Evaluating the Combined Effects of Weather and Real-Time Traffic Conditions on Freeway Crash Risks

CHENGCHENG XU, CHEN WANG, AND PAN LIU

Jiangsu Key Laboratory of Urban ITS, Southeast University, and Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, and School of Transportation, Southeast University, Nanjing, China

(Manuscript received 18 November 2017, in final form 21 June 2018)

ABSTRACT

The study presented in this paper investigated the combined effects of environmental factors and real-time traffic conditions on freeway crash risks. Traffic and weather data were collected from a 35-km freeway segment in the state of California, United States. The weather conditions were classified into five categories: clear, light rain, moderate/heavy rain, haze, and mist/fog. Logistic regression models using unmatched case-control data were developed to link the likelihood of crash occurrences to various traffic and environmental variables. The sample size requirements for case-control studies and the interaction between traffic and environmental variables were considered. The model estimation results showed that the light rain, moderate/heavy rain, and mist/fog significantly increase freeway crash risks. The interaction between light rain and upstream occupancy was also found to be statistically significant. Bootstrap analyses were conducted to quantify the interaction effect between these two variables. The crash risk model was compared to a reduced model in which environmental information was not included. It was found that the inclusion of environmental information improved both goodness of fit and prediction performance of the crash risk prediction model. The inclusion of environmental information in crash risk models improved the prediction accuracy of crash occurrences by 6.8% and reduced the false alarm rate by 1.3%. It was also found that the inclusion of environmental information had minor impacts on the prediction performance of the crash risk model in clear weather conditions.

1. Introduction

It has long been recognized that environmental factors have significant impacts on traffic safety. Driving under adverse weather conditions poses great challenges to drivers due to the reduced visibility, pavement friction, and vehicle stability. In the past few decades, a number of studies have investigated the relationship between traffic safety and adverse weather conditions (Bertness 1980; Andrey and Yagar 1993; Brown and Baass 1997; Brodsky and Hakkert 1988; Knapp et al. 2000; Khattak and Knapp 2001; Eisenberg 2004; Qiu and Nixon 2008). Even though the estimated impacts of environmental factors on crash risks varied greatly, previous studies generally confirmed that the crash frequency increased significantly under adverse weather conditions.

A number of existing studies have demonstrated that rainy conditions significantly affected the occurrences of crashes (Bertness 1980; Brodsky and Hakkert 1988; Andrey and Yagar 1993; Eisenberg 2004; Qiu and Nixon 2008). Brodsky and Hakkert (1988) found that the injury crash rates in rainy conditions were 2–3 times greater than those in dry conditions. Qiu and Nixon (2008) found that rainy weather could increase crash rates by 71% and injury rates by 49%. The risks of crashes on rainy days increased with an increase in the precipitation intensity. Several studies also compared the crash rates and occupant injuries on snowy and dry days. The findings of existing studies showed that crash rates increased significantly in snowy conditions (Khattak and Knapp 2001; Brown and Baass 1997; Knapp et al. 2000). Khattak and Knapp (2001) found that snowy conditions increased noninjury crash rates by 21 times and injury crash rates by 11 times. A study conducted by Knapp et al. (2000) showed that the crash rates in snowy conditions were about 13 times greater than those on dry days. Most of the existing studies have focused on establishing a relationship between adverse weather conditions and crash count data that were aggregated over a relatively long period of time, such as one year.

Corresponding author: Chen Wang, wkobec@hotmail.com

DOI: 10.1175/WCAS-D-17-0124.1

During the past decade, the rise of dynamic freeway traffic control techniques, such as ramp metering systems and variable speed limit systems, provided a proactive way to improve traffic safety on freeway mainlines (El Khoury and Hobeika 2006; Hossain and Muromachi 2011). The prediction of the risks of crash occurrences on freeways is an essential task for a freeway dynamic safety management system as it helps identify if crash prevention strategies are needed to eliminate the hazardous conditions that potentially lead to crashes. In recent years, increased attention has been directed toward the development of crash risk prediction models using freeway real-time traffic surveillance data (Oh et al. 2005; Abdel-Aty et al. 2004; Xu et al. 2012, 2018; Pirdavani et al. 2015; You et al. 2017; Wang and Kim 2018). The main idea of most existing studies was to establish a relationship between freeway crash risks and various factors such as real-time traffic flow variables and freeway geometric characteristics.

Oh et al. (2005) found that the standard deviation of speed in 5-min time intervals was a good indictor to identify hazardous traffic conditions from normal traffic conditions. A probabilistic neural network model was developed to link crash risks with detector occupancy and standard deviation of speed. Wang et al. (2017) proposed a real-time crash risk modeling approach by integrating crash frequency and real-time crash risk. The results suggested that the average speed, standard deviation of speed and occupancy, and truck percentage are the main traffic variables contributing to crash risks. Theofilatos et al. (2018) investigated the impacts of real-time traffic characteristics on crash occurrence by using rare-event logit regression. The results revealed a negative relationship between crash occurrence and speed in crash locations. Wang and Kim (2018) developed a spatial real-time crash risk model using road link speed data. They divided the city of Brisbane, Australia, into several hundreds of spatial units. The crash risks at each spatial unit were related with the normalized speeds on road links.

In addition to modeling the likelihood of crash occurrence, a number of studies have also investigated the effects of traffic flow conditions on collision type and crash severity (Xie et al. 2017; Xu et al. 2018; Dimitriou et al. 2018). Xie et al. (2017) used the hierarchical Bayesian model to evaluate the impacts of real-time traffic conditions on the likelihood of hit-and-run crashes. Vehicle speed, roadway segment length, and weekend days were found to be positively correlated with the hit-and-run crash risks. Dimitriou et al. (2018) developed a multinomial logit model to assess rear-end crash potential based on vehicle-by-vehicle interactions, geometric characteristics, and operational conditions. The results showed that speed-related traffic variables and vehicle headway are the main traffic variables contributing to the likelihood of rear-end crash potential.

The loop detector data were the most commonly used traffic flow data for real-time crash risk modeling. In recent studies, traffic flow data collected from other traffic surveillance equipment have also been used for developing real-time crash risk models (Wang et al. 2015; Basso et al. 2018; Yuan et al. 2018). For example, Wang et al. (2015) developed a multilevel Bayesian logistic regression model for real-time crash risk assessment at expressway weaving segments. The traffic flow data collected from the microwave vehicle detection system were used for model development. Basso et al. (2018) developed a real-time crash risk model using traffic flow data measured by the Automatic Vehicle Identification system. Yuan et al. (2018) used conditional logistic regression to link the crash likelihood with real-time traffic conditions and adaptive signal phasing on arterials. The traffic data collected from Bluetooth were used. The results showed that the average speed and upstream volume are the main traffic flow variables affecting crash risks on arterials.

Relatively fewer studies have considered the impacts of environmental conditions in real-time crash risk prediction models. Golob and Recker (2004) and Golob et al. (2004) classified freeway traffic flow into different states by using clustering analyses. Nonlinear canonical correlation analyses were further conducted to identify the relationship between the types of crashes and the traffic states in different weather conditions. A more recent study by Yuan et al. (2018) linked crash risks on arterials with traffic flow conditions and weather characteristics. An indicator variable of rainy weather was used in the model. The rainy weather was found to be significantly correlated with crash risks on arterials. The results of these studies indicated that weather is an important factor affecting crash risks.

However, previous studies generally did not consider the interaction between traffic flow and weather conditions. The impacts of traffic conditions on crash risks were assumed to be the same across different weather. In fact, the effects of traffic flow conditions on crash risks might be distinct between different weather conditions due to the reduced pavement friction and visibility. Moreover, previous studies generally used an indicator variable such as rainy weather or not to describe weather conditions. Such indicator-based weather variables overlooked the impacts of visibility or precipitation intensity on crash risks. This study fills gaps in understanding the interactive relationship between traffic flow and weather conditions. The weather condition classification in meteorology based on visibility and precipitation intensity was adopted in this study, with the purpose of better modeling the effects of weather conditions on crash risks.

Crashes are random events, the occurrences of which are affected by numerous factors such as human behaviors, traffic flow characteristics, and environmental factors. Existing studies have confirmed that adverse weather conditions significantly affect traffic safety on freeways. Theoretically, the incorporation of weather information may improve the predictive performance of crash risk prediction models. This study aimed to investigate the combined effects of weather factors and real-time traffic conditions on freeway crash risks. This study has the potential to contribute to the field of real-time crash risk assessment by 1) better modeling the effects of weather conditions on crash risks based on the weather condition classification used in meteorology, 2) investigating the interactions between traffic flow characteristics and weather conditions, and 3) incorporating weather information in crash risk models considering the interactions between traffic flow and weather conditions.

2. Design of the case-control study

a. Logistic regression analysis using case-control data

Logistic regression analysis using case-control data has been widely used in epidemiology to identify the impacts of risk factors on a disease. Recently, the method has been extended to predicting the crash risks on freeways given real-time traffic surveillance data (Abdel-Aty et al. 2004; Xu et al. 2012). Two general approaches are available for conducting a case-control study, including the matched case-control study design and the unmatched case-control design. The major difference between the matched and unmatched case-control design is that for the matched case-control design, controls (noncrash cases) were matched with cases (crash cases) according to some confounding factors such as the time and locations of crashes, while for the unmatched case-control design the control samples were randomly selected.

A matched case-control study design requires a matched analysis, such as the conditional logistic regression model, to account for the selection bias introduced into the study (Rothman and Greenland 1998). Unlike ordinary logistic regression models, the intercepts of a conditional logistic regression model are not constants because the conditional likelihood function is independent of the intercepts. Thus, the conditional logistic regression model cannot be directly used to predict the probability of crash occurrence. To address this concern, previous researchers have suggested using the log odds ratios to approximate the relative risk of a crash as compared to normal traffic conditions. The normal traffic conditions were defined as the mean values of the traffic flow parameters associated with noncrash cases in each matched pair, which sometimes may be difficult to be estimate in practical engineering applications.

The purpose of using the matched case-control study design was to account for the impacts of confounding factors. More recent studies in epidemiology suggested that both matched and unmatched case-control study design can control for the impacts of confounding variables (Rothman and Greenland 1998). The unmatched case-control analysis was used in this study to develop a real-time crash risk model. Traffic and weather data before crash occurrences were taken as cases while the randomly selected traffic and weather data on crash-free days were taken as noncrash cases, that is, the controls. A control-to-case ratio of 4:1 was selected in this study because previous studies in epidemiology have suggested that there was only a small increase in the statistical power beyond a control-to-case ratio of 4:1 (Rothman and Greenland 1998). With the unmatched case-control data, binary logistic regression models were developed to predict the probability crash occurrence on freeways.

The binary logistic regression model has been used in previous studies for predicting a binary-dependent variable as a function of predictor variables (Xu and Tian 2008; Hunter et al. 2011; Xu et al. 2017b). Using the binary logistic regression model, the probability of the occurrence of a crash can be estimated using the following equation:

$$P(x_i) = \frac{1}{1 + e^{-g(x_i)}} \quad (i = 1, 2, \ldots, n), \tag{1}$$

where $P(x_i)$ denotes the probability of the occurrence of a crash and $g(x)$ is the multiple linear combination of explanatory variables, which can be expressed as

$$g(x) = \ln\frac{P(x_i)}{1 - P(x_i)} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, \tag{2}$$

where x_{ki} denotes the value of variable k for sample i and β_k is the coefficient of variable k. The parameters β_0, β_1, $\beta_2, \ldots, \beta_k$ can be estimated by solving the log-likelihood function (Xu and Tian 2008; Hunter et al. 2011; Xu et al. 2017a), which is given by

$$\ln L(\beta, x_i) = \sum_{i=1}^{n} [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{ki} x_{ki} - \ln(1 + e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{ki} x_{ki}})]. \tag{3}$$

The binary logistic regression model classifies an observation as an event (a crash) if the predicted probability of the observation is greater than or equal to a given probability threshold. Otherwise, it will be classified as a nonevent, or in other words, a noncrash case. The probability threshold for distinguishing an event and a nonevent is usually set to 0.5 for a balanced sample in which the number of events is equal to the number of nonevents. However, the value of 0.5 may produce biased prediction performance when dealing with an unbalanced sample, in which the number of observations in one class is greater than that in the other class. For an unbalanced sample, the probability threshold is usually set to the percentage of the actual events in the dataset (Jung et al. 2010), that is, the sample size of crash cases divided by the total sample size of crash and noncrash samples. Since the control-to-case ratio is 4:1 in this study, a probability threshold of 0.2 was adopted for identifying a crash case in the following analyses.

b. *Sample size requirements for case-control studies*

Before a crash risk prediction model is developed, it is sometimes important to identify if the sample size of the case-control data is adequate to ensure that a reasonable level of statistical power can be achieved for detecting an effect at a significance level. This is particularly true for the present research because the sample size of the crash and noncrash cases in adverse weather conditions is very likely to be much smaller than those in clear weather conditions. Statistical power is the probability of detecting the effect of a variable as statistically significant, given the fact that the effect of the variable is real. In case-control studies, the minimum power is often set to 80% (Macfarlane 2003; Freidlin et al. 2002). The sample size required to detect the effect of a continuous variable with a specific level of statistical power can be estimated using the following equation (Kelsey et al. 1996):

$$N_i = \frac{(Z_{\alpha/2} + Z_\beta)^2 \sigma^2 (r+1)}{(d)^2 r}, \tag{4}$$

where N_i represents the required sample size in the case group with respect to variable x_i; Z is the statistic for the standard normal distribution; α represents the level of significance; β represents the probability of failing to detect the effect of a variable as statistically significant, given that the effect of the variable is real; $(1 - \beta)$ represents the desired statistical power; r represents the control-to-case ratio; d represents the difference in means between the case and control group; and σ represents the standard deviation of the continuous independent variable.

The required sample size for an indicator variable can be estimated using the following equation (Kelsey et al. 1996):

$$N_i = \frac{(Z_{\alpha/2} + Z_\beta)^2 \overline{p}(1-\overline{p})(r+1)}{(p_1 - p_0)^2 r}, \tag{5}$$

where p_1 represents the proportion of cases; p_0 represents the proportion of controls; and $\overline{p}$ represents the weighted average of p_1 and p_0, which equals $(p_1 + rp_0)/(1 + r)$.

c. *Interactions in logistic regression models*

The impacts of the interaction between traffic flow and environmental variables must be considered if environmental information is to be included in a crash risk prediction model. Interaction refers to the situation in which the effects of two independent variables on crash risks are not independent. As a result, the effects of one independent variable on crash risks may vary by the other independent variable. If the interactions between traffic flow and environmental variables are not properly considered, the model may provide biased results on the effects of independent variables on crash risks.

The interaction effect between two independent variables can be quantified on either a multiplicative scale or an additive scale. The interaction on a multiplicative scale measures the difference between the interaction effect of two independent variables and the product of the individual effect of each independent variable. The interaction on a multiplicative scale between two independent variables x_1 and x_2 can be estimated by including a product term in the logistic regression model. With the inclusion of the product term, the logit of $P(x_i)$ is given as

$$\ln\left[\frac{P(x_i)}{1 - P(x_i)}\right] = \ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 . \tag{6}$$

The individual effect of x_1 on crash risks (OR_{x_1}) can be estimated as

$$\text{OR}_{x_1} = \frac{\text{odds}(x_1 = 1, x_2 = 0)}{\text{odds}(x_1 = 0, x_2 = 0)} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} . \tag{7}$$

The individual effect of x_2 on crash risks (OR_{x_2}) can be estimated as

$$\text{OR}_{x_2} = \frac{\text{odds}(x_1 = 0, x_2 = 1)}{\text{odds}(x_1 = 0, x_2 = 0)} = \frac{e^{\beta_0 + \beta_2}}{e^{\beta_0}} = e^{\beta_2} . \tag{8}$$

The interaction effect of x_1 and x_2 on crash risks (OR_{x_1,x_2}) can be estimated as

$$\mathrm{OR}_{x_1,x_2} = \frac{\mathrm{odds}(x_1 = 1, x_2 = 1)}{\mathrm{odds}(x_1 = 0, x_2 = 0)} = \frac{e^{\beta_0+\beta_1+\beta_2+\beta_3}}{e^{\beta_0}} = e^{\beta_1+\beta_2+\beta_3} = \mathrm{OR}_{x_1} \times \mathrm{OR}_{x_2} \times e^{\beta_3} . \quad (9)$$

Thus, the coefficient of the product term represents the interaction effect on a multiplicative scale. The interaction on an additive scale measures the difference between the combined effect of two independent variables and the sum of their individual effects. The interaction on an additive scale between x_1 and x_2 in a binary logistic regression model can be estimated by the relative excess risk due to interaction (RERI), which can be calculated using the following equation (Hosmer and Lemeshow 1992):

$$\begin{aligned}\mathrm{RERI} &= (\mathrm{OR}_{x_1,x_2} - 1) - (\mathrm{OR}_{x_1} - 1) - (\mathrm{OR}_{x_2} - 1)\\ &= \mathrm{OR}_{x_1,x_2} - \mathrm{OR}_{x_1} - \mathrm{OR}_{x_2} + 1\\ &= e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1 . \quad (10)\end{aligned}$$

3. Data sources

Data were obtained from a 35-km segment of the northbound lanes of Interstate 880 (I-880 N), a freeway in the San Francisco Bay area of the United States. As shown in Fig. 1, a total of 42 loop detector stations and three weather stations were located along the selected freeway segment. The average spacing between loop detector stations is about 0.8 km, and the average spacing between weather stations is about 11 km. The standard deviation of the spacing between loop detector stations is around 0.5 km. The mean of the distance between upstream detector station and crash location is about 0.5 km. The lower and upper quartiles of this distance are 0.16 and 0.8 km. Crash data, traffic data, and weather data were collected from 1 January to 31 December 2008 and from 1 January to 31 December 2010. A total of 887 crashes were identified and used for further data analyses.

The traffic data were collected from the Highway Performance Measurement System (PeMS), which is maintained by the California Department of Transportation (Caltrans). The 30-s raw loop detector data for each lane were collected from the Caltrans PeMS database. Invalid traffic data were excluded from further data analyses. Note that the traffic data were considered invalid when 1) the average speed was greater than 160 km h^{-1}, 2) the average occupancy was greater than 100%, 3) the flow rate was greater than 0 vehicles per hour (vph) while the occupancy equaled 0%, 4) the average speed was greater than 0 km h^{-1} while the flow rate equaled 0 vph, or 5) the occupancy was greater than 0% while the flow rate was equal to 0 vph. The 30-s raw data were further aggregated into 5-min time intervals to obtain the averages and standard deviations.

For each crash, the researchers extracted traffic data in the time interval between 10 and 15 min prior to crash occurrence. It was suggested by previous studies that using data from at least 5 min before the time of crash occurrences would help to identify a high likelihood of crashes ahead of time so as to make preemptive measures possible (Pande et al. 2011; Xu et al. 2018). Traffic data were extracted for each crash from the two nearest loop detector stations, including an upstream and a downstream loop detector station, as shown in Fig. 1. Weather data were extracted from the National Climatic Data Center (NCDC) website, which provided hourly weather information obtained from weather stations across the United States. Weather data for each crash were extracted based on the time of the crash from the weather station nearest to its location. The environmental information collected for each crash included the visibility and precipitation intensity. Note that the weather stations used in this study for extracting weather data were all located within 1.6 km from the I-880 N freeway. The mean of the distance between weather station and crash location is about 4.3 km. The lower and upper quartiles of this distance are 2.7 and 6.9 km.

Weather conditions were defined based on visibility or precipitation intensity according to the definitions used in meteorology. The low-visibility weather was classified as haze, mist, and fog based on the magnitude of visibility, and the intensity of rainfall was classified as light, moderate, and heavy based on the precipitation intensity (Vautard et al. 2009; Glickman 2000). Considering the sample size in each category, the moderate and heavy rain conditions and the mist and fog conditions were combined. As a result, the study considered five different weather conditions: clear, light rain, moderate/heavy rain, haze, and mist/fog. The definitions for different weather conditions are given in Table 1.

The study presented in this paper was based on an unmatched case-control data structure where the cases were traffic and weather data before crash occurrences, while the noncrash cases were traffic and weather data that were randomly selected in crash-free conditions. For each crash case, four observations of noncrash cases were randomly selected from crash-free days. To be more specific, the location and time of each noncrash case were first generated randomly. The upstream and downstream detector stations of each noncrash case were randomly selected from the 42 loop detector stations. The time of each noncrash case was randomly selected from 210 528 time periods, which is the total number of 5-min intervals in 2008 and 2010. Accordingly, the

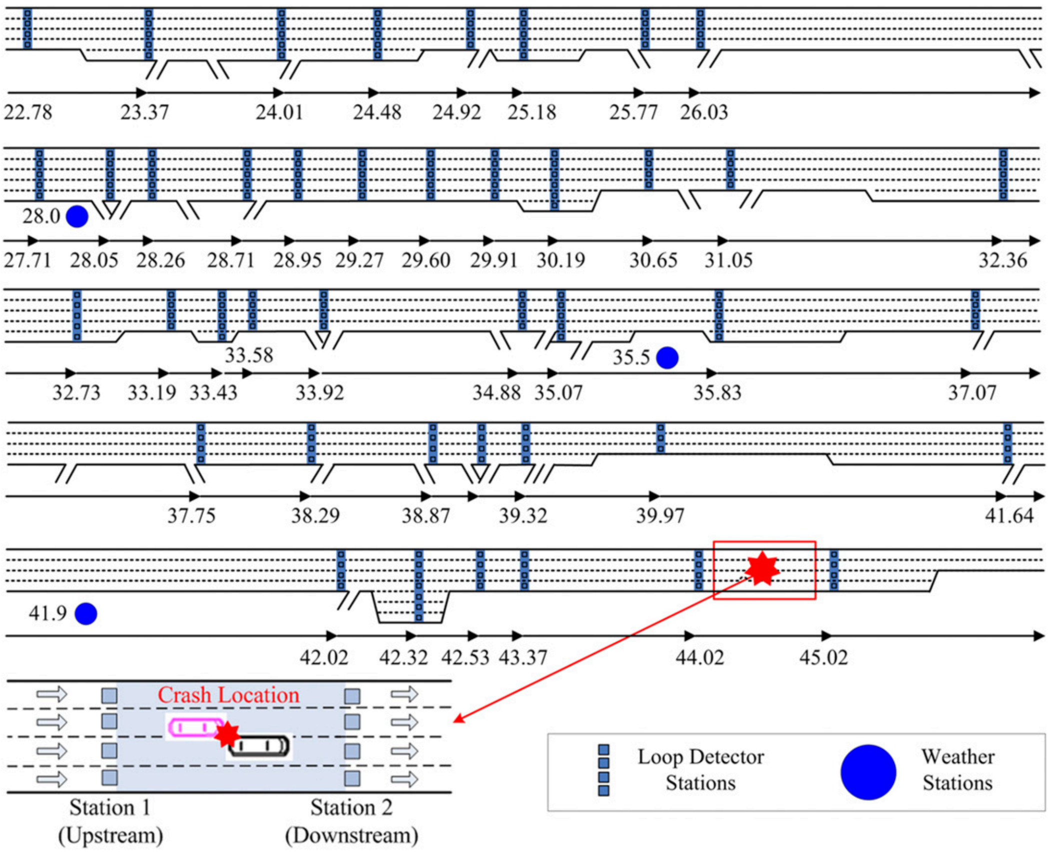

FIG. 1. Location of the loop detector and weather stations along the selected segment on the I-880 N freeway.

sampling universe contains 8.631×10^6 (= 210 528 × 41) potential time–space points. The observations of noncrash cases were randomly selected from the sampling universe. Moreover, it was ensured that no crashes occurred at the location of each noncrash case during the whole day. The traffic flow and weather data for each noncrash case were then extracted based on the randomly generated time and loop detector stations.

The distributions of crash and noncrash cases under different weather conditions are summarized in Table 1. A total of 887 crash cases and 3548 noncrash cases were included in our database. The original dataset was randomly separated into a training set and a testing set with a ratio of 4:1. The training dataset was used to develop a real-time crash risk prediction model, while the validation dataset was used to test the prediction

TABLE 1. Distributions of crash and noncrash cases under different weather conditions.

	Crash		Noncrash	
Weather conditions	Frequency	Percentage	Frequency	Percentage
Light rain[a]	84	9.5%	107	3.0%
Moderate[b]/heavy rain[c]	25	2.8%	36	1.0%
Haze[d]	26	2.9%	75	2.1%
Mist[e]/fog[f]	31	3.5%	65	1.8%
Clear	721	81.3%	3265	92.0%
Total	887	100.0%	3548	100.0%

[a] Light rain: precipitation intensity is less than 2.5 mm h^{-1}.
[b] Moderate rain: precipitation intensity is between 2.5 and 7.6 mm h^{-1}.
[c] Heavy rain: precipitation intensity is greater than 7.6 mm h^{-1}.
[d] Haze: visibility is between 2 and 5 km.
[e] Mist: visibility is between 1 and 2 km.
[f] Fog: visibility is less than 1 km.

TABLE 2. Candidate variables used in logistic regression models.

Symbol	Variables
$\mathrm{Occ_{up}}$	Average occupancy at the upstream station (%)
$\mathrm{Speed_{up}}$	Average speed at upstream station (km h^{-1})
Std dev of $\mathrm{Occ_{up}}$	Std dev of occupancy at upstream station (%)
Std dev of $\mathrm{Speed_{up}}$	Std dev of speed at upstream station (km h^{-1})
$\mathrm{Occ_{down}}$	Average occupancy at downstream station (%)
$\mathrm{Speed_{down}}$	Average speed at downstream station (km h^{-1})
Std dev of $\mathrm{Occ_{down}}$	Std dev of occupancy at downstream station (%)
Std dev of $\mathrm{Speed_{down}}$	Std dev of speed at downstream station (km h^{-1})
$\mathrm{Occ_{ud}}$	Average speed difference between upstream and downstream station (%)
$\mathrm{Speed_{ud}}$	Average occupancy difference between upstream and downstream station (km h^{-1})
Std dev of $\mathrm{Occ_{ud}}$	Standard deviation of speed difference between upstream and downstream station (%)
Std dev of $\mathrm{Speed_{ud}}$	Standard deviation of occupancy difference between upstream and downstream station (km h^{-1})
Light rain	1 = Light rain; 0 = others
Moderate/heavy rain	1 = Moderate/heavy rain; 0 = others
Haze	1 = Haze; 0 = others
Mist/fog	1 = Mist/fog; 0 = others
Length	Distance between upstream and downstream loop detector stations (km)
Lane	Number of lanes at the upstream stations
$\mathrm{Ramp_{on}}$	1 = if there is an on-ramp between upstream and downstream stations; 0 = otherwise
$\mathrm{Ramp_{off}}$	1 = if there is an off-ramp between upstream and downstream stations; 0 = otherwise
Curve	1 = curved section; 0 = otherwise

performance of the model. The training dataset included 710 crash cases and 2840 noncrash cases, while the validation dataset includes 177 crash cases and 708 noncrash cases.

As shown in Table 2, 21 candidate variables were considered for developing real-time crash risk prediction models. Equations (4) and (5) were then used to identify if the sample size of the training dataset was large enough to detect the impacts of candidate variables on crash risks. The statistical power and the level of significance were set to 80% and 0.05, respectively. It was found that the sample size of the training dataset satisfied the sample size requirements for most of the candidate variables, except for haze. As mentioned before, haze is a binary variable that represents the weather condition in which the visibility is between 2 and 5 km. As shown in Table 1, the difference in the proportions of the observations of haze between the crash and noncrash cases is very small (0.8%), indicating that the impacts of haze on crash risks are negligible. To detect such a small difference, a large sample size of 16 915 is needed for haze, which is not cost effective or feasible.

4. Results of data analysis

a. Model specification

Binary logistic regression models were fitted to the unmatched case-control data to identify how traffic flow characteristics and weather conditions affected the crash risks on freeways. For the purpose of comparison, we also developed a reduced model in which the weather information was not included. To account for the possible correlations between candidate variables, the Pearson correlation parameters were calculated between different pairs of candidate-independent variables and generated several combinations that included the maximum number of uncorrelated variables. Stepwise variable selection was then conducted to select the independent variables that should be included in crash risk models. The log-likelihood at the convergence of each model was compared. The model with the highest log-likelihood was considered the best model.

The crash risk prediction models were developed using the training dataset. The models were specified using the Statistical Analysis Software package SAS 9.2. Regression results of the full model and the reduced model were given in Tables 3 and 4, respectively. A likelihood ratio test was conducted to identify if the inclusion of weather information significantly improved the goodness-of-fit of the crash risk prediction model. The test statistic is given by (Washington et al. 2003):

$$\chi^2 = -2[\mathrm{LL}(\beta_T) - \mathrm{LL}(\beta_k)]\,, \tag{11}$$

where $\mathrm{LL}(\beta_T)$ is the log-likelihood of the reduced model in which the weather information was not included and $\mathrm{LL}(\beta_k)$ is the log-likelihood of the full model in which both traffic flow and weather variables were included. The test statistic is χ^2 distributed with the degree of freedom being equal to the number of parameters in the

TABLE 3. Crash risk model with both traffic and weather variables.

Variables	Coefficient	Std error	χ^2	Pr > χ^2
Intercept	−1.085	0.243	19.876	<0.0001
$\mathrm{Occ}_{\mathrm{up}}$	0.022	0.006	11.531	0.001
$\mathrm{Speed}_{\mathrm{down}}$	−0.014	0.002	56.852	<0.0001
Std dev of $\mathrm{Speed}_{\mathrm{down}}$	0.021	0.008	6.553	0.011
Light rain	2.309	0.301	58.935	<0.0001
Moderate/heavy rain	1.056	0.312	11.487	0.001
Mist/fog	0.742	0.240	9.538	0.002
$\mathrm{Occ}_{\mathrm{up}}$ × light rain	−0.117	0.023	26.604	<0.0001
Length	0.786	0.196	16.103	<0.0001

Summary statistics

Number of observations = 3550

$-2L(c) = 3552.857$; $-2L(\beta) = 3341.240$

$-2[L(c) - L(\beta)] = 211.617$ (8 df); $P < 0.0001$

TABLE 4. Crash risk model with traffic variables only.

Variables	Coefficient	Std error	χ^2	Pr > χ^2
Intercept	−0.891	0.240	13.847	0.0002
$\mathrm{Occ}_{\mathrm{up}}$	0.015	0.006	5.548	0.0185
$\mathrm{Speed}_{\mathrm{down}}$	−0.014	0.002	54.848	<0.0001
Std dev of $\mathrm{Speed}_{\mathrm{down}}$	0.020	0.008	6.039	0.0140
Length	0.714	0.193	13.666	0.0002

Summary statistics

Number of observations = 3550

$-2L(c) = 3552.857;-2L(\beta) = 3416.045$

$-2[L(c) - L(\beta)] = 136.813$ (4 df); $P < 0.0001$

full model minus the number of parameters in the reduced model. It can be concluded that the full model produced better goodness of fit if the p value of test statistic χ^2 is lower than the significance level of 0.05.

As shown in Tables 3 and 4, the log-likelihoods of the full and reduced models are −1670.62 and −1708.023, respectively. The test statistic χ^2 can be calculated as $-2 \times (-1708.023 + 1670.62) = 74.8$. This value is much greater than the critical value of 23.51 for χ^2 distribution with 4 degrees of freedom at the significance level of 0.0001, indicating that the full model produces significantly better fitness than does the reduced model. Accordingly, including weather information significantly improved the goodness of fit of the crash risk model.

The best model has seven independent variables, including three traffic flow variables, three indicator variables for weather conditions, a product term, and the spacing between upstream and downstream stations. The research team tested the interactions between different traffic flow and weather variables when developing the crash risk prediction model. Only the interaction between the light rain and upstream occupancy was found to be statistically significant. A possible explanation is that during other adverse weather conditions considered in this study, the speeds that drivers choose on freeways are not sensitive to traffic conditions. Instead, drivers' behavior in those adverse weather conditions such as mist/fog and moderate/heavy rain is largely determined by the factors such as the precipitation intensity and visibility.

As shown in Table 3, the upstream occupancy, downstream speed, and downstream speed variance significantly affected crash risks. Both the upstream occupancy and downstream speed variance have positive coefficients, indicating that the crash risk increases as the upstream traffic density and the downstream speed variance increase. The negative coefficient of the downstream speed indicates that the crash risk increases as the downstream speed decreases. These results are consistent with the findings of previous studies (Abdel-Aty et al. 2004; Lee et al. 2003). The spacing between the upstream and downstream stations was the only geometric parameter that was significant in the crash risk model. The positive coefficient of this variable suggests that the crash risk increases as the length between the upstream and downstream stations increases. The segment length can be considered as an exposure to crash risks. When meeting hazardous traffic flow conditions with high crash potential, the vehicles on the longer segments are more exposed to such hazardous traffic flow conditions than those on the shorter segments. Accordingly, for the same traffic conditions, the crash risks on the longer segments are higher than the crash risks on the shorter segments. The other geometric variables were not found to be statistically significant. A possible explanation is that the traffic flow variables included in the model partly reflected the impacts of geometric characteristics on real-time crash risks, because the real-time traffic flow data such as speed variance was heavily affected by the geometric design of the selected freeway sections.

Of the four indicator variables for weather conditions, light rain, moderate/heavy rain, and mist/fog were found to be statistically significant. The coefficients for these three variables are all positive, implying that these adverse weather conditions significantly increase the likelihood of crash occurrences on freeways. As expected, the impacts of haze on freeway crash risks were not found to be statistically significant.

The production term between upstream occupancy and light rain indicates the complex interactive relationship between weather and traffic flow conditions. On the one hand, the light rain affects safety by reducing visibility and increasing vehicle stopping distance. The coefficient of the product term between the upstream occupancy and light rain is negative, indicating that the impacts of light rain on crash risks increase as the upstream occupancy decreases. The finding is intuitive because traffic in free-flow states tends to have higher

TABLE 5. The odds ratios estimates for different variables.

Variables	Point estimate	95% confidence limits	
		Lower	Upper
$\text{Speed}_{\text{down}}$	0.986	0.982	0.990
Std dev of $\text{Speed}_{\text{down}}$	1.021	1.005	1.038
Moderate/heavy rain	2.875	1.561	5.296
Mist/fog	2.100	1.311	3.362
Length	2.195	1.495	3.224

speeds, which increases the crash risks under light rain conditions in which the pavement friction and visibility are reduced.

One the other hand, light rain can also affect safety by reducing travel speed. The negative coefficient of the product term between occupancy and light rain also indicates that the light rain decreases the impact of the occupancy on crash risks. Drivers tend to drive more cautiously by reducing vehicle speed in light rain. Vehicles generally have lower speed in light rain than in clear weather under the same occupancy conditions. Accordingly, the impact of occupancy on crash risks in light rain is lower than that in clear weather.

Odds ratio analyses were further conducted to quantitatively evaluate the impacts of each variable on crash risks. The odds ratios associated with different independent variables are given in Table 5. The odds ratio represents the increase in the odds of the outcome with one unit increase in the independent variable. The results of the odds ratio analyses showed that the risks of crashes would increase by 1.875 times during moderate/heavy rains and by 1.1 times in mist/fog as compared to the clear weather condition, given that other traffic flow variables remained the same. The odds ratio for the downstream speed is 0.986, implying that the crash risk will decrease 1.4% with each unit increase in the downstream speed. The odds ratio for the downstream speed variance is 1.021, indicating that one unit increase in downstream speed variance increases the crash risk by 2.1%.

As shown in Table 3, the coefficient of light rain is greater than that of moderate/heavy rain. However, it does not necessarily mean that light rain has a greater effect on crash risks than moderate/heavy rain. Considering the interaction between the upstream occupancy and light rain, the odds ratios for each of these two variables can be estimated only when the value of one of the variables is given. For example, when the upstream occupancy is equal to 15%, the odds ratios associated with light rain can be estimated as follows:

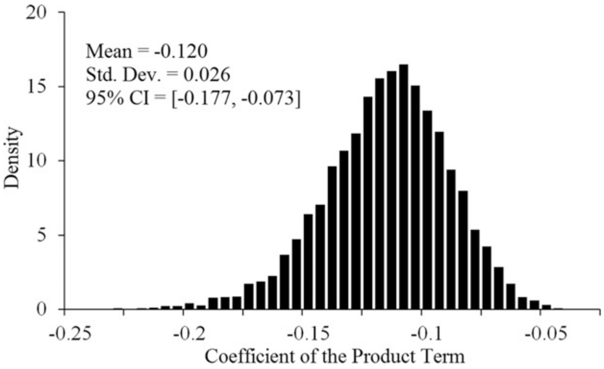

FIG. 2. Bootstrap sampling distribution of the coefficient of the product term.

$$\text{OR} = \frac{\text{odds}(\text{Light rain} = 1,\ \text{Occ}_{\text{up}} = 15\%)}{\text{odds}(\text{Light rain} = 0,\ \text{Occ}_{\text{up}} = 15\%)} = e^{2.309-15\times0.117} = 1.740\,.$$

Bootstrap analyses were further conducted to quantify the interaction effect between the upstream occupancy and the light rain. The purpose of bootstrap analyses lies in two aspects. First, the significance of interaction effects such as the RERI cannot be directly estimated by the logistic regression. The bootstrap analysis was used to investigate whether there is a significant interaction effect between these two variables. It quantifies the difference between the combined effect and the individual effect of these two interacted variables. Moreover, bootstrap analysis is less expensive and more efficient than collecting a large sample of crashes that occurred in adverse weather. Both crashes and adverse weather are rare events. The crashes occurring in adverse weather are extremely rare events.

As mentioned above, the RERI and the coefficient of the product term were used to estimate the interaction effect on an additive scale and on a multiplicative scale, respectively. From the original training dataset, 10 000 bootstrap samples were created by randomly drawing observations with replacement. Each bootstrap sample had the equal sample size with the original training dataset. Crash risk prediction models were developed for different bootstrap samples. The RERI was then calculated for each of the crash risks model using Eq. (10). The bootstrap sampling distributions of the coefficient of the product term and the RERI for the 10 000 crash risk models were illustrated in Figs. 2 and 3. The research team calculated the mean and the 95% confidence interval for the RERI and the coefficient of the product term. The 95% confidence intervals were estimated as the 2.5nd and the 97.5th percentiles of the bootstrap sampling distribution.

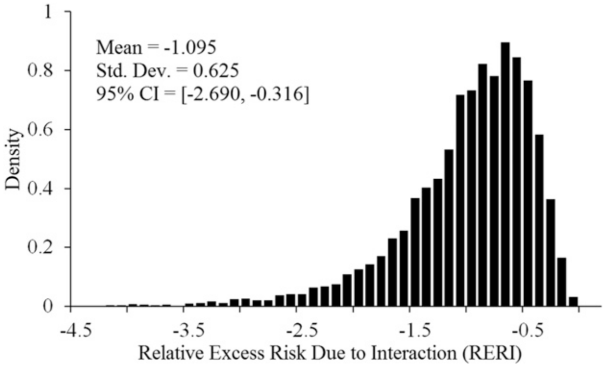

FIG. 3. Bootstrap sampling distribution of the RERI.

As mentioned before, the coefficient of the product term represents the interaction effect on a multiplicative scale. The expected value of the coefficient of the product term is −0.120, indicating the fact that the combined effect of the upstream occupancy and the light rain on crash risks is 1 − exp(−0.120) = 0.113 times less than the product of their individual effects. As shown in Fig. 3, the expected value of the RERI is given by −1.095, indicating the fact that the combined effect of the upstream occupancy and the light rain on crash risks is 1.095 times less than the sum of their individual effects.

b. *Prediction performance*

The prediction performance of the full model with both traffic and weather variables was tested using both the training and validation dataset. The results are given in Table 6. For the training dataset, 61.9% (440 of 710) of crash cases and 67.1% (1906 of 2840) of noncrash cases were correctly identified by the full model, indicating that the crash risk prediction model with weather variables has reasonable goodness of fit to the training data. For the validation dataset that was not used for model specification, 62.7% (111 of 177) crash cases and 66.0% (467 of 708) noncrash cases were correctly predicted. An overall prediction accuracy of 65.3% was achieved for both crash and noncrash cases in the validation dataset, indicating a reasonable prediction capability of the crash risk prediction model.

For comparison purposes, we also tested the prediction accuracy of the reduced model in which weather information was not included. For the training dataset, 56.5% crash cases and 65.2% noncrash cases were correctly identified by the reduced model. For the validation dataset, 55.9% crash cases and 64.7% noncrash cases were correctly predicted by the reduced model. From Table 6 it is clear that the full model provides better prediction performance than the reduced model. The incorporation of weather information in crash risk models improved the prediction accuracy of crash occurrences by 6.8% and reduced the false alarm rate by 1.3%. The McNemar test was further applied to examine whether the prediction performance of the full model is significantly higher than that of the reduced model. The McNemar test is a nonparametric test of which the null hypothesis is that two related dichotomous variables have the same mean (Etemadi et al. 2009). The test was often used for comparing the prediction accuracy of two prediction models (Lensberg et al. 2006; Etemadi et al. 2009). In this study, the McNemar tests were conducted using both training and validation datasets. The results showed that the prediction performance of two models was significantly different at 0.001 significant level ($\chi^2 = 18.356$, $P < 0.0001$). Therefore, the prediction accuracy of the full model is significantly better than that of the reduced model.

The prediction accuracy of the crash risk models is affected by the predetermined probability threshold. The research team further developed receiver operating characteristic (ROC) curves to compare the prediction performance of the full and reduced models with different probability thresholds. The ROC curves for the full and reduced models on the training and validation datasets were illustrated in Figs. 4 and 5. The ROC curves illustrate the relationship between the true

TABLE 6. Comparison of the prediction performance of the full and reduced models.

			Predicted group membership		
Sample	Model	Actual group membership	Crash	Noncrash	Total
Training dataset	Full	Crash	440 (61.9%)	270 (38.1%)	710 (100%)
	Model	Noncrash	934 (32.9%)	1906 (67.1%)	2840 (100%)
	Reduced	Crash	401 (56.5%)	309 (43.5%)	710 (100%)
	Model	Noncrash	988 (34.8%)	1852 (65.2%)	2840 (100%)
Validation sample	Full	Crash	111 (62.7%)	66 (37.3%)	177 (100%)
	Model	Noncrash	241 (34.0%)	467 (66.0%)	708 (100%)
	Reduced	Crash	99 (55.9%)	78 (44.1%)	177 (100%)
	Model	Noncrash	250 (35.3%)	458 (64.7%)	708 (100%)

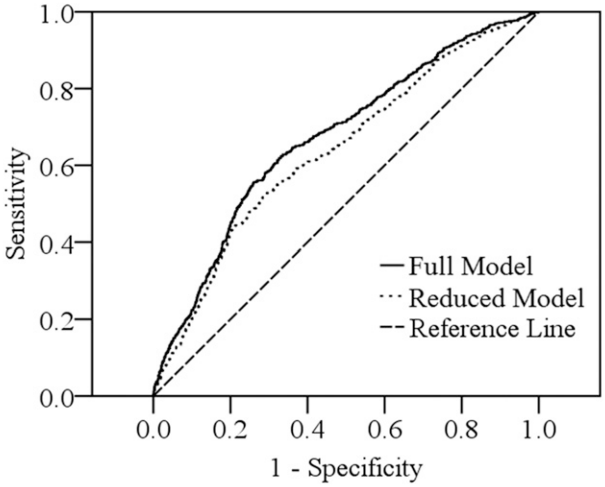

FIG. 4. ROC curves for the full and reduced models for training dataset.

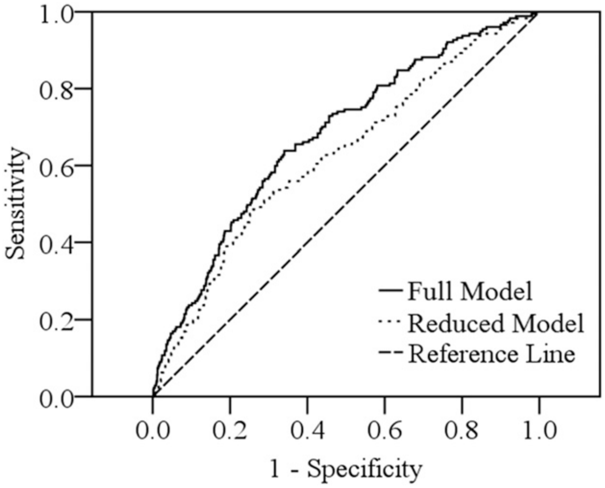

FIG. 5. ROC curves for the full and reduced models for validation dataset.

positive rate (sensitivity) and the false alarm rate (1 − specificity) for a given threshold from 0 to 1. For both training and validation dataset, the ROC curves for the full model are always to the left of the curves for the reduced model, indicating the fact that the prediction accuracy of the full model is always greater than that of the reduced model no matter what probability threshold is selected.

The threshold used in Table 6 was selected to achieve a balance between prediction accuracy of crash cases and noncrash cases. Such a selection of threshold is commonly used to test the theoretical prediction accuracy of event and nonevent. For practical engineering applications, the threshold value should be carefully selected to achieve a proper false alarm rate according to the characteristics and the requirement of practical implementations. As shown in Fig. 4, there is a trade-off between crash prediction accuracy and false alarm rate. The crash prediction accuracy increases with an increase in the false alarm rate. When the models are used to estimate crash risks, drivers can be warned about the traffic conditions with high crash potential. A reasonably low false alarm rate should be selected to reduce the danger of losing drivers' responsiveness to the alerts. Even though a low false alarm rate leads to a significant reduction in the prediction accuracy of crash case, the developed model can still achieve about 18% prediction accuracy of crash case at the false alarm rate of 0.05.

When the model is used in the advanced dynamic safety management systems such as the variable speed limit and ramp metering systems, the false alarm rate is not as critical as it is for warning-based systems. In these applications, drivers do not know the reason of the change in a speed limit or a ramp metering rate. Besides, mitigating the turbulent traffic flow conditions can also provide operational improvements even though it might not lead to a crash (Ahmed et al. 2012a,b). Accordingly, in these applications, the thresholds can be selected to achieve a reasonable balance between prediction accuracy of crash cases and noncrash cases.

ROC curves were also developed for both models under different weather conditions using the whole dataset. The purpose of doing so was to compare the prediction performance of the full and reduced models under different weather conditions. Figure 6 illustrates the ROC curves for both models under adverse weather conditions, including light rain, moderate/heavy rain, haze, and mist/fog. The curve for the full model is always to the left of the curve for the reduced model, indicating the fact that the full model increases the prediction performance of the crash risk prediction model under adverse weather conditions. The ROC curves for the full and reduced models under clear weather conditions are illustrated in Fig. 7. The ROC curve for the full model almost coincides with the curve for the reduced model, indicating that the full model does not change the prediction performance of the crash risk prediction model under clear weather conditions. The results suggest that the inclusion of weather information in crash risk prediction models increases the overall prediction performance mainly by increasing the prediction accuracy for crashes during adverse weather conditions.

5. Summary and discussions

The study presented in this paper investigated the combined effects of weather and traffic flow characteristics on freeway crash risks. We also studied the benefits

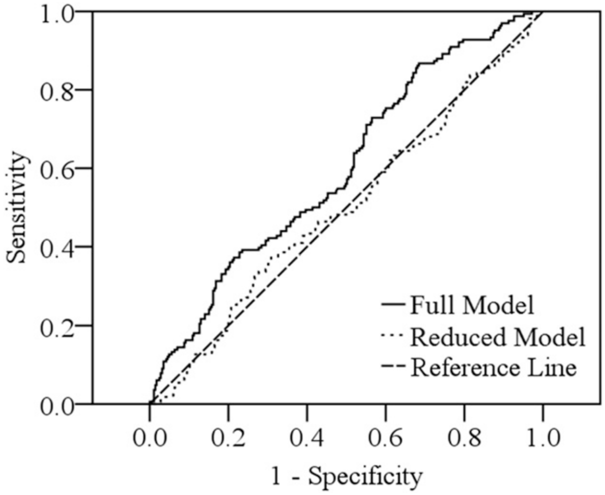

FIG. 6. ROC curves for the full and reduced models under adverse weather conditions.

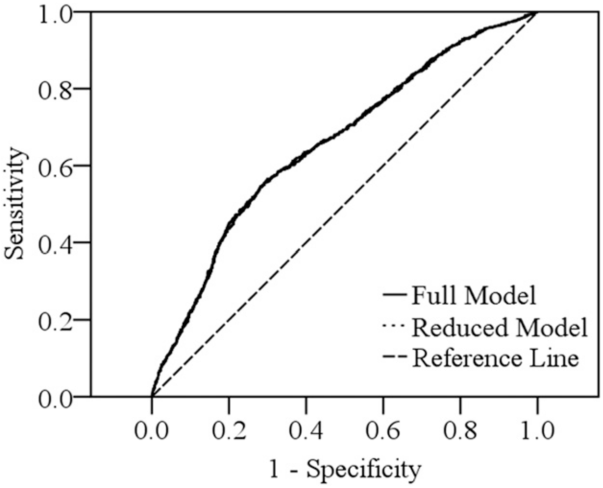

FIG. 7. ROC curves for the full and reduced models under clear weather conditions.

of incorporating weather information in freeway real-time crash risk prediction models that have traditionally focused on traffic flow parameters. Traffic and weather data were collected from a 35-km segment on the I-880 N freeway in the San Francisco Bay area. Based on visibility and precipitation intensity, the weather conditions were classified into five categories: clear, light rain, moderate/heavy rain, haze, and mist/fog. Logistic regression models using unmatched case-control data were developed to relate crash risks to various explanatory variables. Two issues were discussed when the crash risk models were developed. The first issue was related to the sample size requirements for case-control studies, and the second issue was related to the possible interaction between the weather and traffic flow variables.

The model specification results showed that adverse weather conditions, that is, light rain, moderate/heavy rain, and mist/fog, significantly increased the risks of crash occurrences on freeways. The results of the odds ratio analyses showed that the risks of crashes would increase by 1.875 times during moderate/heavy rains, and by 1.1 times in mist/fog as compared to the clear weather condition, given that other traffic flow variables remained the same. The interaction between the upstream occupancy and the light rain was found to be statistically significant. The bootstrap analysis results showed that the interaction effect of the upstream occupancy and the light rain on crash risks were 0.113 times less than the product of their individual effects and 1.095 times less than the sum of their individual effects.

For the purposes of comparison, a reduced crash risk model was also developed in which the weather information was not included. The results of the χ^2 test showed that the inclusion of weather variables significantly improved the goodness of fit of the crash risk model. The prediction performance of the full model and the reduced model were also compared. It was found that the full model produced better prediction performance for crash risks than did the reduced model. The incorporation of weather information in crash risk models improved the prediction accuracy of crash occurrences by 6.8% and reduced the false alarm rate by 1.3%. Assuming that 100 crashes will occur on a freeway in a future period, the full model can predict 7 more crashes than the reduced model. This is a considerable improvement for practical engineering applications. At the same time, the full model also provides improved prediction accuracy of noncrash cases. Note that the improvements in prediction performance were based on the dataset in which adverse weather accounted for only 10% of the total observations. It can be expected that the improvements in prediction performance will be more significant for the dataset with larger proportions of the observations in adverse weather conditions. The McNemar test indicates that the prediction performance of the full model is significantly higher than that of the reduced model. Accordingly, the inclusion of weather information leads to a considered improvement in predictive performance of the crash risk model.

ROC curves were further developed to compare the prediction performance of the full model and the reduced model across various probability thresholds and under different weather conditions. It was found that the full model always produced better prediction accuracy than the reduced model, no matter what probability threshold was selected. It was also found that the inclusion of weather information in crash risk prediction models increased the overall prediction performance mainly by increasing the prediction accuracy for crashes during adverse weather conditions. The inclusion of weather information had minor impacts on the prediction performance in clear weather conditions.

The crash risk models developed in this study have the potential to be used in freeway dynamic traffic control systems to help identify hazardous driving conditions. The crash risk models can also help us better understand the impacts of weather conditions and traffic flow characteristics on freeway crash risks. However, before the crash risk model is used in practical engineering applications, research is still needed to study the impacts of other adverse weather conditions, such as ice and snow on freeway crash risks. Because of the limitations of the data sources, the impacts of snowy weather on crash risks cannot be considered in the model. Further research is warranted to evaluate the interactive effect between real-time traffic flow conditions and snowy weather.

Acknowledgments. This research was sponsored by the National Natural Science Foundation of China (Grant 51508093) and Natural Science Foundation of Jiangsu Province (BK20171358). The authors thank the editor and the reviewers for their constructive comments and valuable suggestions to improve the quality of this article.

REFERENCES

Abdel-Aty, M., N. Uddin, F. Abdalla, A. Pande, and L. Hsia, 2004: Predicting freeway crashes from loop detector data using matched case-control logistic regression. *Transp. Res. Rec.*, **1897**, 88–95, https://doi.org/10.3141/1897-12.

Ahmed, M., M. Abdel-Aty, and R. Yu, 2012a: Assessment of the interaction between crash occurrence, mountainous freeway geometry, real-time weather and AVI traffic data. *Transp. Res. Rec.*, **2280**, 51–59, https://doi.org/10.3141/2280-06.

——, ——, and ——, 2012b: Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. *Transp. Res. Rec.*, **2280**, 60–67, https://doi.org/10.3141/2280-07.

Andrey, J., and S. Yagar, 1993: A temporal analysis of rain-related crash risk. *Accid. Anal. Prev.*, **25**, 465–472, https://doi.org/10.1016/0001-4575(93)90076-9.

Basso, F., L. J. Basso, F. Bravo, and R. Pezoa, 2018: Real-time crash prediction in an urban expressway using disaggregated data. *Transp. Res. C*, **86**, 202–219, https://doi.org/10.1016/j.trc.2017.11.014.

Bertness, J., 1980: Rain-related impact on selected transportation activities and utility services in the Chicago area. *J. Appl. Meteor. Climatol.*, **19**, 545–556, https://doi.org/10.1175/1520-0450(1980)019<0545:RRIOST>2.0.CO;2.

Brodsky, H., and S. Hakkert, 1988: Risk of a road accident in rainy weather. *Accid. Anal. Prev.*, **20**, 161–176, https://doi.org/10.1016/0001-4575(88)90001-2.

Brown, B., and K. Baass, 1997: Seasonal variation in frequencies and rates of highway accidents as a function of severity. *Transp. Res. Rec.*, **1581**, 59–65, https://doi.org/10.3141/1581-08.

Dimitriou, L., K. Stylianou, and M. A. Abdel-Aty, 2018: Assessing rear-end crash potential in urban locations based on vehicle-by-vehicle interactions, geometric characteristics and operational conditions. *Accid. Anal. Prev.*, **118**, 221–235, https://doi.org/10.1016/j.aap.2018.02.024.

Eisenberg, D., 2004: The mixed effects of precipitation on traffic crashes. *Accid. Anal. Prev.*, **36**, 637–647, https://doi.org/10.1016/S0001-4575(03)00085-X.

El Khoury, J., and A. Hobeika, 2006: Simulation of an ITS crash prevention technology at a no-passing zone site. *J. Intell. Transp. Syst.*, **10**, 75–87, https://doi.org/10.1080/15472450600626265.

Etemadi, H., A. Rostamy, and H. Dehkordi, 2009: A genetic programming model for bankruptcy prediction: Empirical evidence from Iran. *Expert Syst. Appl.*, **36**, 3199–3207, https://doi.org/10.1016/j.eswa.2008.01.012.

Freidlin, B., G. Zheng, Z. Li, and J. Gastwirth, 2002: Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Hum. Hered.*, **53**, 146–152, https://doi.org/10.1159/000064976.

Glickman, S., Ed., 2000: *Glossary of Meteorology*. 2nd ed. Amer. Meteor. Soc., 855 pp., http://glossary.ametsoc.org/.

Golob, T., and W. Recker, 2004: A method for relating type of crash to traffic flow characteristics on urban freeways. *Transp. Res. A*, **38**, 53–80, https://doi.org/10.1016/j.tra.2003.08.002.

——, ——, and V. Alvarez, 2004: Freeway safety as a function of traffic flow. *Accid. Anal. Prev.*, **36**, 933–946, https://doi.org/10.1016/j.aap.2003.09.006.

Hosmer, D., and S. Lemeshow, 1992: Confidence interval estimation of interaction. *Epidemiology*, **3**, 452–456, https://doi.org/10.1097/00001648-199209000-00012.

Hossain, M., and Y. Muromachi, 2011: Understanding crash mechanism and selecting appropriate interventions for real-time hazard mitigation on urban expressways. *Transp. Res. Rec.*, **2213**, 53–62, https://doi.org/10.3141/2213-08.

Hunter, M., P. Jenior, J. Bansen, and M. Rodger, 2011: Mode of flashing for malfunctioning traffic signals. *J. Transp. Eng.*, **137**, 438–444, https://doi.org/10.1061/(ASCE)TE.1943-5436.0000236.

Jung, S., X. Qin, and D. Noyce, 2010: Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accid. Anal. Prev.*, **42**, 213–224, https://doi.org/10.1016/j.aap.2009.07.020.

Kelsey, J., A. Whittemore, A. Evans, and D. Thompson, 1996: *Methods in Observational Epidemiology*. Oxford University Press, 432 pp.

Khattak, A., and K. Knapp, 2001: Interstate highway crash injuries during winter snow and nonsnow events. *Transp. Res. Rec.*, **1746**, 30–36, https://doi.org/10.3141/1746-05.

Knapp, K., L. Smithson, and A. Khattak, 2000: The mobility and safety impacts of winter storm events in a freeway environment. *Proc. Mid-Continent Transportation Symp.*, Ames, IA, Iowa State University, 67–71, http://www.ctre.iastate.edu/pubs/midcon/Knapp1.pdf.

Lee, C., F. Saccomanno, and B. Hellinga, 2003: Real-time crash prediction model for the application to crash prevention in freeway traffic. *Transp. Res. Rec.*, **1840**, 67–77, https://doi.org/10.3141/1840-08.

Lensberg, T., A. Eilifsen, and E. McKee, 2006: Bankruptcy theory development and classification via genetic programming. *Eur. J. Oper. Res.*, **169**, 677–697, https://doi.org/10.1016/j.ejor.2004.06.013.

Macfarlane, T., 2003: Sample size determination for research projects. *J. Orthod.*, **30**, 99–100, https://doi.org/10.1093/ortho/30.2.99.

Oh, C., J. Oh, and S. Ritchie, 2005: Real-time hazardous traffic condition warning system: Framework and evaluation. *IEEE Trans. Intell. Transp. Syst.*, **6**, 265–272, https://doi.org/10.1109/TITS.2005.853693.

Pande, A., A. Das, M. Abdel-Aty, and H. Hassan, 2011: Real-time crash risk estimation are all freeways created equal? *Transp. Res. Rec.*, **2237**, 60–66, https://doi.org/10.3141/2237-07.

Pirdavani, A., E. D. Pauw, T. Brijs, S. Daniels, M. Magis, T. Bellemans, and G. Wets, 2015: Application of a rule-based approach in real-time crash risk prediction model development using loop detector data. *Traffic Inj. Prev.*, **16**, 786–791, https://doi.org/10.1080/15389588.2015.1017572.

Qiu, L., and A. Nixon, 2008: Effects of adverse weather on traffic crashes systematic review and meta-analysis. *Transp. Res. Rec.*, **2055**, 139–146, https://doi.org/10.3141/2055-16.

Rothman, K. J., and S. Greenland, 1998: *Modern Epidemiology*. 2nd ed. Lippincott Williams and Wilkins, 758 pp.

Theofilatos, A., G. Yannis, P. Kopelias, and F. Papadimitriou, 2018: Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. *Accid. Anal. Prev.*, https://doi.org/10.1016/j.aap.2017.12.018, in press.

Vautard, R., P. Yiou, and J. Oldenborgh, 2009: Decline of fog, mist and haze in Europe over the past 30 years. *Nat. Geosci.*, **2**, 115–119, https://doi.org/10.1038/ngeo414.

Wang, G., and J. Kim, 2018: A large-scale neural network model for real-time crash prediction in urban road networks. *97th Transportation Research Board Annual Meeting*, Washington, DC, Transportation Research Board, 18-06519, https://trid.trb.org/view/1497304.

Wang, L., M. Abdel-Aty, Q. Shi, and J. Park, 2015: Real-time crash prediction for expressway weaving segments. *Transp. Res. C*, **61**, 1–10, https://doi.org/10.1016/j.trc.2015.10.008.

——, ——, and J. Lee, 2017: Safety analytics for integrating crash frequency and real-time risk modeling for expressways. *Accid. Anal. Prev.*, **104**, 58–64, https://doi.org/10.1016/j.aap.2017.04.009.

Washington, S., M. Karlaftis, and F. Mannering, 2003: *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman & Hall/CRC, 544 pp.

Xie, M., W. Cheng, G. S. Gill, R. Falahati, X. Jia, and S. Choi, 2017: Predicting likelihood of hit-and-run crashes using real-time loop detector data and hierarchical Bayesian binary logit model with random effects. *96th Transportation Research Board Annual Meeting*, Washington, DC, Transportation Research Board, 17-06376, https://trid.trb.org/view/1439542.

Xu, C., P. Liu, W. Wang, and Z. Li, 2012: Evaluation of the impacts of traffic states on crash risks on freeways. *Accid. Anal. Prev.*, **47**, 162–171, https://doi.org/10.1016/j.aap.2012.01.020.

——, H. Li, J. Zhao, J. Chen, and W. Wang, 2017a: Investigating the relationship between jobs-housing balance and traffic safety. *Accid. Anal. Prev.*, **107**, 126–136, https://doi.org/10.1016/j.aap.2017.08.013.

——, C. Wang, W. Wang, J. Bao, and M. Yang, 2017b: Investigating spatial interdependence in E-bike choice using spatially autoregressive model. *Promet Traffic Transp.*, **29**, 351–362, https://doi.org/10.7307/ptt.v29i4.2144.

——, Y. Wang, P. Liu, W. Wang, and J. Bao, 2018: Quantitative risk assessment of freeway crash casualty using high-resolution traffic data. *Reliab. Eng. Syst. Saf.*, **169**, 299–311, https://doi.org/10.1016/j.ress.2017.09.005.

Xu, F., and Z. Tian, 2008: Driver behavior and gap-acceptance characteristics at roundabouts in California. *Transp. Res. Rec.*, **2071**, 117–124, https://doi.org/10.3141/2071-14.

You, J., J. Wang, and J. Guo, 2017: Real-time crash prediction on freeways using data mining and emerging techniques. *J. Mod. Transp.*, **25**, 116–123, https://doi.org/10.1007/s40534-017-0129-7.

Yuan, J., M. Abdel-Aty, L. Wang, J. Lee, X. Wang, and R. Yu, 2018: Real-time crash risk analysis of urban arterials incorporating Bluetooth, weather, and adaptive signal control data. *97th Transportation Research Board Annual Meeting*, Washington, DC, Transportation Research Board, 18-00590, https://trid.trb.org/view/1494494.