

An Updated Geospatial Liquefaction Model for Global Application

by Jing Zhu, Laurie G. Baise, and Eric M. Thompson

Abstract We present an updated geospatial approach to estimation of earthquake-induced liquefaction from globally available geospatial proxies. Our previous iteration of the geospatial liquefaction model was based on mapped liquefaction surface effects from four earthquakes in Christchurch, New Zealand, and Kobe, Japan, paired with geospatial explanatory variables including slope-derived V_{S30} , compound topographic index, and magnitude-adjusted peak ground acceleration (PGA) from ShakeMap. The updated geospatial liquefaction model presented herein improves the performance and the generality of the model. The updates include (1) expanding the liquefaction database to 27 earthquake events across six countries, (2) addressing the sampling of non-liquefaction for incomplete liquefaction inventories, (3) testing interaction effects between explanatory variables, and (4) overall improving the model's performance. We inspected 14 geospatial proxies for soil density and soil saturation; the most promising of these are slope-derived V_{S30} , modeled water table depth, distance to coast, distance to river, distance to the closest water body, and precipitation. We found that peak ground velocity (PGV) performs better than PGA as the shaking intensity parameter. We present two models which offer improved performance over prior models. We evaluate model performance using the area under the receiver operating characteristic curve, and the Brier score. The best-performing model in a coastal setting uses distance to the coast but is problematic for regions away from the coast. The second best model, using PGV, V_{S30} , water table depth, distance to the closest water body, and precipitation, performs better in noncoastal regions and thus is the model we recommend for global implementation.

Introduction

Soil liquefaction can lead to significant infrastructure damage after an earthquake, due to lateral ground movements and vertical settlements. Regional liquefaction-hazard maps are important in both planning for earthquake events and guiding relief efforts by positioning resources once the events have occurred. Most liquefaction hazard-mapping techniques rely on detailed geologic maps and geotechnical data, such as standard penetration test (SPT) or cone penetration test (CPT) results, fines content, and water table depth (Holzer *et al.*, 2006, 2009; Brankman and Baise, 2008), which are not always available in at-risk regions or with sufficient density and coverage.

We developed a regional liquefaction-mapping approach that relies on broadly available geospatial parameters (Baise *et al.*, 2012; Zhu *et al.*, 2013, 2014, 2015). The motivation for the work comes from the rapid response and loss estimation communities, in which there is a need to predict regional liquefaction extent for any earthquake around the globe. Our work builds on previous works, such as Youd and Perkins (1978), which characterized the relationship between geologic depositional environments and liquefaction susceptibility, and Wald and Allen (2007), which identified a first-order

approximation of soil conditions from topography. As a direct precursor to our work, Knudsen and Bolt (2011) found that liquefaction occurrences commonly coincide with simple geospatial features such as topographic slope and distance to the closest river. In our previous work (Zhu *et al.*, 2015), we developed a liquefaction occurrence/nonoccurrence database that was unbiased with respect to the spatial extent (i.e., complete coverage of liquefaction and nonliquefaction occurrence over the mapped area) using data from Christchurch, New Zealand, and Kobe, Japan. We tested geospatial parameters as proxies for earthquake loading, soil density, and soil saturation, and developed a logistic model (hereafter termed GLM-Zea15g for geospatial liquefaction model by Zhu *et al.*, 2015; the “g” specifies the global model from that paper) to predict the probability of liquefaction after an event. The model provides a first-order estimate of the spatial coverage of liquefaction from simple geospatial parameters and can be implemented for loss estimation and rapid response. Recent work by Matsuoka *et al.* (2015) has followed a similar approach for a Japanese liquefaction dataset, but their work relies on the geomorphological classification map of Japan, which is not available globally.

Although the results of [Zhu *et al.* \(2015\)](#) demonstrate the feasibility of the geospatial approach for predicting regional liquefaction extent, further improvements and refinements can be achieved with additional data and analysis. GLM-Zeal5g was derived from liquefaction inventories in two regions, both of which were coastal sedimentary basins, with an additional qualitative comparison with the liquefaction that occurred in Port-au-Prince, Haiti. For empirical model development, the quality of a database greatly influences the performance of the model. Increasing the number of samples and sampling a broader range of explanatory variables improves the generality of the model and therefore improves the performance when applying it to make future prediction ([Hastie *et al.*, 2001](#)). The results in recent updates to the empirical correlations of liquefaction with *in situ* soil indexes such as the SPT, CPT, and shear-wave velocity ([Cetin *et al.*, 2004](#); [Kayen *et al.*, 2013](#); [Boulanger and Idriss, 2015](#)) consistently show that an expanded case history database with increased sample size and diversity from different earthquake regions can result in an improved model with reduced overall model uncertainty. For example, a model may perform well when developed and assessed with all large-magnitude earthquakes, but it may perform poorly if it is then applied to small-magnitude earthquakes. Although these geotechnical liquefaction models are most relevant for site-specific studies, the lessons learned also apply to our development of a regional geospatial liquefaction model.

The objective of this article is to improve the performance of the geospatial liquefaction model, especially for generalization across broad geographic regions. The [Zhu *et al.* \(2015\)](#) dataset was intentionally limited to spatially complete inventories so that the probabilities from the model could inherently represent the spatial extent of the surface expression of liquefaction. In the updated efforts presented herein, we expand the liquefaction database for testing and improving the model. We compiled liquefaction data from journal articles and reconnaissance reports, from 23 additional earthquakes from the United States, Japan, China, Taiwan, and India. We added a combination of datasets with extensively mapped liquefaction, as well as specific events from underrepresented geologic regions (inland earthquakes or dry regions), and events with little to no liquefaction. For example, the Northridge and Hector Mine, California, earthquakes provide samples in relatively dry areas that rarely experience liquefaction due to the arid climate. We included earthquakes in which no liquefaction was observed to further explore the parameter space.

To expand the database, we include inventories that are not spatially complete. The majority of the added events were documented as incidences of liquefaction (either as points or limited polygons) without information on spatial completeness. This is an artifact of the data collection efforts from these historical events. A majority of the data collected is from field investigations without the more systematic coverage provided with remote-sensing techniques. The spatially incomplete nature of the newly added data restricts

our ability to preserve the actual class imbalance (i.e., ratio of liquefaction occurrence to nonoccurrence), and class imbalance has a strong influence on the probabilities of the model ([Oommen *et al.*, 2011](#); [Zhu *et al.*, 2015](#); [Thompson *et al.*, 2016](#)). To address the lack of observations of nonliquefaction in many of the newly added inventories, we use a sampling scheme to sample nonliquefaction data, which is similar to the one that has been successfully applied in landslide-hazard mapping by [Van Den Eeckhaut *et al.* \(2012\)](#) to address the incompleteness of landslide inventories. By addressing class imbalance, we significantly expanded our database, which, in turn, provides the opportunity to constrain a more complex functional form with additional model parameters.

In this article, we first provide the details of the expanded liquefaction database, including the geospatial parameters that we compile as candidate explanatory variables. Then, we describe the modeling process and present two alternative models. We evaluate our models in terms of mapped liquefaction extent and receiver operating characteristic (ROC) curves and quantitatively compare the models using statistical goodness-of-fit measures. Additionally, we compare the model results with prior regional susceptibility studies in San Francisco and Seattle.

Data

Liquefaction Database

The expanded liquefaction database includes 27 earthquakes from the United States, Japan, New Zealand, China, Taiwan, and India. [Figure 1](#) shows maps with the locations of the earthquake epicenters. The details of each event included in the database are summarized in [Table 1](#). It is important to note that the liquefaction database consists of inventories from earthquakes that triggered liquefaction as well as earthquakes with insignificant or no liquefaction. Building a well-distributed database of liquefaction and nonliquefaction locations across a multidimensional parameter space and geographic space is important for developing a useful and general model. We ensure a well-distributed database by sampling a variety of earthquakes in terms of magnitude as well as geologic setting.

As evident from [Figure 1](#), a majority of the events are located in coastal areas. We define a coastal event as one where the liquefaction occurrences are, on average, within 20 km of the coast; or, for earthquakes with insignificant or no liquefaction, epicentral distances less than 50 km. Based on these criteria, there are five noncoastal events with observed liquefaction in this database: the 1994 M_w 6.6 Northridge (number 9), 2001 M_w 7.7 Bhuj (number 21), 1999 M_w 7.6 Chi-Chi (number 20), 1999 M_w 7.1 Hector Mine (number 26—an earthquake with insignificant or no liquefaction), and 2008 M_w 7.9 Wenchuan (number 27) events. The data from the Bhuj earthquake were only used for verification, but not included in the model development,

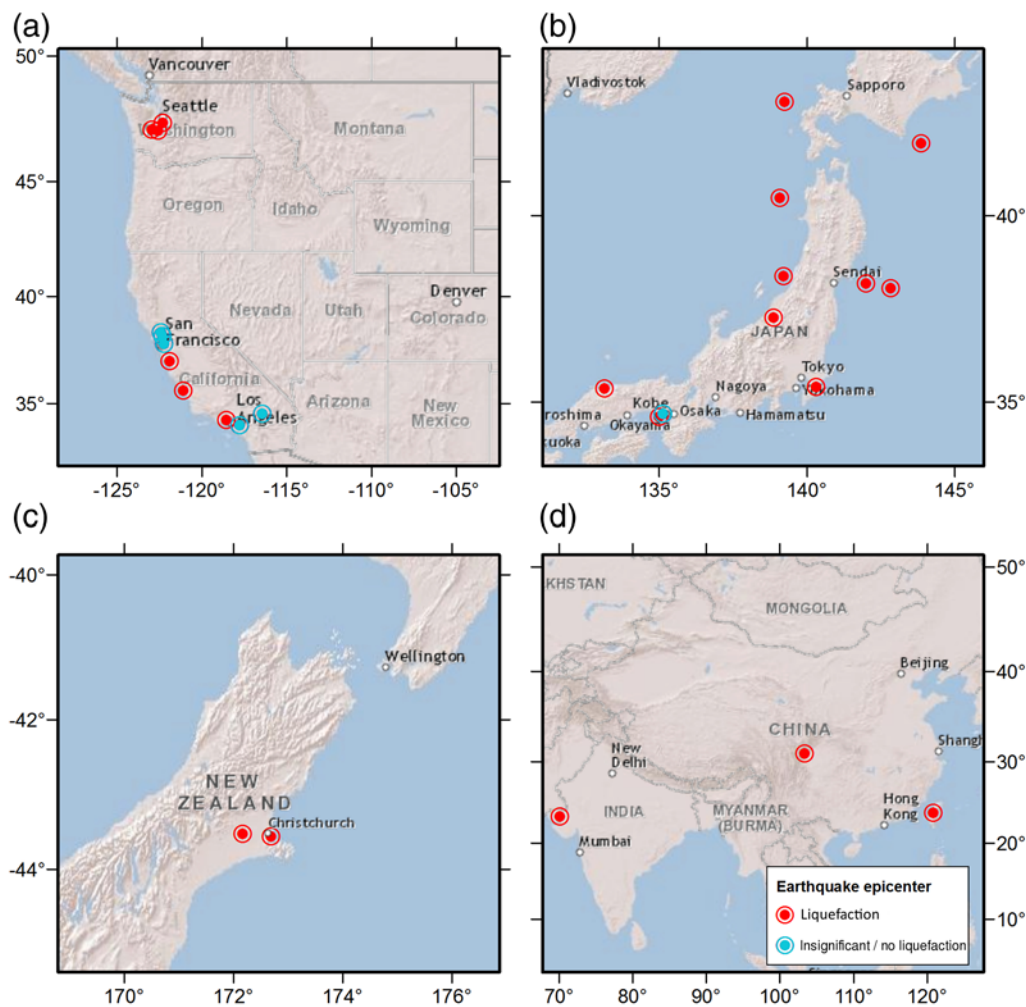


Figure 1. Maps of earthquakes in the liquefaction database from (a) the United States, (b) Japan, (c) New Zealand, and (d) India, China, and Taiwan. The color version of this figure is available only in the electronic edition.

because they were from a remote sensing study (Singh *et al.*, 2002) that lacks validation. Although the Northridge earthquake led to shaking in the coastal environment, the mean distance to the coast for the liquefaction observations was 25 km. A sixth noncoastal event, the 2015 M_w 7.8 Nepal earthquake, is used for verification but not included in the database. The low number of noncoastal events is not entirely surprising because the majority of tectonically active regions are coastal, and liquefaction is known to occur in coastal sediments (e.g., artificial fill, beach deposits, and alluvial and marine sands) as documented by Youd and Perkins (1978), and numerous well-studied earthquakes, such as the 1989 Loma Prieta (Tinsley *et al.*, 1998), 1995 Hyogo-ken Nanbu (Hamada *et al.*, 1995), and 2011 Tohoku (Ministry of Land, Infrastructure, Transport and Tourism [MLITT], 2011) events. Because the database is biased toward coastal events, we will investigate the portability of our results outside of the coastal setting in the Models section.

For the purpose of this article, events can be classified as complete or incomplete datasets. This is an important distinc-

tion in terms of how the datasets are sampled for liquefaction and nonliquefaction points. A complete dataset includes events like the 1995 Kobe (number 3) and the 2010–2011 Darfield and Christchurch (numbers 1 and 2) earthquakes that are extensively mapped as polygons of liquefaction occurrence. In a complete dataset, nonliquefaction points can be sampled anywhere within the mapped extent that is not covered by a liquefaction polygon. For our purposes, complete datasets also include well-studied events in which liquefaction is insignificant or absent. Examples of this include the 2003 M_w 4.2 Kobe event (number 4) and the 2014 M_w 6.0 Napa (number 24) event. These earthquakes with insignificant or no liquefaction provide complete datasets that are very important for providing coverage of the parameter space. The first four earthquakes (numbers 1–4) in the database were used for developing the prior model (Zhu *et al.*, 2015). Except for the 2010–2011 Darfield and Christchurch (numbers 1 and 2), and some Japanese events (numbers 3, 11–18), in which we have access to data in digital format (see Data and Resources; Wakamatsu, 2011), we

Table 1
Summary of Earthquakes in the Liquefaction Database

Number	Earthquake	Date (yyyy/mm/dd)	M_w	Liquefaction Number of Points	Liquefaction Area of Polygons (km ²)	Spatial Category	Liquefaction Occurrence	Coastal or Noncoastal	Reference
1	Christchurch	2011/02/22	6.1	—	71.53	Complete	Yes	Coastal	CGD*
2	Darfield	2010/09/04	7	—	69.18	Complete	Yes	Coastal	CGD*
3	Hyogo-ken Nambu	1995/01/17	6.9	1883	11.44	Complete	Yes	Coastal	Wakamatsu (2011)
4	Kobe	2003/10/08	4.2	—	—	Complete	No [†]	Coastal	
5	Loma Prieta	1989/10/17	6.9	123	7.72	Incomplete	Yes	Coastal	Tinsley <i>et al.</i> (1998)
6	Puget Sound	1949/04/13	6.9	153	—	Incomplete	Yes	Coastal	Chleborad and Schuster (1998)
7	Puget Sound	1965/04/29	6.7	229	—	Incomplete	Yes	Coastal	Chleborad and Schuster (1998)
8	Nisqually	2001/02/28	6.8	69	—	Incomplete	Yes	Coastal	Bray <i>et al.</i> (2001)
9	Northridge	1994/01/17	6.6	36	—	Incomplete	Yes	Noncoastal	Stewart <i>et al.</i> (1994, 1996), Moehle <i>et al.</i> (1994)
10	San Simeon	2003/12/22	6.6	10	0.02	Incomplete	Yes	Coastal	Holzer <i>et al.</i> (2005)
11	Hokkaido Nansei-oki	1993/07/12	7.7	376	4.19	Incomplete	Yes	Coastal	Wakamatsu (2011)
12	Chiba-ken Toho-oki	1987/12/17	6.5	67	7.39	Incomplete	Yes	Coastal	Wakamatsu (2011)
13	Miyagi-ken-oki	1978/06/12	7.6	50	0.29	Incomplete	Yes	Coastal	Wakamatsu (2011)
14	Niigata	1964/06/16	7.6	124	71.64	Incomplete	Yes	Coastal	Wakamatsu (2011)
15	Nihonkai Chubu	1983/05/26	7.7	12	54.44	Incomplete	Yes	Coastal	Wakamatsu (2011)
16	Niigata-ken Chuetsu	2004/10/23	6.6	313	14.7	Incomplete	Yes	Coastal	Wakamatsu (2011)
17	Tottori-ken Seibu	2000/10/06	6.7	120	2.61	Incomplete	Yes	Coastal	Wakamatsu (2011)
18	Tokachi-oki	2003/09/26	8.3	139	0.02	Incomplete	Yes	Coastal	Wakamatsu (2011)
19	Tohoku	2011/03/11	9.1	—	107.07	Locally complete	Yes	Coastal	Ministry of Land, Infrastructure, Transport and Tourism (MLITT) (2011)
20	Chi-Chi	1999/09/21	7.6	170	2.52	Incomplete	Yes	Noncoastal	Chu <i>et al.</i> (2004)
21	Bhuj [‡]	2001/01/26	7.6	—	377.4	Complete	Yes	Noncoastal	Singh <i>et al.</i> (2002)
22	Yountville	2000/09/03	5.0	—	—	Complete	No [†]	Coastal	
23	Piedmont	2015/08/17	4.0	—	—	Complete	No [†]	Coastal	
24	Napa	2014/08/24	6.0	2	—	Complete	No [†]	Coastal	Geo-Engineering Extreme Events Reconnaissance (GEER) Association (2014)
25	Chino Hills	2008/07/29	5.4	—	—	Complete	No [†]	Coastal	Geo-Engineering Extreme Events Reconnaissance (GEER) Association (2008)
26	Hector Mine	1999/10/16	7.1	—	—	Complete	No [†]	Noncoastal	
27	Wenchuan	2008/05/12	7.9	116	—	Incomplete	Yes	Noncoastal	Cao <i>et al.</i> (2010)

*Canterbury geotechnical database (see [Data and Resources](#))

[†]Insignificant to no liquefaction.

[‡]Not included in the model development.

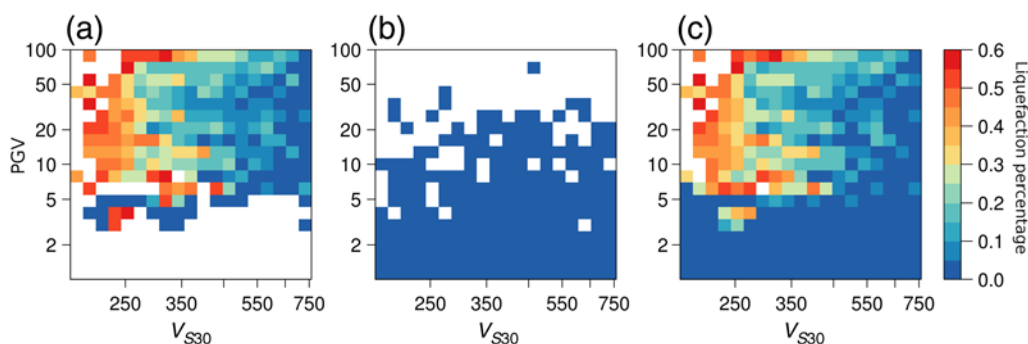


Figure 2. Percentage of liquefaction versus nonliquefaction as a function of peak ground velocity (PGV) and V_{S30} for (a) liquefaction events, (b) events with insignificant or no liquefaction, and (c) all events in the database. The color version of this figure is available only in the electronic edition.

obtained the rest of the data in Table 1 by digitizing published liquefaction maps (see Table 1 for references).

To illustrate the benefit of adding the earthquakes with insignificant or no liquefaction to the database, we illustrate the data space represented by plotting the proportion of liquefaction occurrences as a function of peak ground velocity (PGV) and the slope-derived time-averaged shear-wave velocity to 30 m depth (V_{S30}) in Figure 2. To construct this figure, we sampled the liquefaction inventories using the same sampling method described in the [Sampling](#) section. In this figure, white cells indicate no data points in the interval. As shown in Figure 2a, when the database only contains earthquakes with extensive liquefaction, there are almost no data in which PGV is less than 3 cm/s. The lack of data at low PGV may result in false model predictions for small events (low PGV) in which liquefaction is not expected to occur. After adding the earthquakes with insignificant or no liquefaction (Fig. 2b), the data space for low-PGV values is sufficiently filled, and the boundary between liquefaction and nonliquefaction becomes better differentiated (Fig. 2c).

As discussed above, the liquefaction data can be divided into two groups based on the spatial completeness, and this categorization impacts how liquefaction and nonliquefaction points are sampled. The group that is spatially complete consists of the four events used for developing the prior model (numbers 1–4) and the insignificant-to-no liquefaction complete events (numbers 22–26). The 2011 Tohoku earthquake (number 19) was spatially complete in a limited region (MLITT, 2011) but not for the entire affected area, so we treated it as an incomplete dataset. For this group, liquefaction observations are represented by polygons and the mapped extent associated with the liquefaction data is documented. Because the regions were well studied as evidenced by the reconnaissance reports and the postevent literature, we can assume that the liquefaction was completely mapped, and therefore nonliquefaction can be assumed outside the liquefaction polygons and within the mapped extent. This type of data is only available for a few events, generally for which remote sensing has been incorporated into the postearthquake data collection or in well-studied regions like the San Francisco

Bay area in the United States or large urban regions in Japan where the field reconnaissance was extensive.

Regrettably, a majority of events in Table 1 (numbers 5–21) are spatially incomplete and result from limited field investigations described in the literature. In general, the available data are mapped liquefaction occurrences, predominantly represented by points (limited polygons may exist for some local areas) with limited information about the extent of mapping. As a result, additional assumptions are required to obtain liquefaction nonoccurrence data, as discussed in the [Sampling](#) section.

Geospatial Predictors

In parallel to the liquefaction/nonliquefaction occurrence data, we assemble a database of explanatory variables for use in building the liquefaction models. We consider explanatory variables that can approximate some of the governing factors for whether liquefaction will or will not occur: soil density, soil saturation, and earthquake loading. We only consider variables that can be easily derived at a global scale, and so some factors that affect the occurrence of liquefaction (e.g., soil plasticity) are not considered in our approach. Table 2 summarizes the explanatory variables that we tested for the geospatial liquefaction model, which are expanded from those tested in [Zhu et al. \(2015\)](#). The spatial resolution for all variables is 30 arcsec.

Soil density is an important liquefaction susceptibility factor; loose soils are more susceptible than dense soils. [Wald and Allen \(2007\)](#) demonstrated that soil density is correlated with topographic gradient using California, Taiwan, Utah, and the Mississippi embayment as test cases. Although recent work has demonstrated that the [Wald and Allen \(2007\)](#) correlations are less effective in some regions (e.g., [Magistrale et al., 2012](#)), the method is still appropriate for broad applicability and is used globally with the U.S. Geological Survey ShakeMap to estimate soil amplification after earthquakes ([Worden et al., 2010](#)). We use the digital elevation model (DEM) from the Global Multi-resolution Terrain Elevation Data 2010 (see [Data and Resources](#)). Several geospatial variables were computed directly from the elevation

Table 2
Summary of All Candidate Explanatory Variables

Variable Description	Variable Name	Density	Saturation	Load
Shear-wave velocity over the first 30 m (slope derived)	V_{S30}	•		
Elevation	<i>elev</i>	•		
Topographic slope	<i>slope</i>	•		
Roughness	<i>rough</i>	•		
Topographic position index	TPI	•		
Terrain roughness index	TRI	•		
Distance to the nearest coast	<i>dc</i>	•	•	
Compound topographic index	CTI		•	
Global water table depth	<i>wtd</i>		•	
Distance to the nearest river	<i>dr</i>		•	
Distance to the nearest water body	<i>dw</i>		•	
Elevation above the nearest water body	<i>hwater</i>		•	
Mean annual precipitation	<i>precip</i>		•	
Aridity index	AI		•	
Peak ground acceleration	PGA			•
Peak ground velocity	PGV			•
Magnitude	M_w			•
Magnitude-scaling factor	MSF			•

data, such as slope and V_{S30} . Slope is calculated using the *gdgradient* command in Generic Mapping Tools software (see [Data and Resources](#)). V_{S30} is estimated from slope using the method described in [Wald and Allen \(2007\)](#). We use the coefficients for active tectonic regions because all the earthquakes in the database are in active tectonic regions.

Because surface texture is often used in landform classifications, we consider three indexes such as roughness, topographic position index (TPI), and terrain roughness index (TRI). Roughness is defined as the largest intercell difference of a central pixel and its eight surrounding cells. TPI is defined as the difference between a central pixel and the mean of its eight surrounding cells. TRI is defined as the mean difference between a central pixel and its eight surrounding cells. They are computed from the DEM using the *gdaldem* command in the Geospatial Data Abstraction Library (see [Data and Resources](#)) based on the definitions described in [Wilson et al. \(2007\)](#). As with the gradient, these roughness measurements are dependent upon the resolution of the DEM from which they are derived.

Soil saturation is an important parameter in liquefaction analyses because the soil has to be saturated or partially saturated in order to liquefy. Because water flows downhill and accumulates in streams, rivers, lakes, and oceans, soil saturation is generally correlated with proximity to water bodies and regional climate conditions. We used saturation proxies that are derived from topography, climate data, and groundwater models. We compute distance to the river (*dr*) using the rivers from the Hydrological data and maps based on Shuttle Elevation Derivatives at multiple Scales (HydroSHEDS) database (see [Data and Resources](#)), which were de-

rived from topography. We compute the distance to the nearest coast (*dc*) from a global dataset computed by National Aeronautics and Space Administration (NASA)’s Ocean Color Group (see [Data and Resources](#)). The distance to the nearest water body (*dw*) is calculated as the minimum value of *dr* and *dc*. We do not explicitly include proximity to lakes because it is approximately accounted for in the HydroSHEDS river data ([Lehner et al., 2006](#)). We derive elevation above the nearest water body (*hwater*) from the river layer and elevation data. We use the compound topographic index (CTI; [Beven and Kirkby, 1979](#)) from the HydroSHEDS database (see [Data and Resources](#)). CTI (i.e., wetness index) is calculated as the flow accumulation divided by the tangent of the topographic slope.

In addition to the saturation proxies derived from topography, we also include globally available climate information as an input to our estimation to differentiate the soil saturation across different climate regions, such as the relatively dry region affected by the 1999 M_w 7.1 Hector Mine earthquake (no liquefaction), versus the wet region affected by the 2001 M_w 6.8 Nisqually earthquake (extensive liquefaction). The mean annual precipitation for the former region is ~ 200 mm, whereas the mean annual precipitation for the latter region is ~ 1200 mm. We consider mean annual precipitation (*precip*) from the WorldClim database (see [Data and Resources](#)) and aridity index (AI) from the Consultative Group on International Agriculture Research-Consortium for Spatial Information (CGIAR-CSI) Global Aridity dataset (see [Data and Resources](#)). The global precipitation dataset was developed by interpolating from over 40,000 weather stations across the world and averaging over the 1959–2000 time periods. To incorporate the recent efforts of large-scale groundwater modeling, we use a global dataset of water table depth (*wtd*) from [Fan et al. \(2013\)](#), who modeled groundwater flow using a model constrained by climate, terrain, and sea level, and calibrated it with over 1.5 million published records.

Finally, earthquake magnitude and intensity parameters are important in liquefaction analyses because ground shaking of a contractive soil can lead to pore-water pressure increase, which is a necessary component of liquefaction. The effects of earthquake loading are modeled by ground-shaking parameters and proxies for earthquake duration. For ground-shaking parameters, we consider peak ground acceleration (PGA) and PGV from ShakeMap (see [Data and Resources](#)), which provides near-real-time estimates of ground shaking that incorporates macroseismic data as well as available ground-motion records with ground-motion prediction equation (GMPE) estimates ([Worden et al., 2010](#)). To approximate earthquake duration, we consider the magnitude-scaling factor (MSF) in [Youd et al. \(2001\)](#).

To demonstrate the typical patterns that these geospatial explanatory variables exhibit, maps for eight potential explanatory variables for the San Francisco Bay area and two shaking variables for the 1989 Loma Prieta earthquake are shown in Figure 3. Parameters such as V_{S30} , *wtd*, CTI, and *hwater* are correlated with topography and show similar

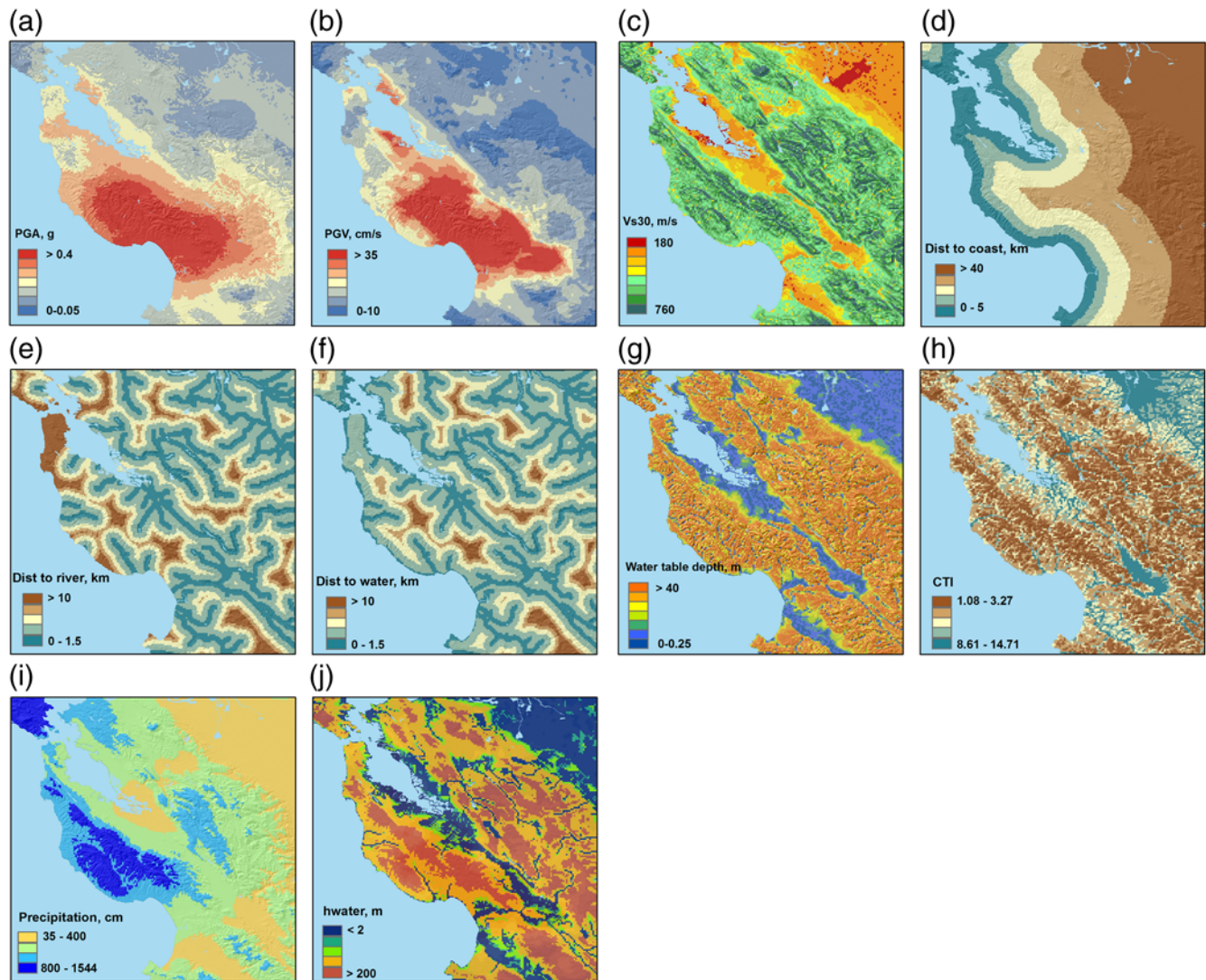


Figure 3. Maps of candidate explanatory variables: (a) peak ground acceleration (PGA), (b) PGV, (c) V_{S30} , (d) distance to the nearest coast (dc), (e) distance to river (dr), (f) distance to the nearest water body (dw), (g) water table depth (wtd), (h) compound topographic index (CTI), (i) mean annual precipitation ($precip$), and (j) elevation above the nearest water body ($hwater$) for the San Francisco Bay area. PGA and PGV are shown for the 1989 Loma Prieta earthquake; all other variables are event independent. The color version of this figure is available only in the electronic edition.

patterns. Also, parameters like dr and dc vary slowly and provide soil saturation proxy at a lower resolution, as compared to parameters such as CTI or wtd which vary much more locally.

Methods

Sampling

To create the liquefaction database with liquefaction occurrence and nonoccurrence, as well as all relevant geospatial explanatory variables, sampling grids are populated at 100 m spacing for each earthquake region in a local Cartesian coordinate system. A grid pixel is labeled as liquefaction when there is a liquefaction point inside the pixel, or 30% of the grid pixel is covered by a liquefaction polygon. The 30% threshold is selected to rasterize polygons because

we found that for the spatially complete data, the 30% threshold retains the same liquefaction to nonliquefaction ratio before and after the rasterization (Zhu *et al.*, 2015). Because the absence of liquefaction is generally not documented in earthquake inventories, we developed a strategy to sample nonliquefaction data. For a complete dataset in which the liquefaction data are documented within a mapping extent, we randomly sample nonliquefaction from the pixels not covered by a liquefaction polygon. For an incomplete dataset in which the mapping extent is unknown, we apply circular buffers around known liquefaction locations to sample nonliquefaction (illustrated in Fig. 4). This allows for holes in the sampled regions, which will be determined by the distribution of the liquefaction locations and the buffer size. We use a nonsampling buffer immediately around each liquefaction

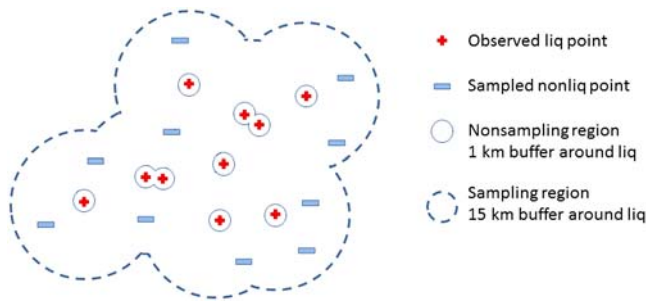


Figure 4. Sampling nonliquefaction events using spatial buffers. The color version of this figure is available only in the electronic edition.

point because we do not know the spatial extent of the recorded liquefaction feature. The sampling region is defined as the area outside of the nonsampling buffer and within the sampling region. After testing a range of buffer widths (see the [Sensitivity Analyses of Sampling Choices](#) section), we sample nonliquefaction pixels within an area that is 1–15 km from an observed liquefaction pixel. Figure 4 is a schematic illustration of the spatial buffer, with a solid circle showing the boundary of the inner buffer and the dashed line showing the boundary of the outer buffer. For events with insignificant or no liquefaction, the sampling region is determined by the area where the ShakeMap intensities are available.

Logistic Model

We use logistic regression to model the probability of liquefaction. Logistic regression is a statistical approach that can be used to describe the relationship of several independent variables to a binary dependent variable. The use of the logistic equation ensures that the resultant probability lies in the range between zero and one:

$$P(X) = \frac{1}{1 + e^{-X}}, \quad (1)$$

in which $X = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$, x_1, x_2, \dots, x_k are the explanatory variables, and $\beta_0, \beta_1, \dots, \beta_k$ are the coefficients estimated from the regression. We use the maximum-likelihood method to obtain these estimates ([Kleinbaum and Klein, 2010](#)).

Sampling Strategy

Because the events in the database were mapped with different levels of detail (polygons vs. points), the number of data points for an event in the database does not necessarily correlate with the extent of liquefaction that occurred. An event that includes mapped polygons results in considerably more data points than an event that only includes liquefaction points. For example, the 2011 Christchurch earthquake has 8867 data points, whereas the 1989 Loma Prieta event has 789 liquefaction points. To prevent a specific event from dominating the regression results, we randomly sample 1000 liquefac-

tion points from each event for model development. For an event in which the number of liquefaction points is less than 1000, we use all the liquefaction points available for that event. To improve model stability and avoid overfitting, we resample 50 times for each model and average the model coefficients. Investigation of the variability of model performance and coefficients across different random samples constrains the sensitivity to the sampling scheme.

In [Zhu *et al.* \(2015\)](#), we chose to use an imbalanced dataset ($\sim 1:13$ liquefaction:nonliquefaction) because we aimed to develop a probability estimator that predicts the areal extent of liquefaction. In other words, we wanted the resulting probability to correlate with spatial extent (e.g., areas labeled 10% probability of liquefaction will contain about 10% liquefaction by area). This is only possible for complete inventories. As discussed in the [Introduction](#), class imbalance significantly influences the resulting probabilities from logistic regression models. However, in the current approach, we include incomplete datasets to improve the generality of the model. As a result, the class imbalance in the current database no longer represents the actual class imbalance. As an alternative approach, we can optimize the model as a model classifier (i.e., discriminating between occurrences and nonoccurrences). Many studies have shown that for several classifiers, a balanced dataset provides improved overall classification performance compared with an imbalanced dataset ([Laurikkala, 2001](#); [Weiss and Provost, 2001](#); [Estabrooks *et al.*, 2004](#)). To minimize the effect of class imbalance and optimize the database for the development of a classifier, we sample equal numbers of liquefaction and nonliquefaction points (e.g., 1000 points each). For events with insignificant or no liquefaction, we sample 1000 nonliquefaction points. Sampling so that the full database has a 1:1 class balance is another reasonable approach, but it is not the choice we made. For events with insignificant or no liquefaction, the sampling region is such that the ShakeMap intensities are available. In the [Interpretation of the Predicted Probabilities](#) section, we evaluate the interpretation of the probabilities in terms of expected spatial extent of liquefaction to understand the impact of using a balanced sample.

Modeling Strategy

Our goal is to develop a model that not only fits the available data well, but will fit new data that were not used to develop the model. Thus, we select a simple model that reflects as much of the underlying physics of the problem as possible. Our modeling strategy involves four stages: (1) exploratory data analysis, (2) interaction assessment, (3) base model selection (i.e., an initial model that does not account for saturation), and (4) a stepwise assessment of alternative saturation parameterizations. Exploratory data analysis is carried out first because we think it is important to understand the distributions and relationships between the liquefaction/nonliquefaction data and explanatory variables. We plot the histograms of the liquefaction and nonliquefaction points (for a single set of

sampled observations), as well as the estimated probability of liquefaction over the range of each candidate variable. We use this to identify gaps in the data space as well as variables that are strongly predictive.

Next, we assess interaction effects of individual pairs of candidate variables. This is addressed prior to the final model selection because if there is evidence of interaction involving certain variables, then the interaction term (e.g., the product of two interacting variables) must be considered in the model selection stage. Then, we establish a base model by selecting variables that show strong correlations with the probability of liquefaction. We prefer using a base model to guide the model selection rather than completely relying on the performance measures. Finally, we test the base model with additional candidate variables to determine if their combinations can improve the model performance.

In the model selection stage, we evaluate candidate functional forms on a dataset sampled from the entire database using the described sampling method. There are various ways to measure the performance of a statistical prediction model. We use the [Brier \(1950\)](#) score to quantify how close predictions are to the actual outcome and area under the ROC curve (AUC), to quantify discrimination (do liquefied locations have higher predicted probabilities than those that did not?). The Brier score measures the mean squared difference between predicted probabilities and actual outcomes. The Brier score for a model can range from 0 for a perfect model to 0.25 for a noninformative model. Useful performance metrics for binary classifications include the true positive rate (TPR), which measures the fraction of positive cases that are correctly classified, and false positive rate (FPR), which measures the fraction of negative cases that are misclassified as positive. The ROC curve has proved a useful tool for evaluating empirical liquefaction models ([Oommen et al., 2010](#); [Maurer et al., 2015](#)). An ROC curve is a plot of the TPR against FPR at various probability thresholds. For a given threshold, a model that perfectly predicts the binary response would have $TPR = 1$ and $FPR = 0$. The closer the ROC curve comes to this ideal case (i.e., the top-left edge of the plot), the better the model performance. Thus, AUC is a scalar measure that quantifies the accuracy of the probabilistic classifier, because as the AUC increases the ROC curve approaches the top-left edge of the plot. The AUC limits are from 0.5 (random classification) to 1.0 (perfect accuracy).

Results

Data Exploration

We create a regression dataset from the entire database using the described sampling method. To explore the data and understand the correlation between individual explanatory variables and the liquefaction and nonliquefaction data, we present histograms of liquefaction and nonliquefaction data over the range of each candidate variable (including

transformations) in [Figure 5](#). On the same plots, the ratio of liquefaction points within each bin is shown as a gray dot in which the darkness of the gray dot increases with the number of data points within the bin (the scale is different for each panel). In other words, the darkness of the point is an indication of weight. The curve represents a univariate logistic model fit to the data.

In the plots, shaking parameters such as PGA and PGV are transformed by taking their natural logarithm, because their distributions are well represented by a lognormal distribution. Similarly, we use the natural logarithm of V_{S30} . We notice that the distributions of a few variables are sharply skewed, such as *dc*, *wtd*, *elev*, TRI, and roughness, with a greater density of data having values close to zero. As a result of the skew, the occurrence of liquefaction is more sensitive to changes in small values than large values. Therefore, we prefer to apply a square root transformation to increase the weights of small values. In the [Models](#) selection, we consider the variables both with and without transformation.

In [Figure 5](#), $\ln(\text{PGA})$, $\ln(\text{PGV})$, $\ln(V_{S30})$, *dw*, and precipitation show strong correlations with the probability of liquefaction. We observe that $\ln(\text{PGV})$ is more evenly distributed than $\ln(\text{PGA})$ and $\ln(\text{PGV})$ shows a stronger correlation with the estimated probability of liquefaction from data. The estimated probability of liquefaction seems negatively correlated with $\ln(\text{PGA})$ and $\ln(\text{PGV})$ when PGA is beyond 0.3g and PGV is beyond 50 cm/s. Magnitude, or a function of magnitude, sometimes is considered as a factor to scale PGA (e.g., [Youd et al., 2001](#)). However, with our sampling scheme, the relationship between the probability of liquefaction and magnitude cannot be reliably estimated because the number of points that are sampled for an event is independent from the magnitude (i.e., 1000 liquefaction points if available and 1000 nonliquefaction points). The relationship between the estimated probability of liquefaction and magnitude that appears in [Figure 5](#) appears to be an artifact of our sampling, and cannot be reliably used for prediction. This also appears to be the case for MSF, which is a function of magnitude.

Interaction Assessment

In [Figure 6](#), we assess the interaction effects between variables. We construct a 2D image in which the axes are two of the explanatory variables and the color is the percentage of points that liquefied in each bin (bins are only shown for five or more observations). The bin width is chosen such that each explanatory variable is divided evenly into 19 bins. The observations are compared with the predicted probabilities (black lines) from bivariate models with or without interaction terms, which are represented as probability contour lines on the same image. When this plot is constructed for two explanatory variables without interaction terms, then the probability contour lines are straight. The interaction term allows for the probability contours to curve. Thus, if we add the interaction term and the contour lines are still essentially linear (and unchanged from the model without the interaction

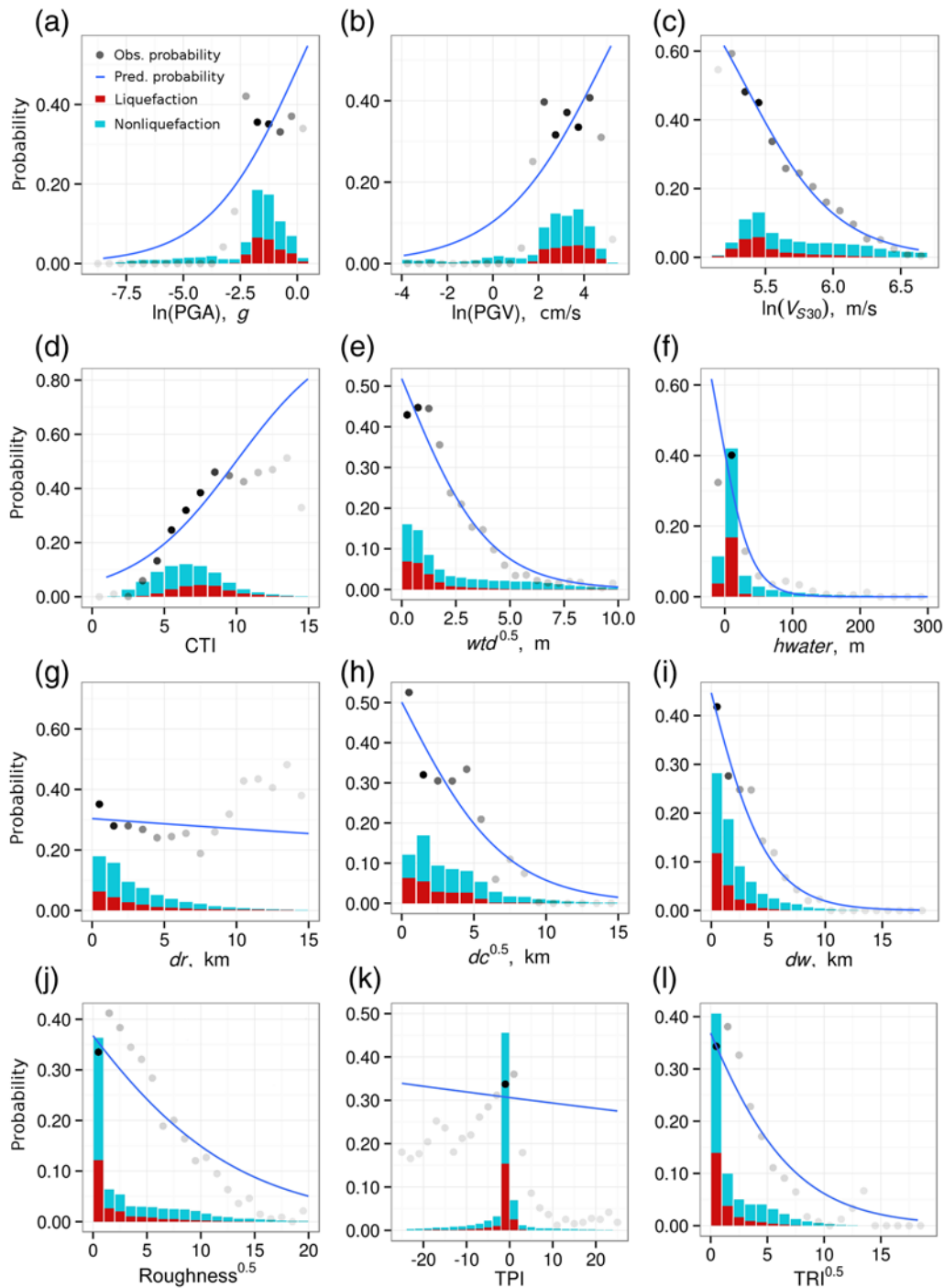


Figure 5. Histograms of liquefaction and nonliquefaction and the probability of liquefaction observed from the data (gray circles) at intervals of a variable's value and the probability predicted from a univariate model. The darkness of the gray dot increases with the number of data points within the bin (the scale is different for each panel). In other words, the darkness of the point is an indication of weight. The color version of this figure is available only in the electronic edition. *(Continued)*

term), then we judge the interaction term to not be significant enough to include in the model.

In this exploration, we only consider multiplicative interaction terms. We use the interaction between dc and dr as an example. When dc is small, dr has little effect on the probability of liquefaction. When dc is large, the probability of liquefaction significantly decreases as dr increases. Note

that dr does not appear to be a good predictor when it is evaluated alone (Fig. 5g), but it becomes valuable when combined with dc . As shown in Figure 6b, after adding the interaction term, the probability contour lines become curved and fit the distribution of data better. This makes sense because we expect both saturation and soil density to change as the river approaches the coast. As a second example, we

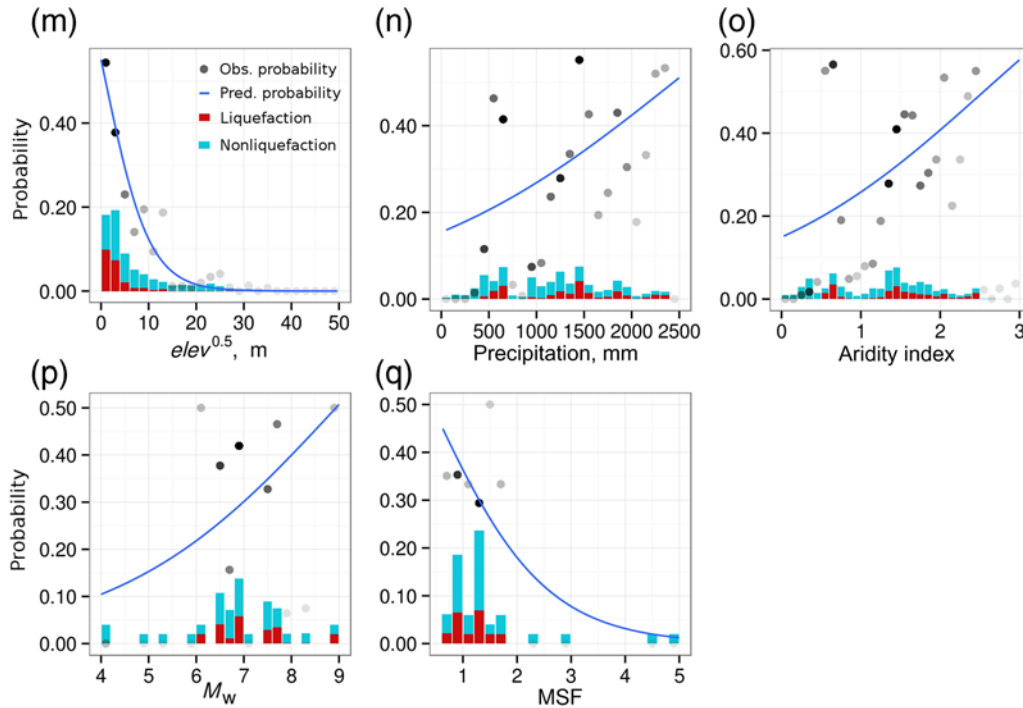


Figure 5. Continued.

expect possible interaction effects between V_{S30} and PGV. When PGV is less than 3 cm/s, the probability of liquefaction is zero, and change in V_{S30} has no effect on the probability of liquefaction. We find adding the $PGV \times V_{S30}$ interaction term does not help because the interaction term does not create curvature in the model contours as shown in Figure 6. Instead, we heuristically assign zero to the predicted probability for both models when $PGV < 3$ cm/s. Similarly, we assign zero to the probability when $V_{S30} > 620$ m/s.

Models

Model Equations. After exploring individual candidate variables and their interaction effects, we select three variables to form a base model, which include $\ln(PGV)$, $\ln(V_{S30})$, and precipitation. We choose these variables because they show strong correlation with the probability of liquefaction (Fig. 5b,c,n) and also can be linked to factors such as earthquake shaking, soil density, and regional climate, which are the primary contributors to the physical process of liquefaction. In addition, the model first relies on a preliminary classification based on PGV and V_{S30} as discussed above. As discussed above, we heuristically assign zero to the predicted probability for both models when $PGV < 3$ cm/s. Similarly, we assign zero to the probability when $V_{S30} > 620$ m/s.

A potential concern with the selection of PGV over the more traditional use of PGA with the MSF correction is that it does not explicitly account for the number of cycles of loading (or duration, which is generally correlated with magni-

tude). Within this context, we would like to note that the saturation of PGV scaling with magnitude is less severe than that for PGA. Thus, one could make the case that because PGV is more sensitive to magnitude than PGA, it indirectly accounts for the additional loading due to the longer durations associated with larger magnitudes. Another concern with the use of PGV is that GMPEs in some regions (especially stable continental regions and subduction zones) rarely include coefficients for evaluating PGV. We still prefer the use of PGV, however, because (1) it performs best when compared with our expanded database, (2) our assessment includes any additional uncertainty in PGV predictions, and (3) inclusion of PGV in GMPEs is becoming relatively standard in modern GMPEs and so we expect this issue to diminish with time.

Another significant contributor to the physical process of liquefaction is soil saturation; however, there were several candidate variables for soil saturation that show good correlation with the probability of liquefaction. Therefore, we focus on assessing the improved performance of the base model when saturation proxies are added as explanatory variables. We assess the performance of the model using the AUC and Brier score calculated from a sampled dataset. We use the same sampled dataset for all three models so that we can directly compare the performance. The coefficients of the best performing model (model 1) are given in Table 3. We also include the coefficients of the global model in [Zhu et al. \(2015\)](#), which uses magnitude-scaled PGA from ShakeMaps ($PGA_{M,SM}$) to account for shaking load. The model performance as summarized by the AUC and Brier scores (shown

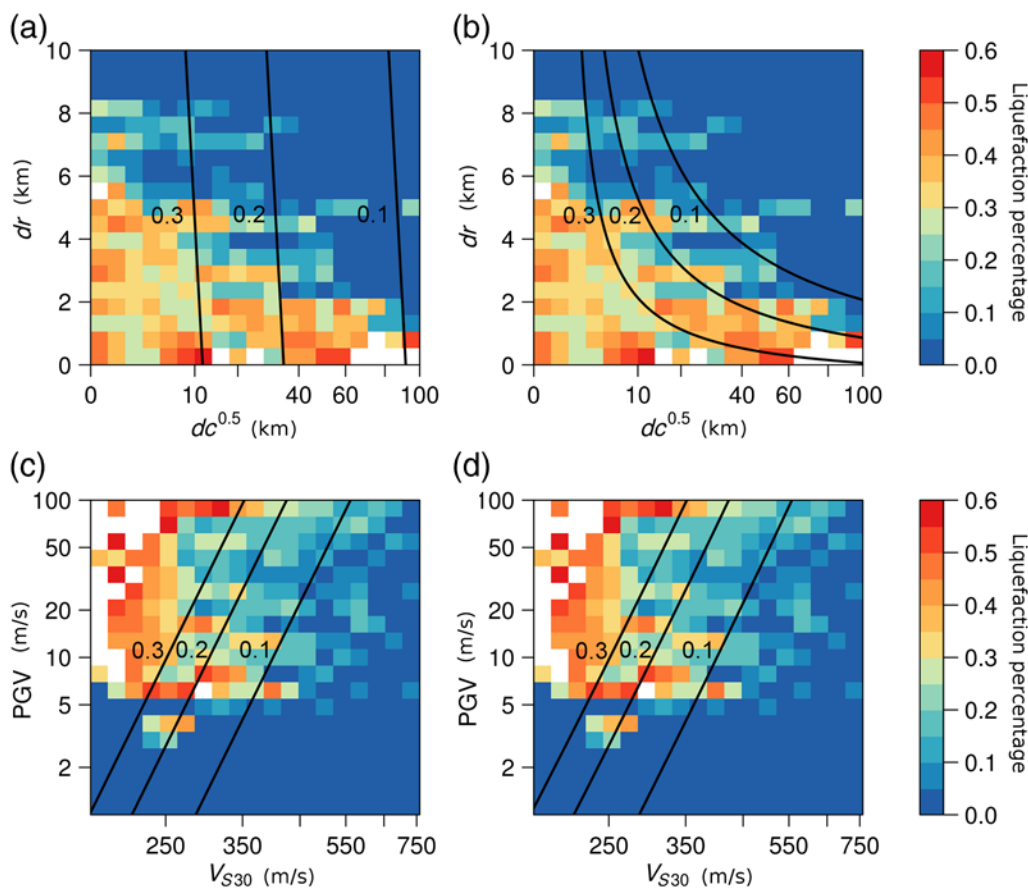


Figure 6. Observed percentage of liquefaction versus nonliquefaction as a function of dr and $dc^{0.5}$ and comparison of probability contours (black lines) from a bivariate model (a) without interaction term and (b) with interaction term. Observed percentage of liquefaction versus nonliquefaction as a function of PGV and V_{S30} , and comparison of probability contours (black lines) from a bivariate model (c) without interaction term and (d) with interaction term. The color version of this figure is available only in the electronic edition.

in Table 3) was comparable across the current dataset: model 1 has the highest AUC (0.801) as compared to model 2 (0.788) and GLM-Zea15g (0.755); model 1 has the lowest Brier score (0.162) as compared to model 2 (0.166) and GLM-Zea15g (0.232).

Goodness of Fit. The liquefaction data in the database are primarily derived from earthquakes that have occurred in coastal environments. There are not many earthquakes with observed liquefaction that occur far from the coast. Although model 1 was developed from a database including both coastal and noncoastal earthquakes, it performs best in the coastal setting. Model 1 relies on the distance to coast parameter as a proxy for saturation and soil density, and we find this proxy can be problematic for the noncoastal setting because as the distance to coast increases, the predicted probabilities approach zero. Therefore, we present a second model (model 2 in Table 3), which is selected from five top-performing models and performs best in noncoastal events (defined earlier): 2008 Wenchuan, 1999 Chi-Chi, 1994 Northridge, and 1999 Hector Mine. Model 2 uses wtd and dw as the saturation proxies.

To compare the models in Table 3 and make sure that they are not biased toward a specific event, we summarize the AUCs of models on individual events in Table 4. Events with no liquefaction are excluded in the table because both liquefaction and nonliquefaction are needed to compute AUC. Both updated models perform well on 16 earthquakes (out of 21 liquefaction earthquakes) with AUC values greater than 0.6. We compare the AUCs of the two updated models, and the bold numbers in the table show the AUC of the better performing model in the two updated models for each event. Although model 1 performs better than model 2 overall, model 2 outperforms model 1 on all noncoastal events. Model GLM-Zea15g outperforms updated models for the 2003 San Simeon, 2004 Niigata, 2003 Tokachi, and 2001 Bhuj events. Both updated models perform poorly (AUC < 0.6) for the 2008 Wenchuan, 2001 Bhuj, and 1994 Northridge events. For the 1965 Puget Sound event, many liquefaction cases with low probabilities from model 1 occurred in the artificially filled areas that were not well captured by the geospatial parameters. In Figure 7, we compare the ROC curve of the updated model with the previous model in Zhu *et al.* (2015). The AUC of the updated model 1 and GLM-Zea15g

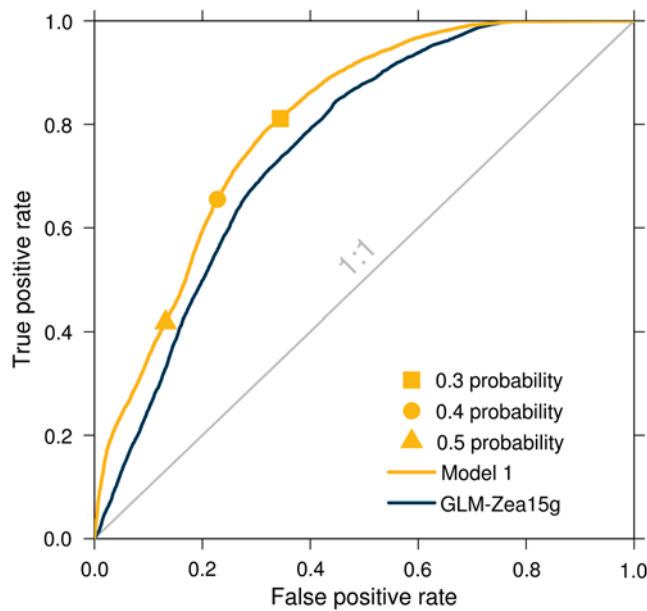


Figure 7. Comparisons of receiver operating characteristic (ROC) curves of GLM-Zea15g and the best-performing model (model 1) with a one-to-one line representing random guessing. The color version of this figure is available only in the electronic edition.

is 0.801 and 0.755, and the Brier score of the updated model 1 and GLM-Zea15g is 0.162 and 0.232, suggesting that the new model provides improved accuracy over the old model.

To convert the model to a classifier, we must select a threshold value to convert from predicted probability to a classification of liquefaction or nonliquefaction. Lower thresholds yield higher TPRs and higher FPRs. The thresh-

old value could be determined based on the highest acceptable FPR. Using the top-performing model (model 1), we present confusion matrices for three thresholds in Table 5. A confusion matrix summarizes statistics for four possible outcomes when comparing a prediction from a binary classifier with an observation: true positive (top left cell; correct positive prediction), true negative (bottom right cell; correct negative prediction), false positive (bottom left cell; incorrect positive prediction), and false negative (top right cell; incorrect negative prediction). For example, using 0.3 as the threshold, 24.2% of data are liquefied and are correctly classified as liquefaction. 5.6% of the data liquefied and are incorrectly classified as nonliquefaction. Using 0.3 as the threshold gives a TPR (the fraction of positive cases that are correctly classified) of 0.81, whereas as a threshold of 0.4 gives a TPR of 0.65 and a threshold of 0.5 reduces the TPR further to 0.42. The FPR (the fraction of negative cases that are incorrectly classified as liquefied) is 0.34 with a threshold of 0.3, 0.23 with a threshold of 0.4, and 0.13 with a threshold of 0.5. These differences in TPR and FPR for different thresholds help to illustrate the effectiveness of the classifier and the meaning of the mapped categories when we use this model to map the probability of liquefaction for an event (Figs. 8 and 9). A threshold of 0.3 is more conservative in that it overpredicts liquefaction, whereas a threshold of 0.4 is a more balanced classifier.

Table 3
Coefficients of Top Performing Coastal Models and GLM-Zea15g

	GLM-Zea15g	Model 1	Model 2	Units
Intercept	24.10	12.435	8.801	
ln(PGV)		0.301	0.334	cm/s
ln(V_{S30})	-4.784	-2.615	-1.918	m/s
precip		5.556×10^{-4}	5.408×10^{-4}	mm
ln(PGA _{M,SM})	2.067			g
\sqrt{dc}		-0.0287		km
dr		0.0666		km
CTI	0.355			
dw			-0.2054	km
wtd			-0.0333	m
$\sqrt{dc} \times dr$		-0.0369		
AUC (all events)*	0.755	0.801	0.788	
Brier score (all events)*	0.232	0.162	0.166	
AUC (noncoastal)	0.655	0.793	0.811	
Brier score (noncoastal)	0.106	0.091	0.104	

We heuristically assign zero to the predicted probability for both models when PGV < 3 cm/s. Similarly, we assign zero to the probability when V_{S30} > 620 m/s. AUC, area under the ROC curve.

*All events in Table 1 except the 2011 Bhuj earthquake.

Table 4

Comparison of the AUCs of Two Updated Models and GLM-Zea15g over Individual Earthquakes

Earthquake	GLM-Zea15g	Model 1	Model 2
1989 Loma Prieta	0.620	0.886	0.848
1995 Kobe	0.808	0.862	0.821
2000 Tottori	0.771	0.806	0.775
2011 Christchurch	0.847	0.801	0.770
1978 Miyagi	0.742	0.791	0.801
2003 San Simeon	0.786	0.757	0.775
2010 Darfield	0.642	0.725	0.692
2001 Nisqually	0.684	0.703	0.751
2004 Niigata	0.728	0.678	0.700
2011 Tohoku	0.532	0.677	0.656
1993 Hokkaido	0.711	0.673	0.720
1964 Niigata	0.668	0.667	0.628
1983 Nihonkai	0.647	0.643	0.649
1987 Chiba	0.613	0.608	0.657
1999 Chi-Chi*	0.568	0.603	0.660
1949 Puget Sound	0.578	0.601	0.610
2003 Tokachi	0.578	0.561	0.571
1965 Puget Sound	0.529	0.558	0.548
1994 Northridge*	0.469	0.535	0.547
2008 Wenchuan*	0.546	0.527	0.551
2001 Bhuj*	0.638	0.515	0.537

The bold numbers indicate the AUC for the better performing model in models 1 and 2. When comparing three models, GLM-Zea15g outperforms the two updated models for six earthquakes.

*Noncoastal events (average distance to coast of liquefaction features > 20 km).

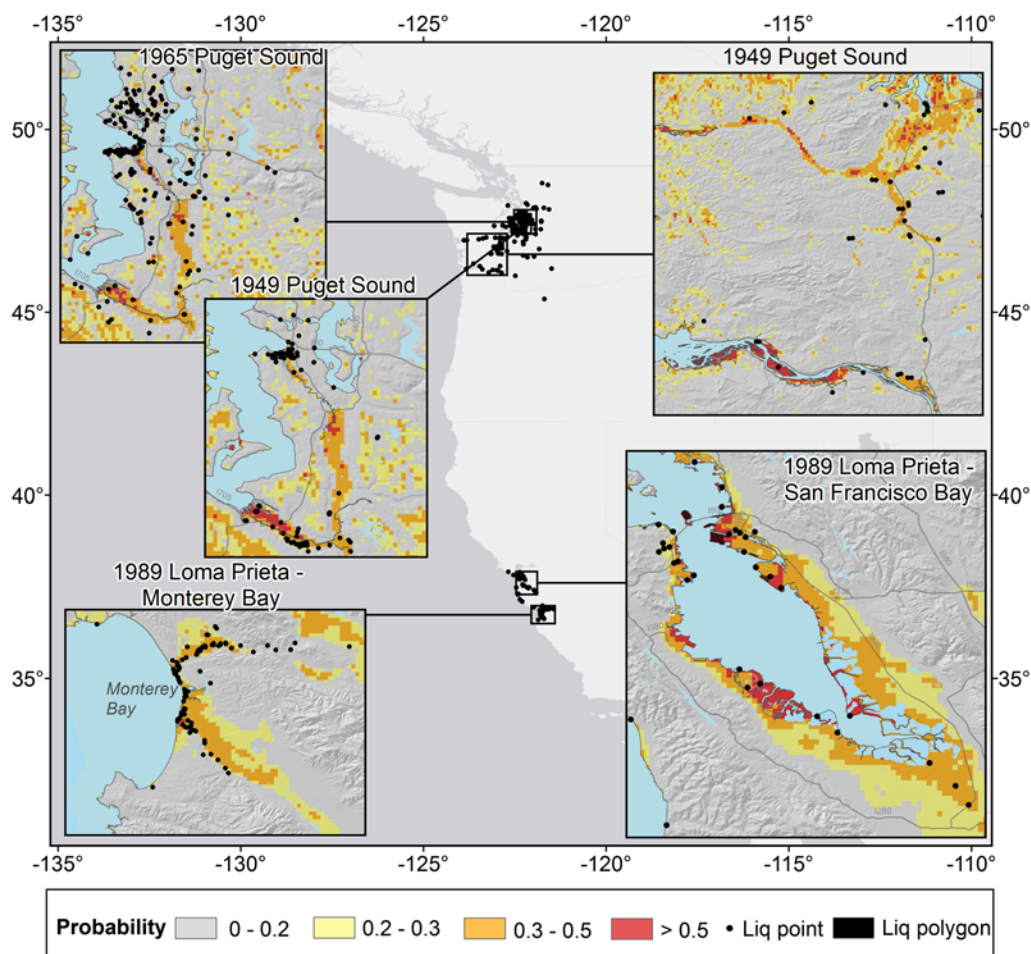


Figure 8. Probability maps predicted from the updated model (model 1) for earthquakes in the United States. The color version of this figure is available only in the electronic edition.

Probability Maps. In addition to the performance metrics, it is also important that the model predicts the spatial distribution of liquefaction for the earthquake event. Figures 8 and 9 show the predicted probability maps (using model 1) for three United States and five Japan earthquakes in which the observed liquefaction points are shown in black. The spatial pattern of liquefaction for each event is well represented by the model. Consistent with the confusion matrix, the categories for probabilities greater than 0.3 show a consistent pattern with the observed liquefaction. We find for earthquakes with very large magnitude such as the 2011 Tohoku earthquake, the model predicts larger areas of high probabilities than the area where liquefaction was observed. This might be related to the fact that the observed probability of liquefaction saturates as a function of $\ln(\text{PGV})$ for large PGV values, but the predicted probability does not (Fig. 5b). The saturation is more severe when using $\ln(\text{PGA})$ as the shaking parameter.

In Figure 10, we show the probability maps calculated using model 2 for the four noncoastal earthquakes in the database (2008 Wenchuan, 1994 Northridge, 1999 Hector Mine, and 1999 Chi-Chi). We also show the model applied

to the 2001 Bhuj and 2015 Nepal earthquakes, which are not in the database as a further verification of the model. The model performs well on the 1999 Hector Mine and the 2015 Nepal events. For the 2008 Wenchuan earthquake, the liquefaction generally occurred near rivers, which coincide with low probability (0.1–0.3) and some medium probability (0.3–0.5) areas predicted by the model. For the 1994 Northridge earthquake, in the epicentral region in the San Fernando Valley, the model appears to overpredict because the liquefaction data in our database for this region are incomplete. Besides the ground failures in the region that are shown in the figure and included in our database, liquefaction was also found to contribute too many liquefaction-related structural failures (Stewart *et al.*, 1994). In the Granada Hills area on the north of the San Fernando Valley where many ground failures were observed, the model predicts low probabilities as a result of the relatively deep water table depth (> 10 m). The observed ground failures in the area might be a result of dynamic ground compaction of loose unsaturated surface material, not liquefaction (Stewart *et al.*, 1994). Outside of the San Fernando Valley, the model

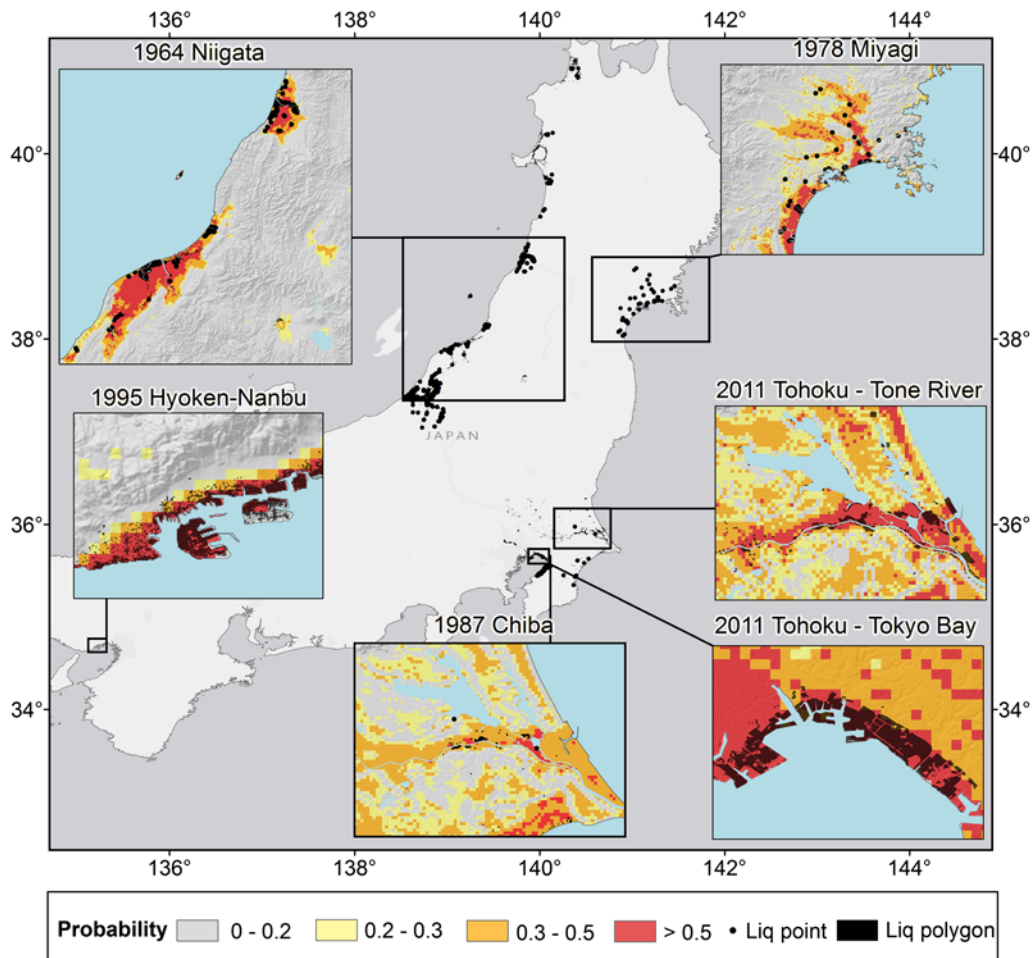


Figure 9. Probability maps predicted from the updated model (model 1) for earthquakes in Japan. The color version of this figure is available only in the electronic edition.

predicts high probabilities in Simi Valley on the west and the coastal area near Marina De Rey on the south, which agree with the liquefaction observations. For the 1999 Chi-Chi event, many observed liquefaction points lie in the area with medium probability. The model overpredicts for coastal areas, where very few liquefaction occurrences were observed. For the 2001 Bhuj event, model prediction in general agrees with the extent of liquefaction estimated from change detection on remote-sensing data. Notwithstanding the above limitations in identifying individual locations of liquefaction observations, it is our interpretation that the aggregate performance of the events in Figure 10 is encouraging for a number of reasons: (1) these are the most challenging events that the model is likely to face, (2) the extent of the observations correlates well with the extent/amplitude of the modeled probabilities (i.e., the model indicates the overall extent of liquefaction for an event even if the exact locations are not identified), and (3) with the exception of prior models developed by our research team, there are currently no feasible alternative models of liquefaction that can be applied globally, and we have shown that this update is a significant improvement over our prior models elsewhere in this article.

Application for Susceptibility. Maps of liquefaction susceptibility which are independent of a specific earthquake scenario may also be useful for regional liquefaction risk estimation. The probability of liquefaction is a function of the set of explanatory variables X in equation (1), which includes the event-specific shaking intensity. To create a susceptibility map, we simply compute X without the intensity (PGV) term. The resulting number is not an estimate of the probability of liquefaction, but simply combines the susceptibility terms together with the coefficients that were determined by our regression analysis. Thus, the absolute values are not directly meaningful. We calculate the susceptibility using model 1, and present the susceptibility as three classes (low, moderate, and high) in Figure 11 for the San Francisco region as compared to a geology-based susceptibility map (Witter *et al.*, 2006). Figure 12 shows a similar comparison between a geospatial susceptibility map versus a geology-based susceptibility map for Seattle (Palmer *et al.*, 2004). Although there are some differences between our model and the geology-based susceptibility maps, we are able to capture similar trends and believe that these geospatial susceptibility maps can be useful as preliminary information for regional-scale planning.

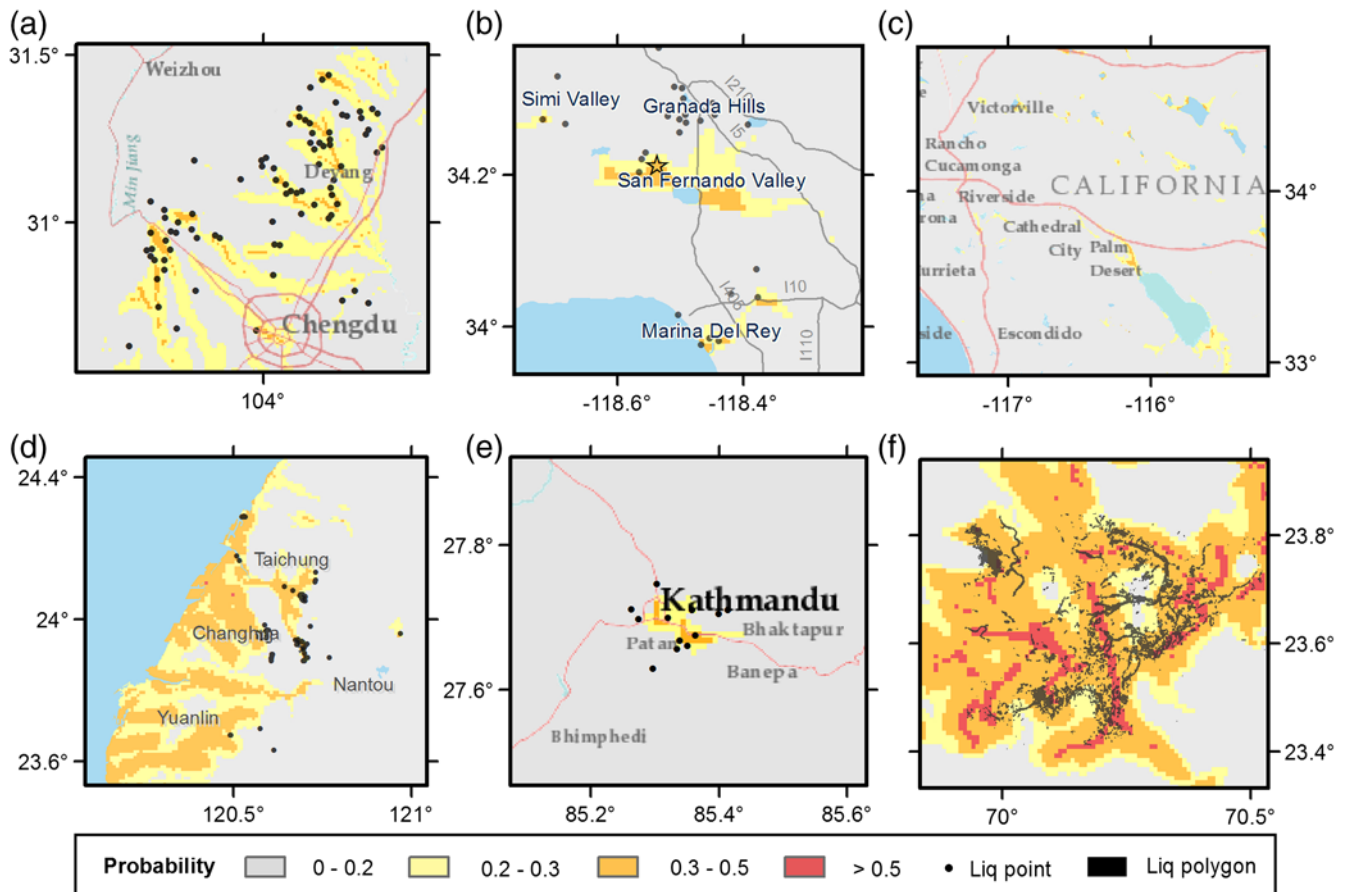


Figure 10. Probability maps predicted from model 2 for the (a) 2008 Wenchuan, (b) 1994 Northridge, (c) 1999 Hector Mine, (d) 1999 Chi-Chi, (e) 2015 Nepal (Moss *et al.*, 2015), and (f) 2001 Bhuj earthquakes. The color version of this figure is available only in the electronic edition.

Discussion

Sensitivity Analyses of Sampling Choices

In this study, we apply a sampling method to combine complete and incomplete datasets into a single database. The resulting method results in a balanced dataset (50:50). To fully understand the implications of our sampling method, we perform sensitivity analyses regarding choices such as the width of the spatial buffer and the class imbalance. In the sensitivity analyses, we use a different dataset than what is described in the [Methods](#) section, which was used for regression. We use a testing dataset that is independent of the sampling method, which consists of all data points from the “complete” datasets as defined in [Table 1](#). For example, 0.5–10 km means we sample nonliquefaction points within the area that is greater than 0.5 km and less than 10 km from observed liquefaction. We develop models using data that are sampled using different buffer widths. We compare the performance of the models in terms of the ROC curve in [Figure 13a](#) and find that the AUCs of the models are not sensitive to the buffer widths. The ROC curves in [Figure 13](#) are different from the ROC curve in [Figure 7](#) because different testing data are used. [Figure 13](#) shows that the model

performs well on spatially complete events that are used for testing. Similarly, we study the sensitivity of model performance to the class imbalance as shown in [Figure 13b](#). We compare the models that are developed using the data sampled using different liquefaction/nonliquefaction ratios, and we find that the AUCs of the models are not sensitive to the class imbalance.

Interpretation of the Predicted Probabilities

The predicted probability from the developed models can be converted to a classification by applying a threshold, as demonstrated with the confusion matrix presented in [Table 5](#). This is useful for predicting liquefaction for a new event. Optimal thresholds can be chosen based on the acceptable false predictions. Another way to interpret the probability is to predict the spatial extent of liquefaction within a probability class. In the development of the [Zhu *et al.* \(2015\)](#) model, which used spatially complete data, the predicted probabilities agreed well with the spatial extent. [Figure 14](#) assesses the relationship between the areal percentage of liquefaction computed with the complete events in the expanded inventory database and the predicted probabilities from GLM-Zea15g and models 1 and 2. For each model, we

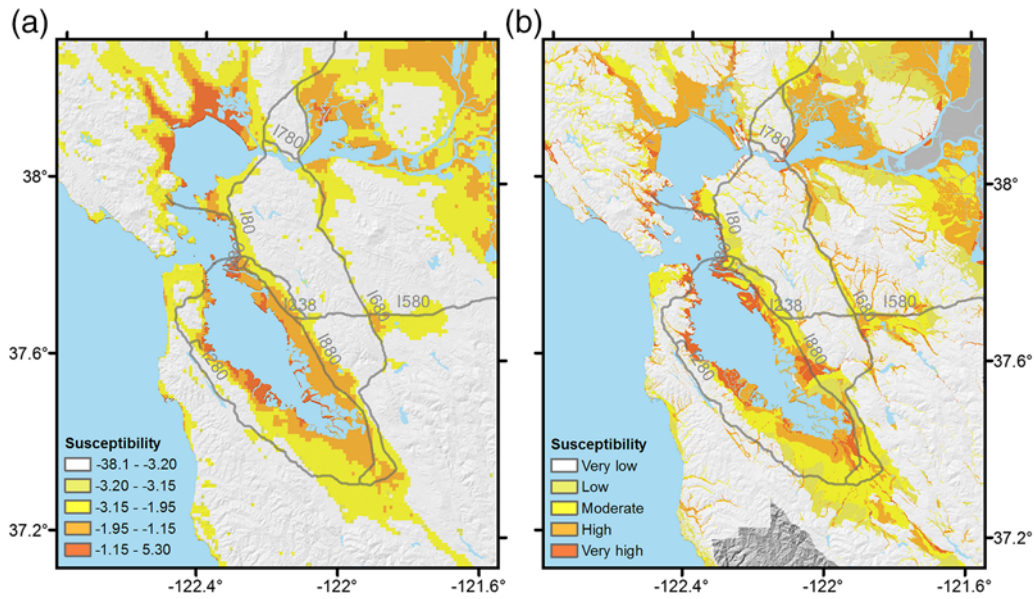


Figure 11. Susceptibility maps for the San Francisco Bay, California, area from (a) the geospatial model (model 1) and (b) geology-based mapping (Witter *et al.*, 2006). The color version of this figure is available only in the electronic edition.

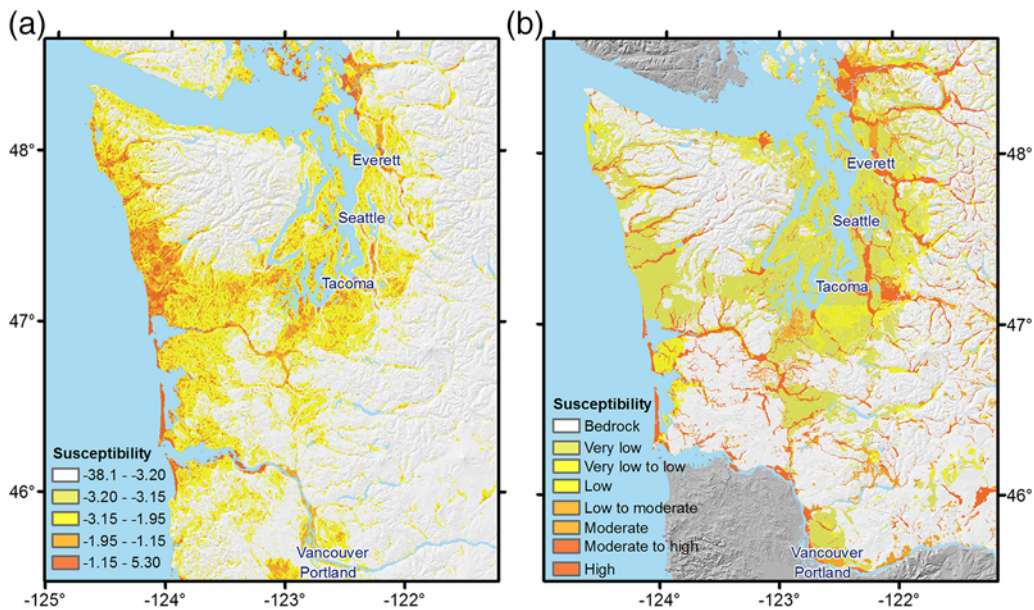


Figure 12. Susceptibility maps for the Seattle, Washington, area from (a) the geospatial model (model 1) and (b) geology-based mapping (Palmer *et al.*, 2004). The color version of this figure is available only in the electronic edition.

bin the predicted probability and compute the liquefaction percent, which is plotted in Figure 14 in the center of each bin, and the 95% confidence interval is illustrated as a vertical line. For GLM-Zea15g, we show a linear model with a 0 intercept and a slope of 0.81. This means that the expanded database indicates that the probabilities from the GLM-Zea15g model should be multiplied by 0.81 to estimate liquefaction percent. We included more events with insignificant or no liquefaction, which explains why the slope of this line is not unity even though that was the target of GLM-

Zea15g. For models 1 and 2, we fit a logistic function which has the same form as equation (1), except that in this case we found that squaring the denominator improves the fit

$$L(P) = \frac{a}{(1 + be^{-cP})^2}, \quad (2)$$

in which L is the areal liquefaction percent, P is the predicted probability, and the parameters a , b , and c are given in Table 6. Equation (2) can either be used to convert the predicted

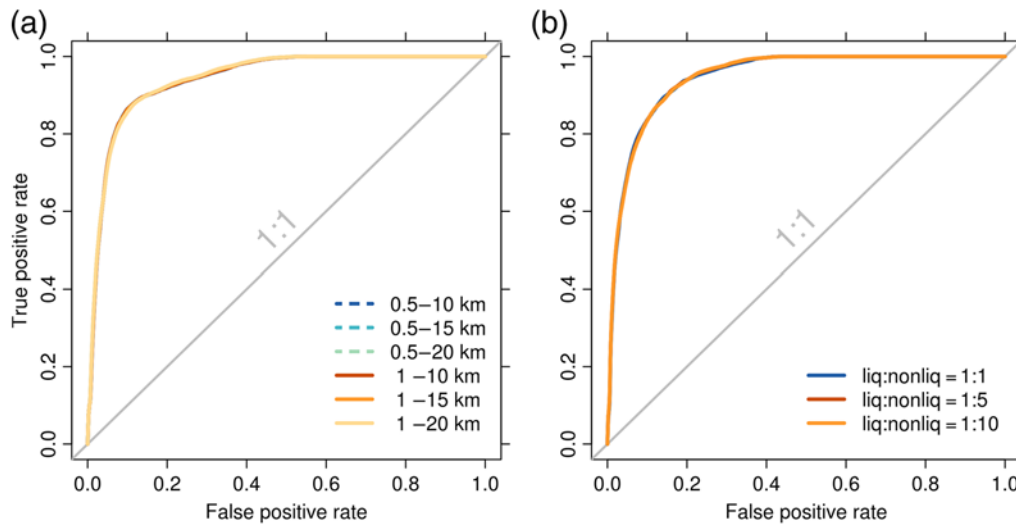


Figure 13. Sensitivity of ROC curves to (a) spatial buffer and (b) class imbalance with one to one lines (1:1) (diagonal lines) representing random guessing. The color version of this figure is available only in the electronic edition.

probability to liquefaction percent or to define simplified classes. For example, to define a class in which the percent liquefaction is between 10% and 20% from the probabilities predicted by model 2, one would insert the value of 10 and 20 for $L(P)$ into equation (2) and solve for P with the model 2 coefficients from Table 6, which would yield probabilities of 0.37–0.47.

Applicability of Model to Noncoastal Regions

Although the goal of the geospatial liquefaction model presented here is for global use, we acknowledge that the model development was based on earthquakes in coastal regions. Our analysis of model 2 in noncoastal regions provides promising results; however, uniform performance across all tectonic environments should not be expected. For example, the accuracy of predicting V_{S30} from slope may be less accurate in glaciated regions (Magistrale *et al.*, 2012). Model validation and development should continue as earthquakes occur and additional data become available.

Table 5

Confusion Matrices for Three Thresholds (0.3, 0.4, and 0.5) for Model 1 Presented in Table 4

Threshold Predicted		$P = 0.3$		$P = 0.4$		$P = 0.5$	
		L	NL	L	NL	L	NL
Actual	L	24.2	5.6	19.6	10.3	12.5	17.4
	NL	24.1	45.9	16.0	54.0	9.2	60.8
		TPR = 0.81		TPR = 0.65		TPR = 0.42	
		FPR = 0.34		FPR = 0.23		FPR = 0.13	

L, liquefaction; NL, nonliquefaction; TPR, true positive rate; FPR, false positive rate.

Conclusions

To predict liquefaction extent immediately after an earthquake worldwide, we need a model that uses widely available geospatial parameters (e.g., Zhu *et al.*, 2015). In this article, we update the Zhu *et al.* (2015) model by (1) expanding the database to include 27 events from six countries, (2) applying a sampling method to add incomplete datasets, (3) evaluating new explanatory variables, and (4) testing interaction terms.

In model development, we compare 18 proxies for earthquake shaking, soil saturation, and soil density. We find PGV performs better than PGA as a shaking parameter. The patterns of saturation proxies show different scales of details. At a regional scale, distance to the water body performs best. We find that considering interaction terms between dr and dc improves the accuracy of the model. The model that performs best over the entire dataset includes PGV, V_{S30} , dr , dc , and precipitation. The model that performs best over the noncoastal dataset includes PGV, V_{S30} , wtd , dw , and precipitation. The updated models offer an improved accuracy as compared to the Zhu *et al.* (2015) model. We validate the models and assess the resulting probability in terms of probability thresholds and the spatial extent of liquefaction. We find that the mapped probability of liquefaction can be used as an estimate

Table 6

Parameters for Relating Model Probabilities to Areal Liquefaction Percent

Parameters	Model 1	Model 2
a	42.08	49.15
b	62.59	42.40
c	11.43	9.165

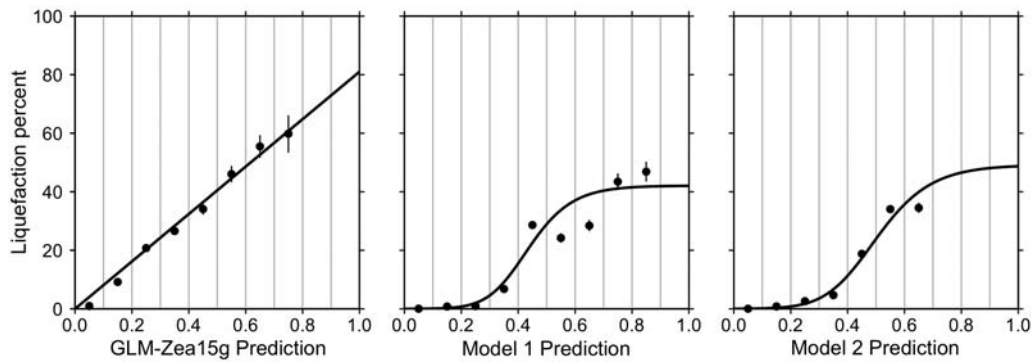


Figure 14. Areal percentage of liquefaction as a function of model probabilities. In each panel, the dots show the estimated liquefaction percent with 95% confidence intervals for binned predicted probabilities. Vertical lines indicate the bin boundaries. The black lines are curves fit to the data.

of spatial extent within classes, but should be adjusted due to the 50:50 class balance used herein. Overall, the footprint and overall degree of liquefaction is successfully recovered for test events to a degree that indicates our models should prove useful for global near-real-time applications.

Data and Resources

The liquefaction data used in this article were all compiled and digitized from published sources (listed in the references in Table 1) except the data for the 2010–2011 Darfield and Christchurch earthquakes from the Canterbury geotechnical database (<https://canterburygeotechnicaldatabase.projectorbit.com>, last accessed July 2014). The liquefaction data that we digitized for 10 earthquakes in the United States, Japan, China, and Taiwan are available from Zhu *et al.* (2016). The electronic data from all of the events in Table 1 in which the reference is Wakamatsu (2011) are available in the CD that accompanies the book. The ShakeMaps were obtained from the U.S. Geological Survey earthquake archives (<http://earthquake.usgs.gov/earthquakes/search/>, last accessed April 2015). The digital elevation model was obtained from the Global Multi-resolution Terrain Elevation Data 2010 (http://topotools.cr.usgs.gov/gmted_viewer/viewer.htm, last accessed December 2013). River networks and compound topographic index data were obtained from the Hydrological data and maps based on Shuttle Elevation Derivatives at multiple Scales (HydroSHEDS) database (<http://hydrosheds.cr.usgs.gov/dataavail.php>, last accessed February 2014). Distance to the nearest coastline data was computed from the Distance to the Nearest Coast dataset (<https://oceancolor.gsfc.nasa.gov/docs/distfromcoast/>, last accessed January 2014). Mean annual precipitation data were obtained from the WorldClim database (<http://WorldClim.org>, last accessed March 2014). The aridity index data were obtained from the Global Aridity and PET dataset (<http://www.cgiar-csi.org/data/global-aridity-and-pet-database>, last accessed September 2014). Analyses on geospatial datasets were performed using the Generic Mapping Tools software (Wessel and Smith, 1998), available at <http://www.soest.hawaii.edu/gmt/> (last accessed October 2014); and the Geospatial Data Abstraction Library, available at <http://www.gdal.org/> (last accessed September 2015). All other computations in this article were completed with the open-source software R (R Development Core Team, 2016) available at <http://www.r-project.org/> (last accessed April 2016). Figures were prepared using R and Geographic Information System program ArcGIS, v. 10.

www.soest.hawaii.edu/gmt/ (last accessed October 2014); and the Geospatial Data Abstraction Library, available at <http://www.gdal.org/> (last accessed September 2015). All other computations in this article were completed with the open-source software R (R Development Core Team, 2016) available at <http://www.r-project.org/> (last accessed April 2016). Figures were prepared using R and Geographic Information System program ArcGIS, v. 10.

Acknowledgments

This work was supported under National Science Foundation Grant Number 1300781 and the U.S. Geologic Survey (USGS), Department of the Interior, under USGS Award Number G16AP00014; we gratefully acknowledge this support. The work has greatly benefited from discussions with Ken Campbell and Keith Knudsen. This work benefited from reviews by Kate Allstadt, David Wald, Janet Slate, and two anonymous reviewers. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

- Baise, L. G., D. Daley, J. Zhu, E. M. Thompson, and K. Knudsen (2012). Geospatial liquefaction hazard model for Kobe, Japan and Christchurch, New Zealand. *Seismol. Res. Lett.* **83**, 458.
- Beven, K. J., and M. J. Kirkby (1979). A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* **24**, 43–69.
- Boulanger, R., and I. Idriss (2015). CPT-based liquefaction triggering procedure. *J. Geotech. Geoenviron. Eng.* **142**, no. 2, doi: [10.1061/\(ASCE\)GT.1943-5606.0001388](https://doi.org/10.1061/(ASCE)GT.1943-5606.0001388).
- Brankman, C. M., and L. G. Baise (2008). Liquefaction susceptibility mapping in Boston, Massachusetts. *Environ. Eng. Geosci.* **14**, 1–16.
- Bray, J. D., R. Sancio, A. M. Kammerer, S. Merry, A. Rodriguez-Marek, B. Khazai, S. Chang, A. Bastani, B. Collins, E. Hausler, *et al.* (2001). *Some Observations of the Geotechnical Aspects of the February 28, 2001, Nisqually Earthquake in Olympia, South Seattle, and Tacoma, Washington*, Pacific Earthquake Engineering Research Center, University of California, Berkeley, available at <http://peer.berkeley.edu/publications/nisqually/geotech/> (last accessed March 2017).
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3.

- Cao, Z., L. Hou, H. Xu, and X. Yuan (2010). Distribution and characteristics of gravelly soil liquefaction in the Wenchuan M_s 8.0 earthquake, *Earthq. Eng. Eng. Vib.* **9**, 167–175.
- Cetin, K. O., R. B. Seed, A. Der Kiureghian, K. Tokimatsu, L. F. Harder Jr., R. E. Kayen, and R. E. Moss (2004). Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential, *J. Geotech. Geoenviron. Eng.* **130**, 1314–1340.
- Chleborad, A. F., and R. L. Schuster (1998). Ground failure associated with the Puget Sound region earthquakes of April 13, 1949, and April 29, 1965, in *Assessing Earthquake Hazards and Reducing Risks in the Pacific Northwest*, U.S. Geol. Surv. Profess. Pap. 1560.
- Chu, D. B., J. P. Stewart, S. Lee, J. S. Tsai, P. S. Lin, B. L. Chu, R. B. Seed, S. C. Hsu, M. S. Yu, and M. C. Wang (2004). Documentation of soil conditions at liquefaction and non-liquefaction sites from 1999 Chi-Chi (Taiwan) earthquake, *Soil Dynam. Earthq. Eng.* **24**, 647–657.
- Estabrooks, A., J. Tacho, and N. Japkowicz (2004). A multiple resampling method for learning from imbalanced data sets, *Comput. Int.* **20**, 18–36.
- Fan, Y., H. Li, and G. Miguez-Macho (2013). Global patterns of groundwater table depth, *Science* **339**, 940–943.
- Geo-Engineering Extreme Events Reconnaissance (GEER) Association (2008). Preliminary Geotechnical Observations of the July 29, 2008 Southern California Earthquake, available at http://www.geerassociation.org/images/GEER_Activities/07-29-2008%20LA%20EQ/ChinoHillsEarthquakeFinal.pdf (last accessed March 2017).
- Geo-Engineering Extreme Events Reconnaissance (GEER) Association (2014). Geotechnical Engineering Reconnaissance of the August 24, 2014 M 6 South Napa Earthquake, *Rept. No. GEER-037*.
- Hamada, M., R. Ioyama, and K. Wakamatsu (1995). *The 1995 Hyogoken-Nanbu (Kobe) earthquake: Liquefaction, ground displacement and soil condition in the Hanshin Area*, Association for the Development of Earthquake Prediction, Tokyo, Japan, 194 pp.
- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, in Springer Series in Statistics, Springer-Verlag, New York, New York.
- Holzer, T. L., M. J. Bennett, T. E. Noce, A. C. Padovani, and J. C. Tinsley III (2006). Liquefaction hazard mapping with LPI in the greater Oakland, California, area, *Earthq. Spectra* **22**, 693–708.
- Holzer, T. L., T. E. Noce, and M. J. Bennett (2009). Scenario liquefaction hazard maps of Santa Clara Valley, northern California, *Bull. Seismol. Soc. Am.* **99**, 367–381.
- Holzer, T. L., T. E. Noce, M. J. Bennett, J. C. Tinsley, and L. I. Rosenberg (2005). Liquefaction at Oceano, California, during the 2003 San Simeon earthquake, *Bull. Seismol. Soc. Am.* **95**, 2396–2411.
- Kayen, R., R. E. S. Moss, E. M. Thompson, R. B. Seed, K. O. Cetin, A. D. Kiureghian, Y. Tanaka, and K. Tokimatsu (2013). Shear-wave velocity-based probabilistic and deterministic assessment of seismic soil liquefaction potential, *J. Geotech. Geoenviron. Eng.* **139**, 407–419.
- Kleinbaum, D. G., and M. Klein (2010). *Logistic Regression: A Self-Learning Text*, Springer Science & Business Media, New York, New York.
- Knudsen, K., and J. Bott (2011). Geologic and geomorphic evaluation of liquefaction case histories for rapid hazard mapping, *Seismol. Res. Lett.* **82**, 334.
- Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution, *Conference on Artificial Intelligence in Medicine in Europe*, Springer, Heidelberg, Germany, 63–66.
- Lehner, B., K. Verdin, and A. Jarvis (2006). *HydroSHEDS Technical Documentation*, World Wildlife Fund US, Washington, D.C.
- Magistrale, H., Y. Rong, W. Silva, and E. Thompson (2012). A site response map of the continental U.S., *Proc. of the Fifteenth World Conference on Earthquake Engineering*, International Association for Earthquake Engineering, Lisbon, Portugal, 24–28 September 2012.
- Matsuoka, M., K. Wakamatsu, M. Hashimoto, S. Senna, and S. Midorikawa (2015). Evaluation of liquefaction potential for large areas based on geomorphologic classification, *Earthq. Spectra* **31**, 2375–2395.
- Maurer, B. W., R. A. Green, M. Cubrinovski, and B. A. Bradley (2015). Assessment of CPT-based methods for liquefaction evaluation in a liquefaction potential index framework, *Geotechnique* **65**, 328–336.
- Ministry of Land, Infrastructure, Transport and Tourism (MLITT) of Kanto Region Development Bureau and Japanese Geotechnical Society (2011). Investigation of liquefaction phenomena of Kanto region in Grate East Japan Earthquake, available at <http://www.ktr.mlitt.go.jp/bousai/bousai00000061.html> (last accessed April 2014) (in Japanese).
- Moehle, J. P., J. Bray, G. Fenves, F. Filippou, S. Mahin, T. McEvilly, M. Riemer, B. Romanowicz, R. Seed, N. Sitar, et al. (1994). Preliminary report on the seismological and engineering aspects of the January 17, 1994 Northridge Earthquake, *Rept. UCB/EERC-94-01*, Earthquake Engineering Research Center, Berkeley, California, 72 pp.
- Moss, R. E., E. M. Thompson, D. S. Kieffer, B. Tiwari, Y. M. Hashash, I. Acharya, B. R. Adhikari, D. Asimaki, K. B. Clahan, B. D. Collins, et al. (2015). Geotechnical effects of the 2015 magnitude 7.8 Gorkha, Nepal, earthquake and aftershocks, *Seismol. Res. Lett.* **86**, 1514–1523.
- Oommen, T., L. G. Baise, and R. M. Vogel (2010). Validation and application of empirical liquefaction models, *J. Geotech. Geoenviron. Eng.* **136**, 1618–1633.
- Oommen, T., L. G. Baise, and R. M. Vogel (2011). Sampling bias and class imbalance in maximum likelihood logistic regression, *Math. Geosci.* **43**, 99–120.
- Palmer, S. P., S. L. Magsino, E. L. Bilderback, J. L. Poelstra, D. S. Folger, and R. A. Niggemann (2004). Liquefaction susceptibility and site class maps of Washington State, by county, *Washington Division of Geology and Earth Resources, U.S. Geol. Surv. Open-File Rept. 2004-20*.
- R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Singh, R. P., S. Bhoi, and A. K. Sahoo (2002). Changes observed in land and ocean after Gujarat earthquake of 26 January 2001 using IRS data, *Int. J. Rem. Sens.* **23**, 3123–3128.
- Stewart, J. P., J. D. Bray, R. B. Seed, and N. Sitar (Editors) (1994). Preliminary report on the principal geotechnical aspects of the January 17, 1994 Northridge earthquake, *Rept. UCB/EERC-94/08*, Earthquake Engineering Research Center, University of California.
- Stewart, J. P., R. B. Seed, and J. D. Bray (1996). Incidents of ground failure from the 1994 Northridge earthquake, *Bull. Seismol. Soc. Am.* **86**, no. 1B, S300–S318.
- Thompson, E. M., D. J. Wald, K. Allstadt, and M. Hearne (2016). Combining case history observations with different completeness levels in empirical ground-failure models, *Seismol. Res. Lett.* **87**, no. 2B, 578.
- Tinsley, J. C., III, J. A. Egan, R. E. Kayen, M. J. Bennett, A. Kropp, and T. L. Holzer (1998). Appendix: Maps and descriptions of liquefaction and associated effects, in *The Loma Prieta, California, Earthquake of October 17, 1989—Liquefaction*, T. L. Holzer (Editor), U.S. Geol. Surv. Profess. Pap. 1151-B.
- Van Den Eeckhaut, M., J. Hervás, C. Jaedicke, J. P. Malet, L. Montanarella, and F. Nadim (2012). Statistical modelling of Europe-wide landslide susceptibility using limited landslide inventory data, *Landslides* **9**, no. 3, 357–369.
- Wakamatsu, K. (2011). *Historic Liquefaction Sites in Japan*, University of Tokyo Press, Tokyo, Japan, 745–2008.
- Wald, D. J., and T. I. Allen (2007). Topographic slope as a proxy for seismic site conditions and amplification, *Bull. Seismol. Soc. Am.* **97**, 1379–1395.
- Weiss, G., and F. Provost (2001). The effect of class distribution on classifier learning: An empirical study, Department of Computer Science, Rutgers University, *Tech. Rept. ML-TR-44*.
- Wessel, P., and W. H. F. Smith (1998). New, improved version of Generic Mapping Tools released, *Eos Trans. AGU* **79**, 579.
- Wilson, M. F., B. O’Connell, C. Brown, J. C. Guinan, and A. J. Grehan (2007). Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope, *Mar. Geod.* **30**, 3–35.
- Witter, R. C., K. L. Knudsen, J. M. Sowers, C. M. Wentworth, R. D. Koehler, C. E. Randolph, S. K. Brooks, and K. D. Gans (2006). Maps of quaternary deposits and liquefaction susceptibility in the central San Francisco Bay

- region, California, *U.S. Geol. Surv. Open-File Rept. 06-1037*, available at <http://pubs.usgs.gov/of/2006/1037/> (last accessed April 2014).
- Worden, C. B., D. J. Wald, T. I. Allen, K. Lin, D. Garcia, and G. Cua (2010). A revised ground-motion and intensity interpolation scheme for ShakeMap, *Bull. Seismol. Soc. Am.* **100**, 3083–3096.
- Youd, T. L., and D. M. Perkins (1978). Mapping liquefaction-induced ground failure potential, *J. Geotech. Eng. Div.* **104**, no. 4, Paper Number 13659, 433–446.
- Youd, T. L., I. M. Idriss, R. D. Andrus, I. Arango, G. Castro, J. T. Christian, R. Dobry, W. D. L. Finn, L. F. Harder, M. E. Hynes, *et al.* (2001). Liquefaction resistance of soils: Summary report from the 1996 NCEER and 1998 NCEER/NSF workshops on evaluation of liquefaction resistance of soils, *J. Geotech. Geoenviron. Eng.* **127**, 817–833.
- Zhu, J., L. G. Baise, and E. M. Thompson (2016). Liquefaction observations from ten earthquakes in the US, Japan, China, and Taiwan, *U.S. Geol. Surv. Data Release*, doi: [10.5066/F7NC5ZB8](https://doi.org/10.5066/F7NC5ZB8).
- Zhu, J., L. G. Baise, E. M. Thompson, and H. Magistrale (2014). Testing national and regional geospatial liquefaction models in the United States, *10th U.S. National Conference on Earthquake Engineering*, Anchorage, Alaska, 21–25 July 2014.
- Zhu, J., L. G. Baise, E. M. Thompson, D. J. Wald, and K. L. Knudsen (2013). A geospatial liquefaction model for rapid response and loss estimation, *Seismol. Res. Lett.* **84**, 313.
- Zhu, J., D. Daley, L. G. Baise, E. M. Thompson, D. J. Wald, and K. L. Knudsen (2015). A geospatial liquefaction model for rapid response and loss estimation, *Earthq. Spectra* **31**, 1813–1837.

Department of Civil and Environmental Engineering
Tufts University
113 Anderson Hall
200 College Avenue
Medford, Massachusetts 02155
Jing.zhu@tufts.edu
Laurie.baise@tufts.edu
(J.Z., L.G.B.)

U.S. Geological Survey
1711 Illinois Street
Golden, Colorado 80401
emthompson@usgs.gov
(E.M.T.)

Manuscript received 22 June 2016;
Published Online 2 May 2017