conclusions into one section (Section 5), where we discuss: (a) the main contributions of the work, (b) relevance to trucking operations, and (c) pinpointing the limitations in the study and opportunities for future work. In our estimation, the reorganization of those two sections and the added discussion have significantly improved the manuscript. The reviewer is referred to our response to Comment 7 and the revised manuscript for additional details.

9. Many sentences are weak / improper and make readers confused.
For instance, page 11: One unit increase in the number of any type of SCEs per 10,000 miles was associated with 8.4% (95% CI: 8-8.8%) increase in the number of crashes per mile. How do you know 8.4%?

In the revision, we have made significant changes in the document that should make our paper easier to follow. Pertaining to the specific comment for the coefficient, we can clarify this as follows. Since the number of SCEs per mile $\mu_i = \exp(\alpha_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ} + \theta_1 z_{i1} + \cdots + \theta_K z_{iK})$, the rate change of μ when the primary predictor x_1 is increased by one unit can be calculated as:

$$\begin{aligned}\frac{\mu'}{\mu} &= \frac{\exp[\alpha_0 + \beta_1(x_{i1}+1) + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ} + \theta_1 z_{i1} + \cdots + \theta_K z_{iK}]}{\exp(\alpha_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ} + \theta_1 z_{i1} + \cdots + \theta_K z_{iK})} \\ &= \frac{\exp(\alpha_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ} + \theta_1 z_{i1} + \cdots + \theta_K z_{iK}) \times \exp(\beta_1)}{\exp(\alpha_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ} + \theta_1 z_{i1} + \cdots + \theta_K z_{iK})} \\ &= \exp(\beta_1) \rightarrow \text{which is the incidence rate ratio (IRR).}\end{aligned}$$

Therefore, if the IRR is 1.084, the number of SCEs per mile μ is increased by 8.4% when the predictor x_1 is increased by one unit.

10. The website of the data sets provided in the manuscript does not work.

In this revision, we provide two links in our supplementary materials.

- **Website containing code and analysis:** **https://caimiao0714.github.io/Github-SCE-crash/.**
- **Sample of masked data: We provide a sample CSV file, where we masked the driver ID as well as the coordinates (by rounding them to one digit) at https://github.com/caimiao0714/Github-SCE-crash/blob/master/data/sample_ping.csv. The purpose of providing the sample file is to provide some insight into the ping data frequency, shape, and speed values.**

These two links were tested on different browsers with no observed issues.

11. Lastly, the review would suggest that the authors re-check their grammar and text as there are many spelling mistakes in the manuscript. For instance, line 2 in page 5: "The found that ……". This manuscript should be revised properly.

The revised paper was checked for grammar and spelling mistakes.

Response to Reviewer 2's Comments

Overall, this is a high-quality paper and is a worthy addition to literatures with some improvement. The study using emerging telematics data and the combination The paper used a large dataset to evaluate the relationship between SCE and crash risk. The topic is certainly worth investigation given the importance of the validity of crash surrogates. Some specific comments are below:

We want to thank you for your careful reading of the paper and the comments and suggestions you have given.

1. Please refrain from using acronyms and abbreviations. There are too many of them and make the paper difficult to follow.

Thank you for your suggestion. We have reduced the frequency of acronyms in our manuscript. Two of the most commonly used acronyms are used throughout the paper: naturalistic driving study (NDS) and safety-critical event (SCE). Other acronyms that have long names if typed out, such as "expected log posterior density leave one out" are used only briefly in Section 3.3; it would be very cumbersome to write (or read) this if we spelled it out each time.

2. Please provide more details on the information about the "ping". For example,
   - Is a ping a single data point or several points?
   - The paper mentioned "ping data every couple of seconds to around 5 minutes." For a particular truck/driver, is the interval fixed or varying?
   - If a vehicle was not on, will a ping still be sent? Figure 1 seems implying so.
   - One important issue is that if a SCE occurred not during at ping period, will it still show in the dataset? Some trigger based system will catch all such events, for example, when the acceleration is above 0.5G, a record will be automatically generated. It is not clear from the description whether this is the case for the data used.

Thank you for your suggestion. Below are our answers to your questions:

- **One ping is one data point giving information about the truck at that time. Variables include latitude, longitude, date, time, real-time speed, and many others.**
- **For a particular truck/driver, the intervals between pings vary.**
- **Yes, the pings are still recorded when a vehicle was not on.**
- **The SCEs were not collected in the vehicle's ping record. The SCEs were collected in a completely different and independent system. These SCEs were collected whenever the kinematic thresholds were triggered by the driver. Therefore, SCE will still be recorded even when there was no ping at that time. The sensor system and kinematic thresholds are introduced in the revised Introduction.**

Our original manuscript did not make it clear that the data were not specifically collected for a safety study. Rather, the data were collected as part of the company's monitoring and surveillance plan. Only afterward was thought given to using the data to assess road safety.

Since we were not involved in the design stage, we had to accept the decisions made regarding the data collection plan. We've added the following paragraph to make this clear.

We must emphasize that these data were collected as part of the company's ongoing monitoring and surveillance plan. The data were not collected specifically as part of a planned NDS. That said, the data are still measurements on factors that could affect safety and are therefore still valid for the purpose of answering the questions we pose in Section 1.2.

3. I have major concerns regarding the hard break, which is defined based on the speed decrease: "the speed decrease within a unit time is larger than a preset threshold value." How long is a unit time? I can only specular it is based on the time interval, couple of seconds to around 5 minutes. In reality a hard break typically only takes less than a second. Speed change over more than a few seconds most likely only represents whether the truck got to a stop instead of a hard brake. Can authors clarify how the Hard Brake and headway SCEs were defined. For example, acceleration threshold?

We describe the sensor system used by the company and some of the thresholds in the revised introduction. We've added the following paragraph:

The sensor-based monitoring system used by the company captures four different kinematic events:

- Headway, which signals an instance of tailgating for at least 118 seconds at an unsafe gap time (a measure of distance between leading and trailing vehicles) of 2.8 seconds or less (Grove et al. 2015).
- Hard brakes, which are defined as instances of a deceleration rate of 9.5 miles per hour per second or more.
- Activation of the rolling stability system, which intervenes by applying brake pressure (in addition to potentially applying trailer pressure) assisting the driver in aligning the vehicle when the system's critical thresholds are approached (Bendix® 2007).
- Activation of the forward collision mitigation system.

4. For the regression model, I suspect the four SCEs could be correlated. Can you check whether multicolineality is an issue? The results on page 15" "Four SCEs" model was not significantly better than the "Pooled" model." Could be a result of this?

This is an important point that we did not consider in the original manuscript. We have added a correlation matrix plot for the four different types of SCEs (Figure 3 on p. 13). We also conducted a variance inflation factor test for potential multi-collinearity. The Pearson correlation coefficients were all less than 0.2 in magnitude for the four SCEs, and the variance inflation factor scores were all less than 1.3 (Table 9). The only correlation that was even moderately strong was between "Driver Type" and "Ping Speed" which was 0.6. The two tests suggest there was very minor positive correlation among the rates of four types of SCEs, but did not cause significant multicollinearity issues.

5. Please provide appropriate summary statistics for the four SCEs, e.g., mean, standard deviation, and correlation among them. Hard Brake should happen at much higher frequency than other types of SCEs, which could dominate the "All SCEs" variable.

We've added the summary table (Table 4 at the bottom of p. 11) for the whole data set, including the SCEs. We've also added correlation matrix (Figure 3) which indicates that there is only minor correlation between different types of SCEs (Pearson correlation coefficient less than 0.2). In our study, the predictor variables are not the number of SCEs, but the rate of SCEs (the number of SCEs per 10,000 miles). This is because the number of events highly depends on miles driven, while the rate of events is less dependent on miles driven. Table 3 shows that the number of hard brakes, headways, collision mitigation, and rolling stability. The SCEs were dominated by hard brakes (48%) and headways (38%), not by hard brakes alone. Table 4 shows the average rates of SCEs per 10,000 miles were 6.86, 5.35, 0.21, 1.74 for hard brakes, headways, rolling stability, and collision mitigation, and there was large variability in these rates (Table 4).

6. Page 2: "Since NDS data are typically collected every 30 seconds to 5 minutes, the amount of NDS data are generally very large, which provides both an opportunity and a challenge for data analytics." This statement is not accurate. Typical NDS studies used continuous data collection method, e.g., SHRP2 NDS and 100-Car NDS collect data from ignition-on to ignition-off at 10HZ for video and acceleration data.

Thank you for catching this. We have completely rewritten the Introduction and the inaccurate statement has been revised.

7. Figure 3 and 4 are difficult to understand. Why only check the percentage of zeros? Careful exam of Figure 4 show that "the observed proportion of zero crashes located almost exactly at the center of the simulated distributions" is not true: the observed is still far from center for "DED LOC" and "DEC OTR". Overall, I don't think Figure 3 and 4 provide indispensable information on the model performance. The benefit seems to be overwhelmed by the difficulty in understanding them. Unless the authors can substantially improve its clarity.

This is a subtle, but important point. Since we are concerned about how likely our model predicts crashes when the driver has SCEs, comparing the the proportion of zeros in simulated and observed data is informative in prediction accuracy. The purpose is to guard against a model that might fit very well the mean number of SCEs but grossly over-predict (or under-predict) the proportions of zero SCEs. Figures 3 and 4 show that the simulated distributions (light blue histograms) are generally to the right of the observed zero proportion (solid black vertical line), suggesting that our Bayesian negative binomial models tend to over-predict non-crashes but under-predict the crashes, and the magnitude of the prediction bias is small. We have revised the paragraph describing the graphical presentation of the posterior predictive checks in subsection 4.3 and refrained from claiming "the observed proportion of zero crashes located almost exactly at the center of the simulated distributions" in Figure 4. Our revised manuscript says:

To investigate the models' predictive accuracy, we adopt the approach of Gelman et al. (2013, Section 6.3) who suggested simulating some function of the data and parameter, and comparing it with the observed value of a particular quantity. For our trucking safety application, we examined the proportion of zero crashes since it corresponds to a crash-free trip, which is of interest to truck drivers and operators alike. The probability of having zero crashes is, of course, an unknown quantity, but its posterior distribution can be estimated by simulating samples using Hamiltonian Monte Carlo. In this section, we limit our analysis to the models whose outcomes were crashes since the accident and fatality models indicated that our observed events were insufficient for statistical inference (based on the size of the credible interval in Table 6).

Figure 4 shows the posterior distributions, which are indicated by the histograms in light blue, for the posterior probability of zero crashes under each of the six models considered in Table 7. The observed proportion of zero crashes is indicated by the vertical line in each part of Figure 4. For all six models, the observed proportion of zero crashes was considerably less than what would be predicted by the model. Note that the magnitude of this prediction bias is small, usually around 0.015. In other words, while both models (with and without business units and driver types) perform reasonably well in predicting the mean numbers of SCEs, the model with business units and driver types does a better job predicting the proportion of zero crashes. This suggests that different business units and driver types should be accounted for in the model.

Based partly on the result from Figure 4, we ran the model with all four SCEs (model 2) separately for each of the seven business units and driver types. The corresponding posterior predictive check for zero crashes is shown in Figure 5. Here, the vertical lines are much closer to the simulated posterior distribution. This suggests that different business units and driver types should be accounted for in the model.

Minor issues:

1. Equation (1), should subscript "i" be in the right of equation for "y"as well?
2. Page 3: "The large sample size can yield statistically significant results and conclusions." Large sample is a major strength of this paper. However, large sample size does not necessarily yield statistically significant results (if there is no relationship, we probably don't want statistically significant results). More accurate statement is needed, for example: The large sample size provides high statistical power to detect potential relationship between SCEs and crashes.
3. Abstract: second line: should it be "used to measure safety" since "outcome" is a little vague.
4. Figure 3: please label the X-axis. What is PPC in the caption of the figure?

Thank you for these suggestions. We have addressed the issues as suggested.

1. **A subscript "i" has been added to y in the right side of Equation (1).**
2. **The sentence has been changed to: "First, the large sample size (66 million driving hours and 2.3 billion driving miles) provides the statistical power to estimate the correlation between SCEs and crashes."**
3. **The corresponding sentence has been changed to "In NDSs, safety-critical events (SCEs) are commonly used to measure safety since crashes are very rare."**
4. **We have added labels for the horizontal axes in Figure 4 (previously, Figure 3). In addition, we now spell out "posterior predictive checks."**

Response to Reviewer 3's Comments

This paper investigated the association of the surrogate safety metrics and crashes using the NDS data collected from instrumented trucks. Below comments can further improve the quality of the manuscript.

Please contact the journal in case of clarifying the data source and data reliability step.

First, we thank you for your careful reading of the paper and your suggestions for improvement. We have clarified the data sources and quality of the collected data in the revised Data section. Due to the proprietary nature of the data, we cannot provide the actual data in the supplementary materials. Instead, we provide a hypothetical data set similar to our ping data in our supplementary materials section.

While authors reviewed some NDS studies, the biggest NDS study in the US and Europe i.e. "SHRP2" and UDRIVE are missing from the review. Below papers can provide more info about the latest and the largest scale NDS studies with more than 3200 drivers. Please update the Table 1, accordingly.

- Complementary Methodologies to Identify Weather Conditions in Naturalistic Driving Study Trips: Lessons Learned from the SHRP2 Naturalistic Driving Study & Roadway Information Database
- Eenink, R., Barnard, Y., Baumann, M., Augros, X., & Utesch, F. (2014). UDRIVE: the European naturalistic driving study. In Proceedings of Transport Research Arena. IFSTTAR.
- The impacts of heavy rain on speed and headway behaviors: an investigation using the SHRP2 naturalistic driving study data (TRC).
- The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters

We have read through the provided three references. These papers did use large NDS data sets, but they did not investigate the association between crashes and SCEs like those papers listed in Table 1. We were careful in the revision to indicate that the table describes studies on the relationships between SCEs and crashes. Here are a brief summary of the recommended papers:

1. Ghasemzadeh et al. (2019) used the SHRP2 NDS database with more than 3,500 drivers and described three methods to merge NDS data sets with weather information: wiper status, National Climate Data Center, and weather-related crashes. They also provided data reduction and processing procedures for the SHRP2 NDS data. The paper did not investigate the relationship between crashes and SCEs.
2. Eenink et al. (2014) and Barnard et al. (2016) introduces the overarching goal, methodology, research questions, data to be collected, and expected outcomes of the European naturalistic Driving and Riding for Infrastructure & Vehicle safety and Environment (UDRIVE) project. This paper is more like a research proposal than an evaluation of the data, and the association between crashes and SCEs is a proposed research question but unanswered.
3. Ahmed & Ghasemzadeh (2018) is a data analytics paper that quantifies the association between weather conditions and driver speed and headway selection behaviors using the SHRP2 NDS database. They did not investigate the association between driver speed, headway selection, and real crashes.

Please introduce the data ping. What frequency of data a data ping is representing?

Thank you for your suggestion. We have updated our description of the data ping, including the frequency of the times between pings in the revised manuscript:

Our study is based on data captured from April 1, 2015 to March 31, 2016 by the company's sensor-based monitoring system on their entire fleet. Our dataset includes intermittently collected real-time driving *ping* records, which ranges from every couple of seconds to approximately 15 minutes. Over 50% of the time intervals between two pings were less than 5 minutes and over 95% of them were less than 15 minutes. The time intervals varied among drivers, places, and trips, and there were no clear patterns explaining the variations in interval lengths. Each ping is a data point that includes the exact date and time of the record (year, month, day, hour, minute, and second), GPS (latitudes and longitudes with five decimal place recordings), GPS quality, speed, and drivers' anonymized unique identification code.

We also pointed out that the data set we used for this study was collected for a different purpose, although we argue that it still provides safety information. Only afterward was thought given to using the data to assess road safety. Because of this, we were unable to have any influence on the design of the study regarding things like choosing the ping intervals. We have added this paragraph in Section 2.2.1:

We must emphasize that these data were collected as part of the company's ongoing monitoring and surveillance plan. The data were not collected specifically as part of a planned NDS. That said, the data are still measurements on factors that could affect safety and are therefore still valid for the purpose of answering the questions we pose in Section 1.2.

In addition, a hypothetical sample of ping data (in comma separated values format) is provided in our supplementary materials. The readers are encouraged to check out the ping data if they feel unclear/confused by our description.

- Please explain the collision mitigation surrogate in Table 1.
- The authors should explain the method that they calculated headway. Were the vehicles instrumented with radar?
- More explanation of the hard brakes and the threshold that was used should be added. What was the threshold for the 231101 hard brakes? Clarify whether this number represents the events or data pings.
- Rolling stability should be defined.
- Description of the headway calculation and the threshold for critical headway SCE should be added.
- What are the present thresholds on page 8?

The definitions of SCEs were determined by our partner company as a part of their own routine monitoring program using the Bendix® Wingman® Advanced™ monitoring system. The definitions of the four SCEs are defined in the revised Introduction section.

- Headway, which signals an instance of tailgating for at least 118 seconds at an unsafe gap time (a measure of distance between leading and trailing vehicles) of 2.8 seconds or less (Grove et al. 2015).

- Hard brakes, which are defined as instances of a deceleration rate of 9.5 miles per hour per second or more.
- Activation of the rolling stability system, which intervenes by applying brake pressure (in addition to potentially applying trailer pressure) assisting the driver in aligning the vehicle when the system's critical thresholds are approached (Bendix® 2007).
- Activation of the forward collision mitigation system.

Regarding rolling stability, any NDS study that consider adverse weather and driver performance?

This is a good point, but it is not central to the primary question of our study: the relationships between SCEs and crashes. We have not found any research articles that study the the association between rolling stability and crashes controlling for adverse weather and driver performance. The relationship between rolling stability, as well as the other three SCEs, and other factors such as weather, traffic, driver characteristics, etc., is a topic for future research.

It would be interesting to see challenges with the data, missing values, etc. to be explained in a paragraph as a data preparation stage.

This is an excellent point. With such a large file (1.2 tera-bytes) data management and cleaning was quite a task. We have added a more complete discussion of missing values and data quality control in the first paragraph of Data Description subsection:

The original dataset provided by the company included 34,348 drivers. We have excluded 2,520 drivers (i.e., 7.4% of the original dataset) from our analysis if any of the following criteria is met: (a) driver inactivity, where we required the driver to have at least 100 GPS pings in the data to be included; (b) the unique identification code for the driver is not found in the provided demographics table; and/or (c) the number of SCEs reported were identified as obvious outliers (we only removed drivers who had an unrealistically large number of SCEs). Hereafter, all reported data will correspond to only those generated by the remaining 31,828 drivers, whose characteristics are summarized in Table 2.

Page 9 the authors mentioned the median distance of the trip and the median number of miles per trip as 2.61 and 77.06. Did the authors only considered trucks in the urban environment?

This data set covers the national truck transportation environment in the United States. Figure 2 shows the geographic spread of the pings. Although most of the ping data were in urban areas, rural areas were also covered. In fact, in rural areas, the trucks were probably on interstate highways where the speed limit is higher. With the same distribution of times between pings, the trucks would seem to ping less often in rural areas because they cover more ground in that amount of time. Note that here by "trip" we refer to continuous driving with no stops longer than 30 minutes, which is independent from trucks origin or destination. The actual dispatch and driver's shift in our data typically consist of a number of "trips" pieced together, separated by, for example, rest stops.

Xk should be xik

We have corrected this subscript issue.

Authors need to explain how did they come up with K values.

We did not propose these Pareto k values. They were proposed by (Vehtari et al. 2015, 2017). We have added citations in the Methods and Results section to clarify this.

It is recommended that page 10 paragraph 1 be summarized in a table and provide stat for each category.

A summary table (Table 3) has been added at the bottom of p. 11.

Page 11 talked about table 2 and table two is presented in page 14. Please keep the tables close to the description, if possible.

We have reformatted the manuscript to keep the tables and figures closer to where they are referenced. We used LaTeXto typeset the document, and sometimes it has a mind of its own about where tables and figures should go. By overriding some of the default settings, we were able to place the references closer to the objects.

References

Ahmed, M. M. & Ghasemzadeh, A. (2018), 'The impacts of heavy rain on speed and headway behaviors: an investigation using the shrp2 naturalistic driving study data', *Transportation Research Part C: Emerging Technologies* **91**, 371–384.

Barnard, Y., Utesch, F., van Nes, N., Eenink, R. & Baumann, M. (2016), 'The study design of udrive: the naturalistic driving study across europe for cars, trucks and scooters', *European Transport Research Review* **8**(2), 14.

Bendix® (2007), 'Bendix® ABS-6 Advanced with ESP® Stability System - frequently asked questions to help you make an intelligent investment in stability', Bendix Commercial Vehicle Systems LLC, a member of the Knorr-Bremse Group. https://www.bendix.com/media/documents/products_1/absstability/truckstractors/StabilityFAQ.pdf. [Published March 2007; accessed April 19, 2020].

Dingus, T. A., Hanowski, R. J. & Klauer, S. G. (2011), 'Estimating crash risk', *Ergonomics in Design* **19**(4), 8–12.

Eenink, R., Barnard, Y., Baumann, M., Augros, X. & Utesch, F. (2014), UDRIVE: the European naturalistic driving study, *in* 'Proceedings of Transport Research Arena', TRA 2014, 14-17 Apr 2014, Paris, France. IFSTTAR.
URL: *http://eprints.whiterose.ac.uk/93078/*

Federal Highway Administration (2019), 'Human factors', U.S. Department of Transportation, https://highways.dot.gov/research/research-programs/safety/human-factors. [Updated December 2, 2019; accessed April 29, 2020].

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, Chapman and Hall/CRC.

Ghasemzadeh, A., Hammit, B. E., Ahmed, M. M. & Eldeeb, H. (2019), 'Complementary methodologies to identify weather conditions in naturalistic driving study trips: Lessons learned from the shrp2 naturalistic driving study & roadway information database', *Safety Science* **119**, 21–28.

Gitelman, V., Bekhor, S., Doveh, E., Pesahov, F., Carmel, R. & Morik, S. (2018), 'Exploring relationships between driving events identified by in-vehicle data recorders, infrastructure characteristics and road crashes', *Transportation Research Part C: Emerging Technologies* **91**, 156–175.

Grove, K., Atwood, J., Hill, P., Fitch, G., DiFonzo, A., Marchese, M. & Blanco, M. (2015), 'Commercial motor vehicle driver performance with adaptive cruise control in adverse weather', *Procedia Manufacturing* **3**, 2777–2783.

Guo, F., Klauer, S. G., Hankey, J. M. & Dingus, T. A. (2010), 'Near crashes as crash surrogate for naturalistic driving studies', *Transportation Research Record* **2147**(1), 66–74.

Hu, Q., Cai, M., Mohabbati-Kalejahi, N., Mehdizadeh, A., Yazdi, A., Ali, M., Vinel, A., Rigdon, S. E., Davis, K. C. & Megahed, F. M. (2020), 'A review of data analytic applications in road traffic safety. part 2: Prescriptive modeling', *Sensors* **20**(4), 1096.

Jun, J., Ogle, J. & Guensler, R. (2007), 'Relationships between crash involvement and temporal-spatial driving behavior activity patterns: use of data for vehicles with global positioning systems', *Transportation Research Record* **2019**(1), 246–255.

McElreath, R. (2020), *Statistical rethinking: A Bayesian course with examples in R and Stan*, CRC press.

Mehdizadeh, A., Cai, M., Hu, Q., Yazdi, A., Ali, M., Mohabbati-Kalejahi, N., Vinel, A., Rigdon, S. E., Davis, K. C. & Megahed, F. M. (2020), 'A review of data analytic applications in road traffic safety. part 1: Descriptive and predictive modeling', *Sensors* **20**(4), 1107.

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A. & Elleman, L. G. (2016), *Web and phone based data collection using planned missing designs*, Sage Publications, Inc, pp. 578–595.

Revelle, W., Wilt, J. & Rosenthal, A. (2010), *Individual Differences in Cognition: New Methods for Examining the Personality-Cognition Link*, Springer New York, New York, NY, pp. 27–49.

Saleh, J. H., Saltmarsh, E. A., Favaro, F. M. & Brevault, L. (2013), 'Accident precursors, near misses, and warning signs: critical review and formal definitions within the framework of discrete event systems', *Reliability Engineering & System Safety* **114**, 148–154.

Vatcheva, K. P., Lee, M., McCormick, J. B. & Rahbar, M. H. (2016), 'Multicollinearity in regression analyses conducted in epidemiologic studies', *Epidemiology (Sunnyvale, Calif.)* **6**(2).

Vehtari, A., Gelman, A. & Gabry, J. (2017), 'Practical bayesian model evaluation using leave-one-out cross-validation and waic', *Statistics and Computing* **27**(5), 1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y. & Gabry, J. (2015), 'Pareto smoothed importance sampling'.