An AI Driven, Mechanistically Grounded Geospatial Liquefaction Model for Rapid Response and Scenario Planning

Geyin, M.[1], Maurer, B.W.[2], and Christofferson, K.[3]

Abstract: Geospatial models for predicting soil liquefaction infer subsurface traits via satellite remote sensing and mapped information, rather than directly measure them with subsurface tests. Field tests of such models have demonstrated both promising potential and severe shortcomings. Informed by these tests, this paper develops geospatial models that are driven by algorithmic learning but pinned to a physical framework, thereby benefiting both from machine and deep learning, or ML/DL, and the knowledge of liquefaction mechanics developed over the last 50 years. With this approach, subsurface cone penetration test (CPT) measurements are predicted remotely within the framing of a popular CPT model for predicting ground failure. This has three potential advantages: (i) a mechanistic underpinning; (ii) a significantly larger training set, with the model principally trained on in-situ test data, rather than on ground failures; and (iii) insights from ML/DL, with greater potential for geospatial data to be exploited. While liquefaction is a phenomenon best predicted by mechanics, subsurface traits lack theoretical links to above-ground parameters, but correlate in complex, interconnected ways - a prime problem for ML/DL. Preliminary models are trained using ML/DL and a modest U.S. dataset of CPTs to predict liquefaction-potential-index values via 12 geospatial variables. The models are tested on recent earthquakes and are shown – to a statistically significant degree – to perform as well as, or better than, the current leading geospatial model. The models are coded in free, simple-to-use Windows software. The only input is a ground-motion raster, downloadable minutes after an earthquake or available for countless future scenarios. Ultimately, the proposed approach and models, which warrant further application and evaluation, could be improved upon using additional training data and new predictor variables. Users of the models should understand key limitations, as discussed in detail herein.

Keywords: soil liquefaction; geospatial modeling; artificial intelligence; software

1. Introduction

As evidenced by earthquakes occurring globally each year, reliable predictions of soil liquefaction are needed both prior to an earthquake for efficient planning and mitigation, and immediately after an earthquake for informing response, reconnaissance, and recovery. Such predictions would thus ideally have the capability of being made: (i) rapidly (e.g., in near-real-time after an event); (ii) at high resolution (e.g., consistent with

[1] Norwegian Geotechnical Institute, Houston, TX USA
[2] Department of Civil and Environmental Engineering, University of Washington, Seattle WA, USA.
[3] Department of Computer Science, University of Toronto, Toronto ON, Canada.

the scale of individual assets); and (iii) over the regional extents impacted by large earthquakes (e.g., that of a metropolis or transportation system). Problematically, state-of-practice liquefaction models require relatively costly geotechnical data, such as that from the cone-penetration-test (CPT). Given the infeasibility of in-situ testing across vast areas, regional-scale predictions of liquefaction have traditionally relied on geologic maps, from which generic areal classifications of liquefaction susceptibility may be assumed (e.g., Youd and Hoose, 1977). Such an approach is simple, but also unacceptably uncertain for most intents. Alternatively, statistical distributions of geotechnical data within geologic units may be developed and used to predict liquefaction (e.g., Holzer et al., 2011). This is more data-driven and likely to be more accurate, but many in-situ tests are still required within each mapped unit of interest. Moreover, predictions using this approach have the coarse resolution of geologic maps and assume intra-unit uniformity (i.e., local conditions are not considered), meaning an entire city could potentially receive the same prediction. It thus remains a persistent challenge to model liquefaction (or any geohazard) in a manner that is regional and rapid, yet high resolution and accurate. The existence of a model having these traits is conceivable, however, given the growth of community geotechnical datasets, remote sensing, and algorithmic learning (i.e., machine and deep learning, or ML/ML).

Towards that end, interest has grown in prediction models that use inputs readily available from satellite remote-sensing and existing mapped information. In contrast to geotechnical methods, "geospatial" models can predict liquefaction rapidly, at infinitely many locations. This is made possible using geospatial proxies of soil properties relevant to liquefaction (i.e., above-ground inferences of below-ground conditions). While the concept of such a model is not new (e.g., Kramer, 2008; FEMA, 2013), the model of Zhu et al. (2017), lightly modified by Rashidian and Baise (2020), is arguably the most rigorously formulated and widely accepted. It is also implemented in the United States Geological Survey (USGS) "PAGER" system, which provides content on possible earthquake impacts (Wald et al., 2008). In a recent study, Geyin et al. (2020) tested the Zhu et al. (2017) model against 18 CPT-based models using ~15,000 liquefaction case histories (essentially all CPT-based case histories globally available to date). These analyses elucidated both the promising potential of geospatial data, as well as significant room, and potential means, for improving existing geospatial models.

Informed by these analyses, this paper aims to develop an improved geospatial model driven by algorithmic learning (benefiting from ML/DL insights) but pinned to a physical framework (benefiting from mechanics and the knowledge of regression modelers). In the following, the typology of liquefaction models is succinctly summarized (to place this paper, and the methods it will utilize, in context). Next, tests of the Zhu et al. (2017) model are summarized and lessons for improvement are discussed, as are the potential advantages of the proposed approach. This approach is then used to develop ML/DL models that predict the probability of liquefaction-induced ground failure. Finally, these models are tested using unbiased data and implemented in *RapidLiq*, a new Windows software program.

1.1 A Succinct Overview of Geotechnical and Geospatial Liquefaction Models

The typology of models for predicting liquefaction roughly consists of three tiers: (Tier 1) fully-empirical models that require only geospatial or geologic information (e.g., Rashidian and Baise, 2020); (Tier 2) semi-mechanistic "stress-based" models that require in-situ test data and are widely used in engineering practice (e.g. Boulanger and Idriss, 2014; Green et al., 2019); and (Tier 3) numerical constitutive models, which require many material and model parameters (e.g., Cubrinovski and Ishihara, 1998; Ziotopoulou and Boulanger, 2016). While improvements to computational throughput have grown the use of "Tier 3" models, their application is still limited to specific sites and special projects, given the required inputs and operator skill. Given the rapid and regional scale aims of the proposed work, "Tier 3" models will not be used herein, which is not to say that such models could not conceivably be implemented at regional scale.

Many "Tier 2" models are popular in engineering practice. These include, among others, Robertson and Wride (1998), Moss et al. (2006), Idriss and Boulanger (2008), Kayen et al. (2013), and Green et al. (2019), which all use subsurface measurements to predict liquefaction as a function of earthquake magnitude (M_w) and peak ground acceleration (PGA). However, because these models predict the factor of safety against liquefaction "triggering" (FS_{liq}) at-depth within a profile, the outputs are often used cooperatively with other models that predict manifestations of liquefaction at the surface (i.e., "ground failure"). One popular manifestation model is the liquefaction potential index (LPI) proposed by Iwasaki et al. (1978):

$$LPI = \int_0^{20\,m} F(FS_{liq}) \cdot w(z)\,\mathrm{d}z \tag{1}$$

where $F(FS_{liq})$ and $w(z)$ weight the respective influences of FS_{liq} and depth, z, on surface manifestation. Specifically, $F(FS_{liq}) = 1 - FS_{liq}$ for $FS_{liq} \leq 1$ and $F(FS_{liq}) = 0$ otherwise; $w(z) = 10 - 0.5z$. LPI thus assumes that surface manifestation depends on the thickness of all liquefied strata in the upper 20 m, the degree to which FS_{liq} in each stratum is less than 1.0, and how near those strata are to the surface. LPI can range from zero to 100, with surface manifestations becoming more likely as LPI increases (e.g., Maurer et al., 2014; Geyin and Maurer, 2020a). Other similar manifestation models include those of van Ballegooy et al. (2014) and Maurer et al. (2015a).

"Tier 1" geospatial models, which aim to predict liquefaction via readily available predictor variables, have recently received renewed attention. Like "Tier 2" models, geospatial models characterize liquefaction demand via ground-motion intensity measures (IMs). But, instead of quantifying liquefaction resistance with in-situ measurements, geospatial models predict below-ground conditions using above-ground information. Examples of such predictors include (among many) the slope and roughness of the surface; the distance to rivers and coasts; and compound-topographic-index, which can be derived from satellite data or existing prediction maps. Geospatial models are well suited for regional scale applications such as: (i) loss estimation and disaster simulation; (ii) city planning and policy development; (iii) emergency response; and (iv) post-

event reconnaissance (e.g., to remotely identify sites of interest). Given these many uses, geospatial models have recently been explored by government agencies in the European Union, United States, and New Zealand for simulation, planning, and response purposes (e.g., MBIE, 2017; Lai et al., 2019; Allstadt et al., 2021).

The geospatial model originally proposed by Zhu et al. (2017) is a logistic regression model of the form $P(X) = (1 + e^{-X})^{-1}$ where X is a sequence of predictor variables and coefficients, and $P(X)$ is the likelihood of ground failure (i.e., surface manifestation). The model, which was trained on observations of ground failure, takes on two forms depending on a site's vicinity to a coastline. The equations for model parameter X are in Table 1. The variables are: PGV = peak ground velocity (cm/s); V_{S30} = shear-wave velocity of the upper 30-m (m/s) predicted from topography (Wald and Allen, 2007); dr = closest distance to a river (km) in the Lehner et al. (2006) dataset; dc = distance to coast (km); dw = the lesser of dr and dc (km); $precip$ = mean annual precipitation (mm) (Fick and Hijmans, 2017); and wtd = predicted water table depth (m) (Fan and Miguez-Macho, 2013). Following additional testing, Rashidian and Baise (2020) proposed two minor modifications to mitigate false positive predictions: (i) the model's output should be reassigned as zero below a PGA of 0.1 g; and (ii) the $precip$ input should be capped at 1700 mm/yr.

Table 1. Geospatial Liquefaction Model Equations (Zhu et al., 2017; Rashidian and Baise, 2020).

Model	Model Parameter X
(Coastal)	$12.435 + 0.301 \cdot \ln(PGV) - 2.615 \cdot \ln(V_{S30}) + 5.556 \times 10^{-4} \cdot precip - 0.0287 \cdot (d_c)^{0.5} + 0.0666 \cdot d_r - 0.0369 \cdot d_r \cdot (d_c)^{0.5}$
(Inland)	$8.801 + 0.334 \cdot \ln(PGV) - 1.918 \cdot \ln(V_{S30}) + 5.408 \times 10^{-4} \cdot precip - 0.2054 \cdot d_w - 0.0333 \cdot wtd$

1.2 A Test of Geospatial Liquefaction-Model Performance

Using approximately 15,000 liquefaction case histories compiled from 23 earthquakes by Geyin et al. (2021) and Geyin and Maurer (2021a), Geyin et al., 2020 tested the Zhu et al. (2017) geospatial model against 18 different CPT methods for predicting liquefaction surface manifestation. These were comprised of six different triggering models used in series with three different manifestation models. Because most of the case histories were sourced from three events in Canterbury, New Zealand, test cases were parsed into the "Canterbury" and "Global" datasets. Performance was quantified via receiver-operating-characteristic (ROC) analyses – specifically the area under the ROC curve, or AUC – which is a popular metric of prediction efficiency (e.g., Fawcett, 2006). Using this metric, a perfectly efficient model achieves an AUC of 1.0 whereas a model on par with random guessing achieves an AUC of 0.5. ROC analyses are also attractive in that they are insensitive to changes in class distribution. If the proportion of negative to positive instances in a test set changes, the AUC results will not change (Fawcett, 2006). P-values were computed per the method of DeLong et al. (1988) to determine whether measured differences in AUC could have arisen by chance (i.e., due to finite-

sample uncertainty) and not because one model is more efficient than another. While the reader is referred to Geyin et al. (2020) for complete details, the most salient results are summarized as follows. First, on the "Canterbury" dataset, the geospatial model performed significantly better than 16 out of 18 CPT models, with a measured *AUC* of 0.84. Against the top two CPT models, it was statistically indifferent, and thus either outperformed or matched all 18 CPT models. This is a surprising result, given the relative costs of the required model inputs. Second, on the "Global" dataset, all geotechnical models performed significantly better than the Zhu et al. (2017) model, with the latter performing only somewhat better than random guessing with an *AUC* of 0.55. This might be expected, given: (i) the variation of geomorphic, topographic, and climatic environs in a global dataset; and (ii) the challenge, given this variation, of accurately predicting below-ground conditions from above-ground parameters. Inherently, the CPT models - being based on subsurface tests - *should* be more portable across environments. Nonetheless, the strong performance of the seminal Zhu et al. (2017) model in Canterbury demonstrates the promising potential of geospatial data for regional-scale purposes, a conclusion similarly reached by Lin et al. (2021a).

1.3 Limitations of Existing Geospatial Modelling Approaches

By way of the study above, possible shortcomings of the Zhu et al. (2017) and Rashidian and Baise (2020) geospatial model, henceforth referred to as RB20, were identified. It should be emphasized that this model has transformed the perception of geospatial modeling for geohazards. Nonetheless, like all models it has shortcomings that could be improved upon. *First*, RB20 was trained directly on outcomes (i.e., observations of ground failure) rather than on the mechanistic causes of those outcomes (i.e., subsurface engineering properties). While this lack of a mechanistic underpinning can be overcome with vast training data (e.g., how voice transcription apps predict words without understanding language), current ground failure inventories are arguably too sparse. Specifically, both "positive" and "negative" cases (i.e., sites with and without observed liquefaction) are needed in which predictor variables span the range of possible values. That is, the parameter space of all predictor variables should be fully populated. Yet, while liquefaction is common in earthquakes, ground failure inventories are slow to grow (relative, for example, to those of in-situ test data). Given the adopted approach, inadequate training data can result in a divergence from mechanistic principles (e.g., prediction of liquefaction given shaking too weak, from a mechanistic perspective, to induce liquefaction).

Second, RB20 uses just 5-variables. Four represent capacity (distance to surface water; precipitation; and mapped V_{S30} and groundwater depth) and one represents demand (*PGV*). Notably, none of these variables are likely to correlate to the type of soil, or by corollary, to the susceptibility of the soil to liquefaction. This was a common cause of mispredictions identified in the Geyin et al. (2020) study, with RB20 expecting susceptibility if the ground is flat, saturated, and near water. However, such profiles can consist mostly of soils less- or un-susceptible to liquefaction (e.g., clays, peats, or gravels). Moreover, we find in many such cases

that geologic maps accurately predict the presence of such soils. An improved model might thus use mapped geologic data, when available, and/or other yet unidentified proxies of soil type. *Third*, and following from the above, RB20 is a traditional regression equation. This method of modeling inherently requires hypotheses of what is believed to matter and how (beliefs that are unnecessary with ML/DL). The efficient prediction of subsurface traits likely requires more than four geospatial variables, yet regression limits the number easily modeled. Algorithmic learning would allow more geospatial predictor variables to be used, with greater potential for those variables/data to be exploited fully.

1.4 The Proposed Modelling Approach and its Potential Benefits

This paper proposes a new geospatial modeling approach that is driven by algorithmic learning but pinned to an established mechanistic framework. Specifically, ML/DL models will be trained to predict *LPI* values in the absence of subsurface test data. Prior to model training, *LPI* values are computed from a national database of in-situ geotechnical tests subjected to a range of hypothetical ground motions. During model training, the ML/DL models will learn to predict these *LPI* values using twelve predictor variables. These variables consist of *PGA* and M_w, which are "demand" variables, and ten geospatial parameters from the geotechnical test site, which are "capacity" variables. The goal of these ten geospatial variables, in effect, is to predict the relationship between *LPI* and seismic loading in the absence of subsurface data. Multiple models will be developed and ensembled, thereby avoiding large "swings" on account of which model is chosen (as is common in prediction of ground motions, hurricane tracks, etc.). When used in the forward direction, the trained models predict *LPI* at sites without geotechnical testing, given *PGA*, M_w, and geospatial variables sampled at the coordinates of the sites. To complete the prediction of ground failure, the predicted *LPI* values will be input to existing fragility functions (Geyin and Maurer, 2020a) that predict the probability of liquefaction manifestation (i.e., "ground failure") as a function of *LPI*. These functions were trained on a large database of well-documented liquefaction case histories compiled from 24 global earthquakes. Thus, the ultimate output is a predicted probability of ground failure (the same as RB20). A synopsis of the proposed approach is shown in Fig. 1.

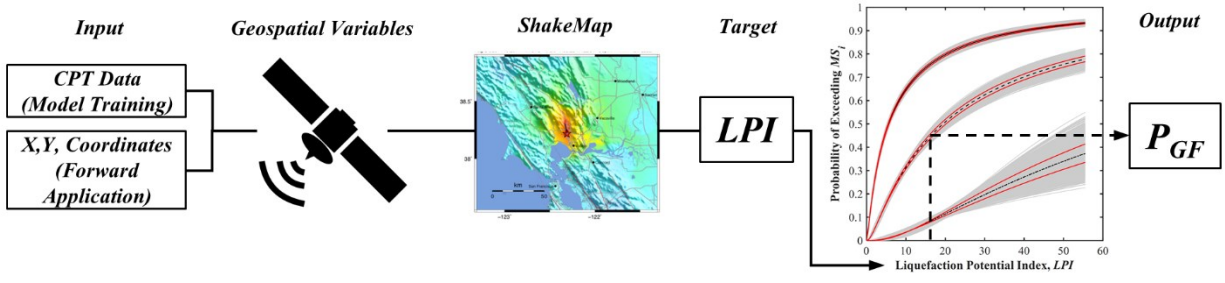

Fig 1. Synopsis of the proposed modeling approach to predict the probability of ground failure (P_{GF}).

This approach has several potential advantages:

- The principal prediction target is transferred from ground failure (with relatively sparse training data) to subsurface measurements (for which the potential training set is vast). Because the location of in-situ tests need not have experienced an earthquake (i.e., be a liquefaction case history), significantly more training data is available. Given the rise of community geotechnical datasets – both internationally and in the U.S. – the gap between the number of subsurface tests and the number of liquefaction case histories will likely grow. We hypothesize this larger training set will be advantageous, both now and in the future.
- Liquefaction is a physical phenomenon best predicted by mechanics. Much has been learned about liquefaction over the last 50 years. This knowledge is embedded in current state-of-practice liquefaction triggering and manifestation models. We hypothesize that anchoring to these models, which provide a mechanistic foundation, will be advantageous, given their validated ability to model liquefaction response as a function of soil and profile traits (e.g., subsurface stratigraphy, soil density, fines-content, plasticity, saturation, and ground motion duration and intensity).
- Whereas liquefaction is best predicted by mechanics, subsurface traits lack theoretical links to above-ground parameters (i.e., geospatial data), but surely correlate to them in complex, interconnected ways. This is a prime problem for ML/DL, which can provide learning insights that are unlikely, if not infeasible, with traditional regression approaches. We hypothesize that ML/DL provides the potential for geospatial data to be exploited more fully.
- The models are updated easily as additional training data (in-situ tests) become available. In the short term, some geospatial variables could be viewed by the learning algorithms as relatively unimportant, either because they truly are unimportant, or because there is insufficient training data to elucidate their predictive value. Existing geospatial models are also retrainable, but we hypothesize that meaningful growth in the ground-failure datasets that they are trained upon will take place at a slower pace (e.g., with data from a few events annually that impact a small fraction of earth), whereas growth in community geotechnical datasets will proceed more quickly.

2. Data and Methodology

In the current effort, two models will be developed using relatively modest sets of training data and predictor variables compiled in the United States (US). As will be discussed, the proposed approach could be extended using additional training data and new predictor variables at regional, national, or global scales. Nonetheless, the developed models, which are applicable to the US, will be shown to be at least as efficient as other geospatial models and thus warrant application and further evaluation, even if preliminary in nature.

While several in-situ geotechnical tests could be used within the proposed approach, we choose CPT data given that: (i) it has inherent advantages over other tests upon which liquefaction models have been based

(NRC, 2016); (ii) the Geyin and Maurer (2020a) fragility functions were trained on CPT-based case histories; and (iii) a US national CPT database is readily available in native digital format. Specifically, the USGS national database of 1,712 CPTs (USGS, 2021) was adopted for analysis. This dataset provides somewhat well-distributed measurements, as mapped in Fig. 2, in a range of environments, generally in high-seismicity regions. Approximately 5% are from sites where liquefaction case histories were compiled following modern earthquakes. Given the limited dataset, some regions of the US are unrepresented in model development, as shown in Fig. 2. Ultimately, however, tests of the derivative geospatial models are not clearly suggestive of regional bias (i.e., the models perform well in regions with no training data). Nevertheless, it stands to reason that an expanded dataset would result in better models. Of the 1,712 CPTs, 20% were randomly selected and reserved for model testing, while the remaining 80% were used for model training.

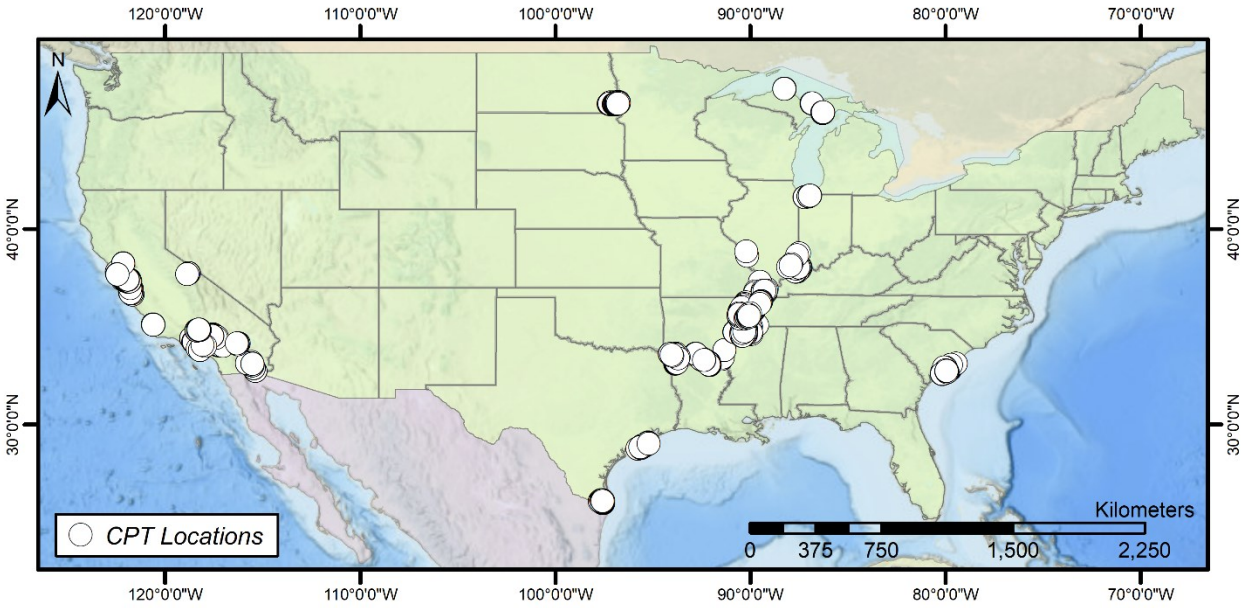

Fig. 2. Spatial distribution of CPT training and test data.

Next, each CPT was subjected to 152 combinations of *PGA* and M_w, with *PGA* ranging from 0.0 g to 1.0 g and M_w ranging from 4.5 to 9.0. These represent loadings that could hypothetically impact a site and for which it would be of interest to predict liquefaction. We exclude $M_w < 4.5$ events based on Green and Bommer (2019). With the proposed approach, however, it is irrelevant whether the CPT sites ever experienced an earthquake or whether a specific combination of *PGA* and M_w could feasibly occur in the future. In other words, we assume that subsurface conditions are generally independent of the seismic hazard (i.e., saturated, loose, cohesionless soils are equally present in high seismicity regions as in low seismicity regions). As such, it is not necessary that CPTs be subjected to site-specific combinations of *PGA* and M_w that are more likely to occur (e.g., according to a probabilistic seismic hazard analysis). For each combination of *PGA* and M_w, the

Idriss and Boulanger (2008) CPT liquefaction model was used to predict FS_{liq} versus depth. These predictions were then input to the *LPI* manifestation model, as defined in Eq. 1. All CPT processing and calculations were performed using the software *Horizon* (Geyin and Maurer, 2020b). While different, or additional, triggering and/or manifestation models could be used, the Idriss and Boulanger (2008) triggering model - when used in conjunction with *LPI* - demonstrated an efficiency that was never bested, to a statistically significant degree, by any other model when tested on global case-history data (Geyin et al., 2020). In addition, the magnitude-scaling factor (*MSF*) inherent to Idriss and Boulanger (2008) is soil-independent, whereas other triggering models (Boulanger and Idriss, 2014; Green et al., 2019) have *MSF*'s that vary with depth depending on the inferred relative density. The more predictable scaling of computed *LPI* with increasing M_w was deemed advantageous for modeling, given the limited training data utilized herein. A subsequent study could explore the use of other CPT models, although prior testing of such models (Geyin et al., 2020) suggests the efficacy of the resulting product would be very similar.

Given the 1,712 CPTs and 152 combinations of seismic loading, a total of 260,224 *LPI* values were computed. These values are plotted in Fig. 3 as a function of magnitude-scaled *PGA* ($PGA_{M7.5}$), as computed by Idriss and Boulanger (2008), and form the primary prediction target of the proposed modeling approach. At sites of high liquefaction hazard (i.e., thick deposits of saturated, loose sand), *LPI* increases rapidly with $PGA_{M7.5}$, whereas at sites of low hazard (i.e., sites devoid of soil susceptible to liquefaction), *LPI* may remain near zero for all $PGA_{M7.5}$. The goal of the geospatial modeling, in effect, is to predict the relationship between *LPI* and seismic loading (*PGA*, M_w) in the absence of subsurface data.

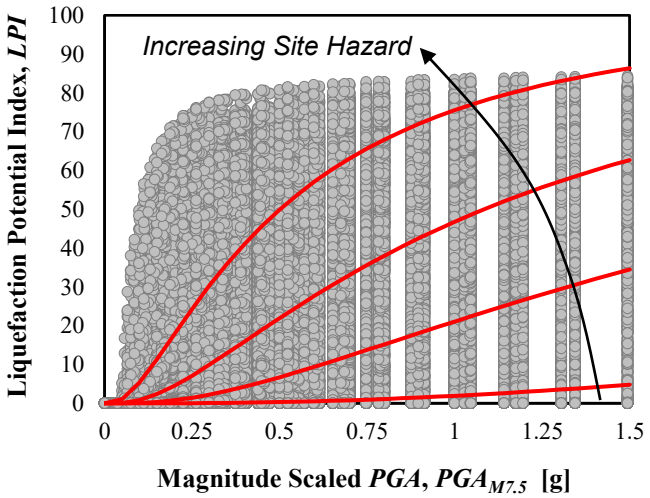

Fig. 3. *LPI* versus $PGA_{M7.5}$. Plotted are 260,224 *LPI* values computed from 1,712 CPTs subjected to 152 different levels of seismic loading.

Ten geospatial predictor variables were next compiled at the coordinates of each CPT. The goal of these ten variables is to correlate to the subsurface conditions which give rise to low or high *LPI*. These consisted

of: predicted V_{S30} (Heath et al., 2020); predicted ground water depth (Fan and Miguez-Macho, 2020); measured distance to river (Lehner et al., 2006) and measured distance to coast (NASA, 2012); predicted depth to bedrock (Shangguan et al. 2017); measured annual precipitation (Fick and Hijmans, 2017); and the predicted (binomial) presence of unconsolidated soil, sandy soil, clayey soil, and silty soil, as obtained from the USGS National Geologic Map compilation (Horton et al. 2017). The intention of the latter four variables is to predict whether soil is present, and if so, whether it has one of these three predominant soil types. Additional mapped soil types were ultimately found not to be useful, as will be further discussed. The range of predictor variables in the dataset and their spatial resolutions are given in Table 2. In lieu of predicted V_{S30}, we also explored the use of measured topographic slope, which ultimately produced models with nearly identical performance. This is unsurprising, given that V_{S30} is most often predicted solely from topographic slope in the Heath et al. (2020) compilation, which merges several regional V_{S30} maps with a general slope-based V_{S30} model. We adopted the predicted V_{S30} from Heath et al. (2020), however, given that it includes region-specific insights into the relationship between topographic slope and subsurface conditions. While the potential benefits of using V_{S30} from Heath et al. (2020) (i.e., versus topographic slope) were not realized during model training and testing, such benefits could conceivably be observed in future, forward applications elsewhere. Notably, models developed without either parameter performed significantly worse. The importance of each predictor variable will be further discussed later in the paper.

Table 2. Range of predictor variables in the dataset and their spatial resolutions.

Variable (Units)	**Range in Dataset**	**Spatial Resolution**
Earthquake Magnitude (M_w)	4.5 to 9	N/A
Peak Ground Acceleration (g)	0 to 1	N/A
Ground Water Table Depth (m)	0 to 216	~1000 m (30 arc-sec)
Distance to River (m)	2 to 6,220	~90 m (3 arc-sec)
Distance to Coastline (km)	0 to 1,210	~90 m (3 arc-sec)
Depth to Bedrock (cm)	379 to 21,717	250 m
Annual Precipitation (mm)	68 to 1,389	~1000 m (30 arc-sec)
V_{S30} (m/s)	92 to 713	~1000 m (30 arc-sec)
Unconsolidated Soil (binomial)	0 or 1	25 m to 500 m (varies)
Dominant Clay (binomial)	0 or 1	
Dominant Silt (binomial)	0 or 1	
Dominant Sand (binomial)	0 or 1	

Using the training set (80% of CPTs), models were next developed to remotely predict *LPI* as a function of *PGA* and M_w, which may be viewed as "demand" variables, and of the ten geospatial variables, which may

be viewed as "capacity" variables. The latter can be compiled at national scale in advance of model application. The former are available at regional scale minutes after an earthquake (e.g., via a single "ShakeMap" file (Wald et al., 2005)) or for various future earthquake scenarios. Like other geospatial models, this gives the model near-real-time functionality, such that ground failure can be predicted at regional scale minutes after an event. Various ML/DL techniques were explored, including Gaussian process models (e.g., Rasmussen and Williams, 2006), support vector machines (SVM) (e.g., Vapnik, 1995), decision trees (e.g., Rokach and Maimon, 2008), model ensembles with bagging, gradient boosting, or random forests (e.g., Breiman, 1996; Piryonesi et al., 2021; Ho, 1998) and neural networks (e.g., Glorot et al., 2010). In general, modeling approaches that are easier to interpret tend to have lower predictive accuracy (e.g., single decision trees, support vector machines), while those with higher accuracy (e.g., neural networks, or ensembles of decision trees) are typically very complex to interpret. Each approach has numerous options and internal parameters (i.e., "hyperparameters") (e.g., neural net optimization algorithm, activation function, and layer quantity and size; regression tree leaf size; Gaussian basis and kernel functions; SVM kernel scale and box constraint). Once promising models were identified, hyperparameter optimization was employed, such that the hyperparameter values that minimized the model error were identified via an automated optimization scheme. 5-fold cross-validation was used to control overfitting, as is common in model development. Additionally, training and test performance metrics were compared for signs of overfitting (i.e., better training performance than test performance), which was inferred when performance metrics from the training and test sets differed by at least 4%. In this regard, models with slightly lower accuracy but without overfitting were favored over models that achieved the highest training accuracy but with suspicion of overfitting. Because many ML/DL algorithms either require or perform better when variables have a Gaussian distribution, all predictors were BoxCox transformed (Box and Cox, 1964) and normalized to have values between 0 and 1. Ultimately, the software in which the prediction models are implemented performs all necessary computations, and as such, no pre-processing of data is required (e.g., predictor variables are input in their native format).

3. Results and Discussion

Using the aforementioned methodology with relatively modest sets of training data and predictor variables, several dozen preliminary models were trained. Of these, two were ultimately adopted for further implementation and testing. The first model is a boosted ensemble of decision trees, wherein numerous relatively weak models are coalesced to form one high-quality model. For brevity, we henceforth refer to it as the "ML model." The theory and algorithm underlying this approach – which is commonly included in machine learning toolkits (e.g., Scipy, TensorFlow) – is explained in detail by Friedman (2001). An excellent overview of its practical implementation is provided by Elith et al. (2008). The growth of a decision tree involves establishment of recursive binary splits, such that specific combinations of model inputs map to a predicted

output. However, because a single tree is prone to overfitting and tends not to be very accurate, models that ensemble many decision trees are preferred. In "gradient boosting", a strong learner is sequentially built from weak learners, wherein each tree attempts to diminish the errors of the previous tree by gradually increasing emphasis on observations poorly predicted by the ensemble. While gradient boosting is slow, it generally produces a more accurate model compared to other assembling algorithms (e.g., bagging or random forests) (Piryonesi et al., 2021). With respect to performance, the ML model achieved a mean absolute error (MAE) (*LPI* units) of 3.58 and 3.72 on the training and test sets, respectively, as summarized in Table 3. As discussed previously, the unbiased test set consists of *LPI* data from CPT sites unknown to the model during training.

The second model is a deep (7-layer) artificial neural network, which we henceforth refer to as the "DL model." With roots in the 1980's (e.g., Hopfield, 1982), this now ubiquitous approach mimics the perceived structure of the human brain, with layers of interconnected nodes. At the most basic level, DL models have four components: inputs, weights, a threshold, and an output. Connections between nodes are modelled as weights, such that positive and negative weights indicate excitatory and inhibitory connections, respectively. If the output from an individual node is above a specified threshold, the node is activated, sending data to the next layer of the network. An activation function then controls the amplitude of the output at each node. As DL models have multiple layers, the above process is repeated multiple times, with each layer potentially passing information from the previous layer to the next. During training, the weights are iteratively adjusted to optimize model performance. Like the ML model, DL models are quite convoluted, rendering simple interpretations of the inner workings infeasible, since single node weights have little physical meaning, and since millions of connections may be present in a model. As shown in Table 3, the DL model achieved a MAE of 4.13 and 4.20 on the same respective datasets (i.e., it performed slightly worse than the ML model). Given the limited training set and preliminary nature of the ML and DL models, we also create a third "Ensemble model" by averaging the outputs of the ML and DL models. The merging of two models with different structures could have the effect of "stabilizing" predictions and, conceivably, provide benefits unrealized during testing. As shown in Table 3, the ensemble performs better than the DL model and worse than the ML model, although all are similarly efficient when considering the range of the *LPI* domain (i.e., zero to 100). The performance of these models, and other results in Table 3, will be further discussed momentarily.

Table 3. Summary of model performance (mean absolute error) on the training, test, and overall datasets.

Model	Mean Absolute Error (*LPI* Units)			Mean Absolute Error (Probability Units)		
	Training	Test	Overall	Training	Test	Overall
ML	3.5814	3.7237	3.6642	7.1691	7.6482	7.3698
DL	4.1329	4.2097	4.175	8.6918	8.9609	8.7894
Ensemble	3.743	3.8499	3.8039	7.6491	8.0338	7.7967

While the convoluted nature of ML/DL models tends to obscure simple interpretations of model function (e.g., relative to traditional regression), insights into the ML decision-tree ensemble can be gained via predictor importance (e.g., Auret and Aldrich, 2011), which may be interpreted as the relative contribution of each variable to model accuracy. Accordingly, the relative importance of each variable was computed and is plotted in Fig. 4, where variables are sorted from most to least important. This approach and presentation mirrors that of Durante and Rathje (2021), who explored the ML prediction of lateral spreads using geospatial data. As could be expected, the magnitude-scaled *PGA* and predicted groundwater depth had the largest importance, given the mechanistic relationship between these inputs and computed *LPI*. Also of relatively large importance were the measured distance to river and predicted depth to bedrock, which correlate to the expected thickness and geomorphology of deposits. Bedrock at shallow depth limits *LPI* while bedrock at very large depth suggests the presence of a sedimentary basin, which tends to collect sands and silts in a low-velocity flow regime. The presence of a nearby river, particularly in combination with flat topography, suggests a similar geomorphology, while also indicating that the deposits are likely young and saturated. Of the compiled geologic data, the predicted predominance of clay was most important, whereas the predicted predominance of silt was least important. This aligns with expectations, given the established use of plasticity index to infer liquefaction susceptibility within mechanistic models. Whereas soils classifying as clay are rarely susceptible, silts are an intermediate soil whose liquefaction response is difficult to predict from name alone. We hypothesize that the overall importance of mapped soil type could increase if the set of training sites was larger and more diverse, given that the set used herein does not span the full range of geologic conditions that may be encountered. While the computed predictor importance gives insights into decision-tree models, we are unaware of any analogous tool for studying neural networks, which thus remain relatively more convoluted.

Following prediction of *LPI* via the ML, DL, or Ensemble models, probabilities of ground failure were computed using the Geyin and Maurer (2020a) fragility functions, which are conditioned on *LPI*. As an example, the test and training set performance is shown in Fig. 5 for the ML model. Here, the "predicted probability" is the output when *LPI* is predicted via the geospatial ML model, whereas the "actual probability" is that when *LPI* is computed from the CPT data. Also shown in Fig. 5 are linear trendlines (green dotted lines), from which assessments of overall prediction bias may be made. The ML model's MAEs of 3.58 and 3.72 (*LPI* units) on the training and test sets translate to MAEs of 7.17% and 7.65% in probability units (Table 3). These comparisons (i.e., predicted vs actual probabilities of ground failure) provide the clearest context of model performance, given that the consequences of an *LPI* error vary widely depending on the *LPI* value. A prediction of *LPI* = 17, for example, is relatively erroneous if the actual *LPI* is 2, since this translates to a ~65% overprediction of ground-failure probability (Geyin and Maurer, 2020). In contrast, a prediction of *LPI* = 87 is very accurate if the actual *LPI* is 72, given that the probability of ground failure is nearly identical whether *LPI* is 87 or 72. For this reason, direct comparison between predicted and actual *LPI* values is arguably less

meaningful. As seen in Fig. 5, the model is generally unbiased on the training and test sets, but it does exhibit relatively more bias on the latter, such that the predicted probability of ground failure has an average tendency to be 2% greater than actual. This might be attributable to the dataset's modest size, such that the test set has features unrepresented in the training set.

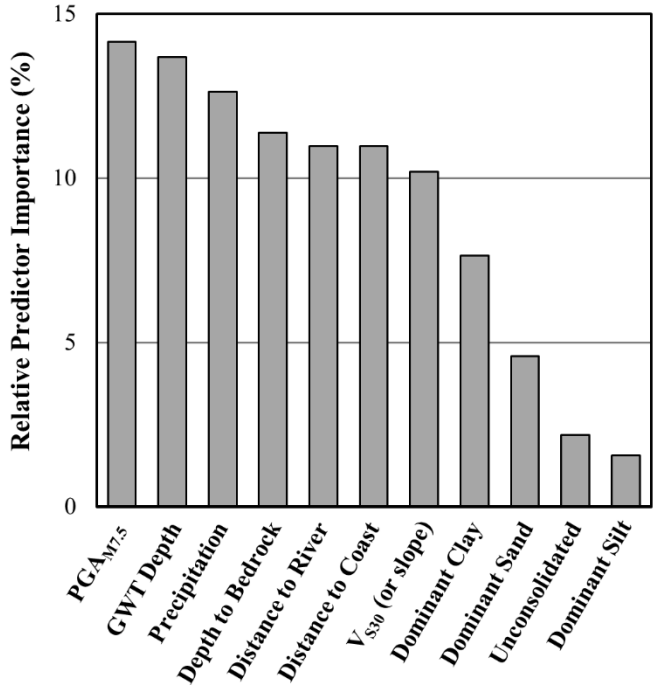

Fig. 4. Relative predictor importance ranking for the ML model.

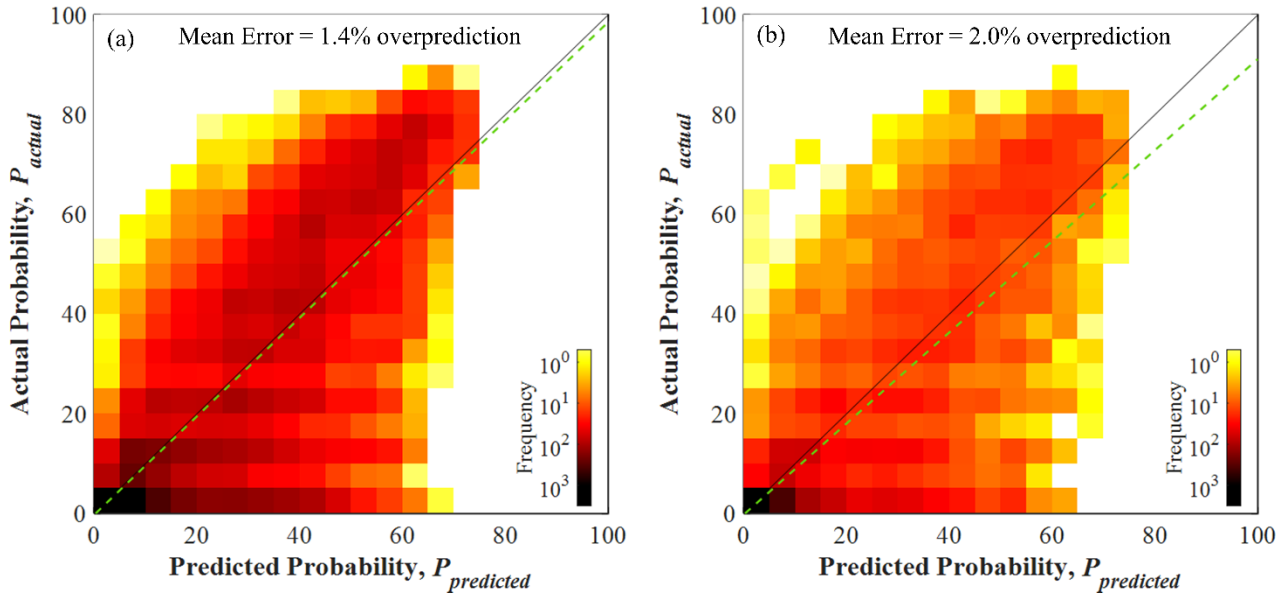

Fig. 5. Probability of ground failure: ML prediction vs. actual for the (a) training dataset; and (b) test dataset. Green dotted lines = linear trendlines, from which prediction bias may be judged.

3.1 Field Application and Testing

To demonstrate and test forward predictions at regional scale, the ML, DL, and Ensemble models were next used to predict ground failure in eleven U.S. earthquakes across two types of datasets. Performance was assessed using field observations and compared against the Rashidian and Baise (2020) model in all events. Due to the paucity of recent, well documented U.S. earthquakes outside of California, these tests cover a relatively narrow geographic range. Further testing on future events in other U.S. regions is thus needed. In the first series of tests, regional scale predictions are compared to mapped observations of ground failure in six events: (i) 1989 M_w6.9 Loma Prieta, California; (ii) 1994 M_w6.7 Northridge, California; (iii) 2001 M_w6.8 Nisqually, Washington; (iv) 2003 M_w6.5 San Simeon, California; (v) 2011 M_w5.8 Mineral, Virginia and (vi) 2016 M_w7.1 Ridgecrest, California. In these events, mapped observations of liquefaction-induced ground failure were obtained from the USGS Ground Failure Database (Schmitt et al., 2017a,b), except for observations in the 2011 Mineral and 2016 Ridgecrest events, which were respectively obtained from Green et al. (2015) and Zimmaro et al. (2020). The quantities of mapped observations in these six events are respectively 129, 41, 44, 12, 35, and 2. In these datasets, mapped observations are exclusively "positive" (i.e., a lack of liquefaction is not explicitly mapped). It was therefore assumed that liquefaction did not manifest if none was documented, as has been previously assumed in the development of geospatial hazard models (e.g., Zhu et al., 2017). While this assumption may at times be invalid and inevitably introduces uncertainty, it facilitates rapid, regional-scale testing across a variety of topographic and geomorphic environments. In this regard, we view performance in the context of model comparisons and not as an absolute measure of efficacy. Later, a separate dataset containing positive *and* negative observations at discrete sites will be studied.

For each earthquake, a USGS ShakeMap file with all requisite seismic data (i.e., M_w and mapped *PGA* and *PGV*) was obtained in .xml format. The adopted geospatial predictor variables were then compiled across the ShakeMap extents (i.e., the area of perceptible shaking). As an example, these inputs are mapped in Fig. 6 for the 1989 Loma Prieta earthquake. It can be seen in the final two panels of Fig. 6 that while unconsolidated soil covers ~40% of the study area, the dominant soil type is infrequently mapped as either sand, silt, or clay. Although soil lithology is always defined in the Horton et al. (2017) compilation, it is not always used by the models developed herein for one of two reasons. First, not all dominant soil types were found to be useful in the early stages of modelling. Some mapped soil types (e.g., marl, gravel, peat) have insufficient in-situ test data to elucidate and quantify the relationship between soil type and liquefaction hazard. Second, the mapped lithology is sometimes not predicted to a useful degree of specificity (e.g., as "sand" or "clay"), but rather, is defined only as "coarse detrital" or "fine detrital." These broad classifications were similarly found not to be useful, which might be expected given that particle gradation is generally not efficient or sufficient for classifying liquefaction hazard. Accordingly, the model benefits from knowledge of the mapped soil type when it is clay, sand, or silt, whereas if the mapped soil type is not one of these classifications, it is inherently treated

as having a general unconsolidated character. It is likely that additional geologic descriptors would be useful to future models that use larger and more diverse sets of training data. Following compilation of the adopted geospatial predictor variables, probabilities of ground failure were computed using the three models developed herein and RB20. These probabilities are mapped in Fig. 7 for the 1989 Loma Prieta earthquake, along with observations of ground failure.

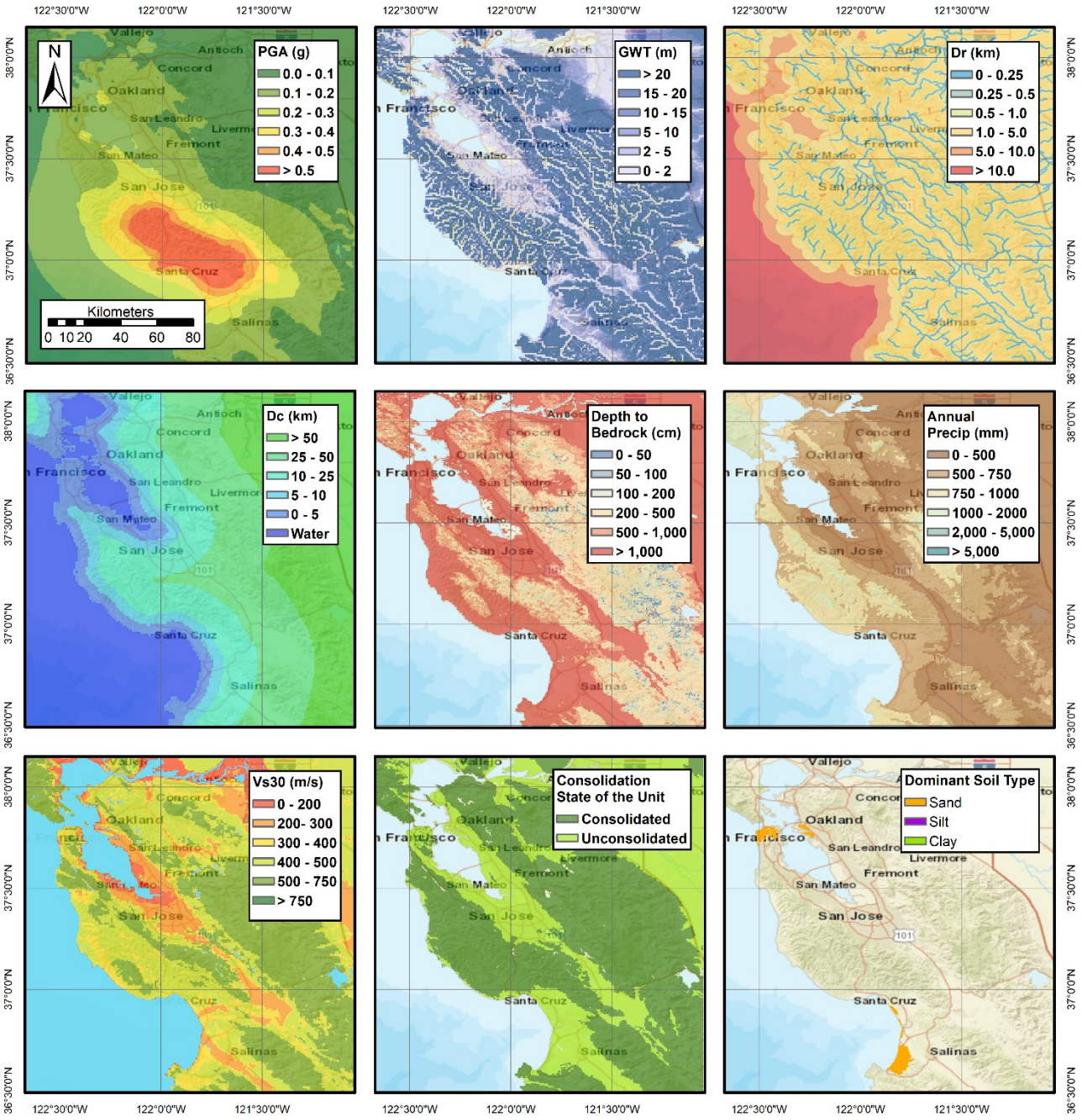

Fig. 6. ML/DL model predictor variables mapped across the area effected by the 1989 Loma Prieta earthquake.

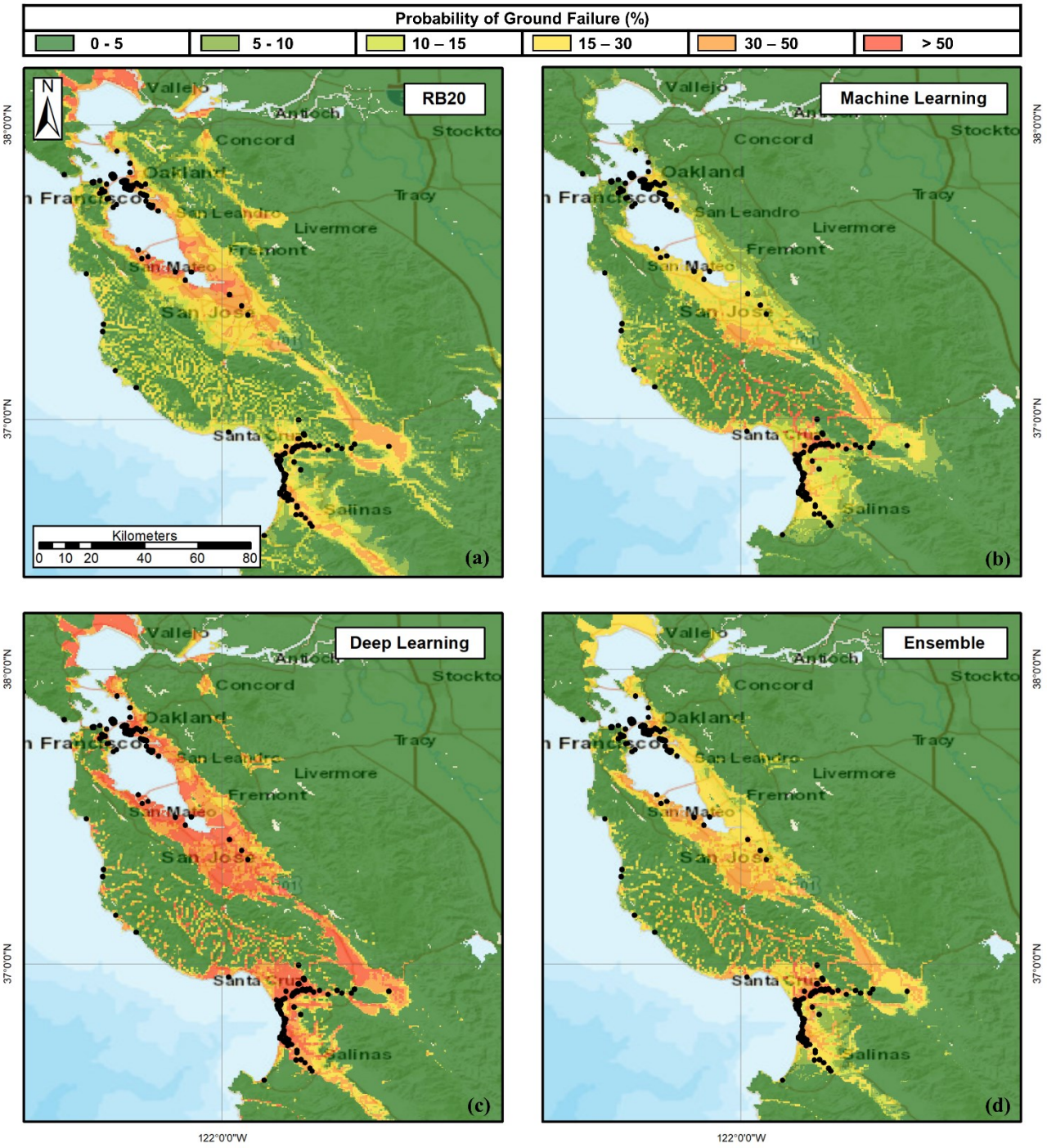

Fig. 7. Probabilities of ground failure in the 1989 Loma Prieta earthquake, as computed by the: (a) RB20; (b) ML; (c) DL; and (d) Ensemble models. Black dots are observed ground failures.

405 Model performance was quantified using ROC *AUC* values, as is common for binomial classifiers, and
406 which give equal weighting to false positive and false negative predictions. Samples were collected on a 100
407 m by 100 m grid across the ShakeMap extents. Grids wherein ground failures were observed were classified

as “positive” and those without any documented evidence as “negative.” This resulted in several million data points per event, although the exact quantity depended on the event’s area of influence. While the method of geospatial sampling has been shown to influence computed *AUC* values (Lin et al., 2021b) (e.g., if an equal number of positive and negative points were sampled instead), we found that *relative* performance was insensitive to this decision (i.e., the best and worst models were the same in each event across a range of sampling techniques). Plotted in Fig. 8 are ROC curves for each model in the 1989 Loma Prieta earthquake from which *AUC* values were computed. Arranged by *AUC*, the best performing models were RB20 (*AUC* = 0.949), Ensemble (*AUC* = 0.945), DL (*AUC* = 0.944), and ML (*AUC* = 0.931). The four models thus exhibited very similar efficiencies, with the Ensemble model slightly outperforming the individual ML and DL models. Following the same methodology, analyses were performed for the 1994 Northridge and 2001 Nisqually events, as mapped in Fig. 9, and for the 2011 Mineral and 2016 Northridge events, as mapped in Fig. 10. A summary of model performance – as quantified by *AUC* – is presented in Table 4 for these events and others yet to be discussed. It can be seen that RB20 outperformed the Ensemble model in three of the six events. Specifically, in Loma Prieta by 0.4%, in Northridge by 3.5%, and in Nisqually by 1.3%. Conversely, the Ensemble model outperformed RB20 in San Simeon by 0.3%, in Mineral by 2.9%, and in Ridgecrest by 1.4%. The models proposed herein thus demonstrate efficacies similar to RB20 for these specific events. While these measured differences in performance fluctuate with different sampling techniques, the overall conclusion of apparently similar performance remains the same. In subsequent analyses, it will be determined whether these measured differences in performance are statistically significant.

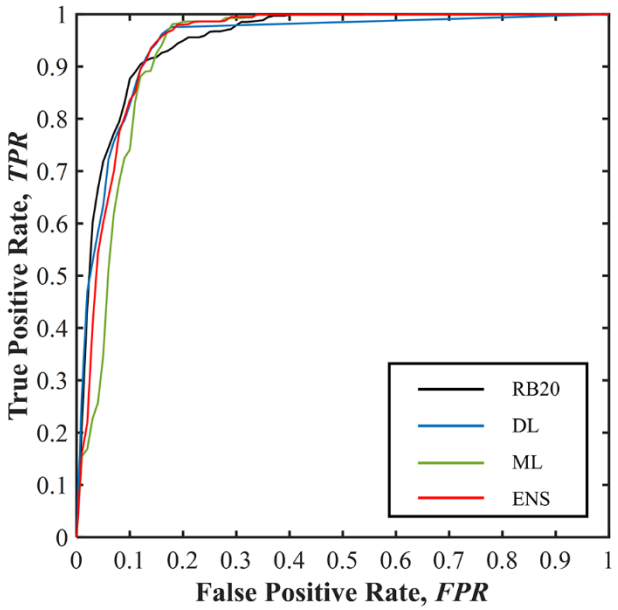

Fig. 8. Receiver Operating Characteristic (ROC) curves for the RB20, DL, ML, and Ensemble models.

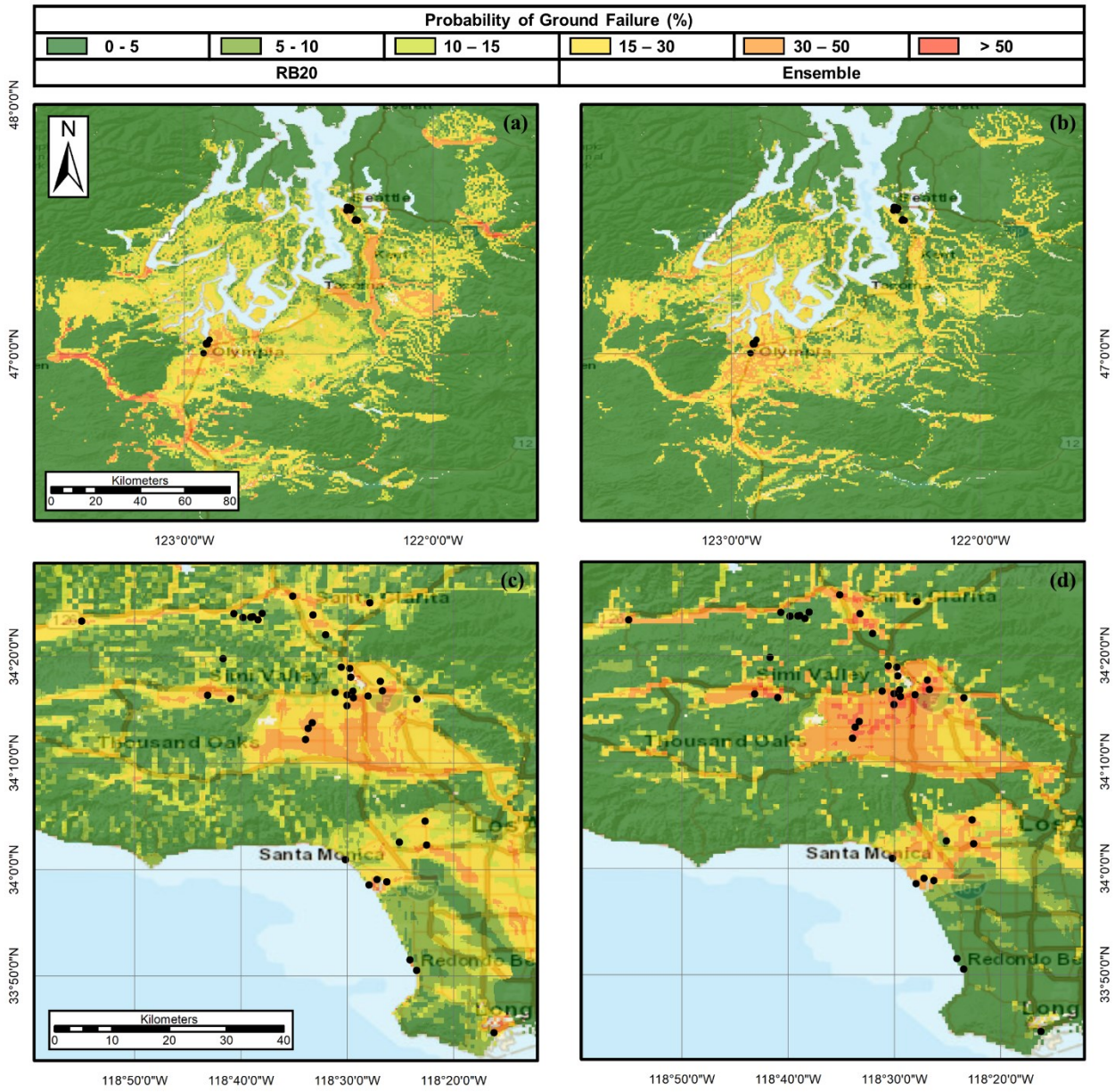

Fig. 9. Probabilities of ground failure in the 2001 Nisqually earthquake, as computed by (a) RB20; and (b) the Ensemble model; and in the 1994 Northridge earthquake, as computed by (c) RB20; and (d) the Ensemble model. Black dots are observed ground failures.

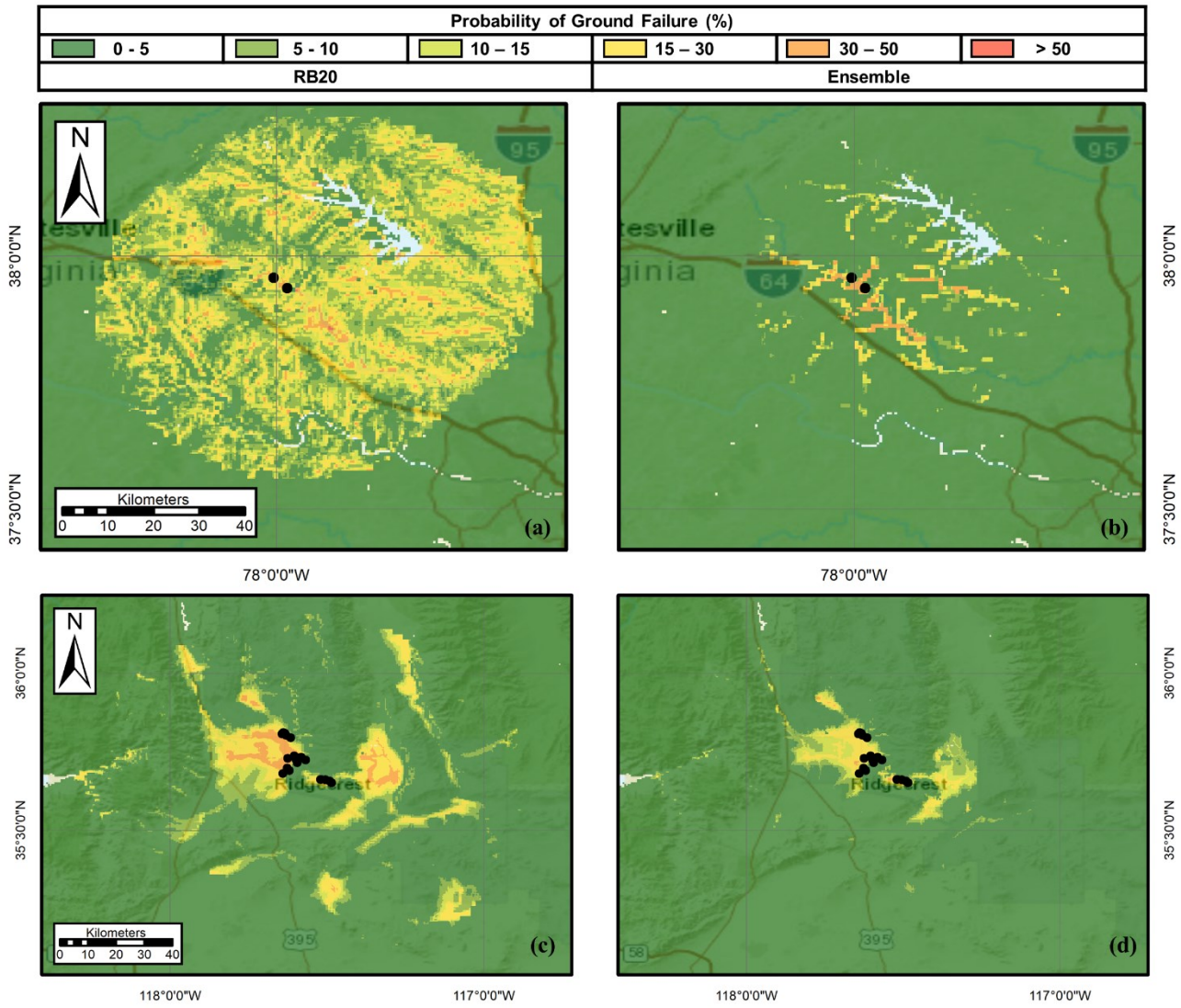

Fig. 10. Probabilities of ground failure in the 2011 Mineral earthquake, as computed by (a) RB20 and (b) the Ensemble model; and in the 2016 Ridgecrest earthquake, as computed by (c) RB20; and (d) the Ensemble model. Black dots are observed ground failures.

Table 4. Summary of *AUC* values for events and datasets described in the text, as computed for the ML, DL, and Ensemble models developed herein, and for the RB20 model.

Model	Dataset 1						Dataset 2
	Loma Prieta	Northridge	Nisqually	San Simeon	Mineral	Ridgecrest	101 Case Histories
DL	0.944	0.803	0.931	0.665	0.655	0.992	0.682
ML	0.931	0.812	0.920	0.980	0.733	0.945	0.765
Ensemble	0.945	0.813	0.933	0.979	0.732	0.992	0.734
RB20	0.949	0.848	0.946	0.976	0.703	0.978	0.504

Additional observations from Figs. 7-10 and Table 4 are as follows. *First*, with respect to model bias, the RB20 model was either originally trained (by Zhu et al., 2017) or later calibrated (by Rashidian and Baise, 2020) using the same observational data adopted herein for testing, except for the Ridgecrest earthquake data which postdated Rashidian and Baise (2020). In contrast, the ML/DL models were trained on CPTs from areas affected by the Loma Prieta, Northridge, and San Simeon events, but not directly on the field observations adopted for testing. Moreover, the Nisqually, Mineral, and Ridgecrest earthquakes provide completely blind tests of the ML/DL models since no data from these events/regions were included in training. While a more rigorous analysis of bias is not undertaken, nor critical to the thesis of this study, we nonetheless note that the preceding tests were generally biased in favor of RB20.

Second, it was observed that the DL model is relatively sensitive to predicted water table depth, as compared to the ML and RB20 models. In this regard, erroneous predictions by the DL and Ensemble models were often associated with erroneous expectations of the groundwater depth. As an example, predictions in the 2003 San Simeon earthquake by the DL and RB20 models are mapped in Fig. 11. Specifically, an area near the towns of Oceano and Grover Beach, California is shown, where numerous ground failures were observed as mapped in Fig. 11. Because the Fan and Miguez-Macho (2020) model predicts a groundwater depth of ~20 m beneath the northernmost features, the DL model predicts a near-zero probability of ground failure, whereas RB20 generally predicts a probability of 5-15%. Due largely to this behavior, the DL and RB20 models had respective *AUC*s of 0.665 and 0.976 for this event. To assess the influence of more accurate inputs, nearby well measurements were obtained from the California Department of Water Resources (DWR, 2020), indicating that groundwater is shallower in this area than expected by Fan and Miguez-Macho (2020). Using this more accurate input, the models were rerun, as mapped in Fig. 11. While the RB20 and ML models correspondingly displayed slight improvements (~1% increase in *AUC*), the DL model's *AUC* increased nearly 30% to 0.990. Similar behavior could be observed in other events at a lesser scale, from which we conclude that the performance of the DL model would likely improve with more accurate groundwater maps.

Third, considering the three models developed herein, the Ensemble model outperformed both the ML and DL models in three events. The ML model performed best in two other events, and in the last event the DL and Ensemble models tied for best performance. Considering all tests (i.e., both the blind prediction of *LPI* and now the regional-scale prediction of ground failure), the DL model lacks statistical support for individual use. Accordingly, and in conjunction with the DL model's sensitivity to groundwater data, we recommend adoption of the ML or Ensemble models. Ultimately, additional tests in other events, and ideally additional model training and improvements, are needed before one model is recommended over another.

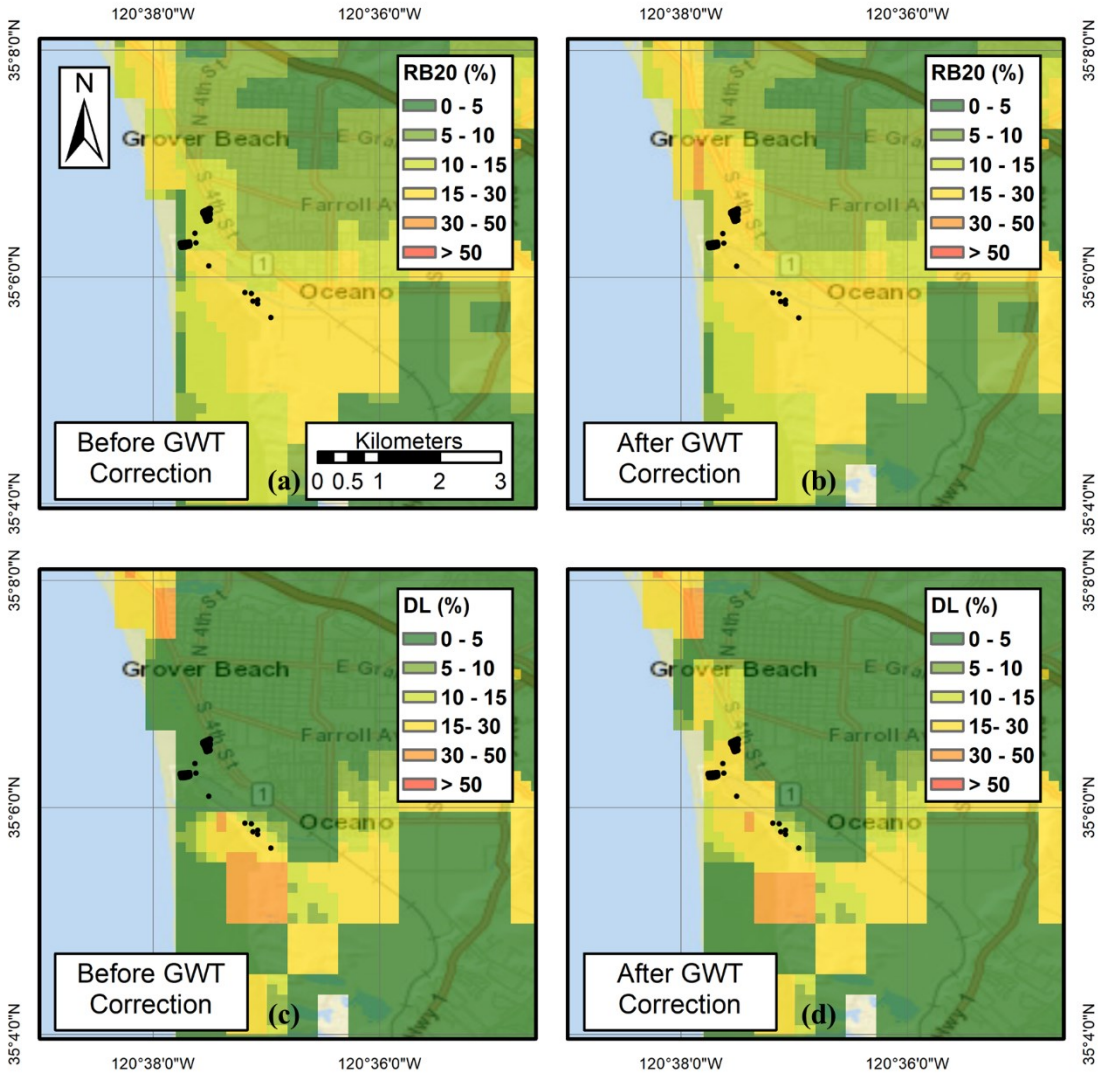

Fig. 11. Probabilities of ground failure in the 2003 San Simeon earthquake, with and without correction of measured ground water table (GWT) depths: (a) RB20 before GWT correction; (b) RB20 after GWT correction; (c) DL model before GWT correction; and (d) DL model after GWT correction. Black dots are observed ground failures.

458 Towards that end, a second dataset of 101 well-documented liquefaction case histories was also used to
459 test performance. These cases, which consist of both positive *and* negative observations, were sourced from
460 Geyin and Maurer (2021a), who compiled from the literature all CPT-based case histories from all earthquakes
461 in the U.S. to-date. Namely, the: (i) 1971 M_w7.6 San Fernando; (ii) 1979 M_w7.6 Imperial Valley; (iii) 1981
462 M_w5.9 Westmoreland; (iv) 1983 M_w6.9 Borah Peak; (v) 1987 M_w6.2 Elmore Ranch; (vi) 1987 M_w6.5
463 Superstition Hills; (vii) 1989 M_w7.6 Imperial Valley; and (viii) 1994 M_w6.9 Northridge earthquakes. In the
464 resulting compilation, liquefaction manifestations were observed in 57% of cases and were not observed in the

remaining 43%. The four geospatial models were applied to each event and *AUC* values were computed for the composite dataset, as given in Table 4. In this analysis, the ML/DL models performed much better than RB20, which had an *AUC* near 0.5, an efficiency akin to random guessing. Of the models developed in this study, the ML model performed best (*AUC* = 0.765). In contrast to the initial series of tests, however, those using this second dataset might be biased in favor of the ML/DL models, since some of the 101 sites held CPTs included in the dataset of 1,712 used in model development. While these tests provide another datapoint for consideration, wherein it is known with confidence that the field observations are correctly classified, we prefer not to glean definitive new conclusions, given the possibility of bias and small size of the dataset.

Lastly, to assess whether the findings presented thus far might change with consideration of finite-sample uncertainty, P-values were computed using the nonparametric method of DeLong et al. (1988) to assess whether differences in *AUC* could result by chance (i.e., due to limited field data) and not because one model is more efficient than another. The P-values computed by this approach are probabilities that two *AUC* samples could have come from the same distribution. Since this approach requires *AUC* normality, Anderson-Darling and Lilliefors tests (Anderson and Darling 1952; Lilliefors 1967) were used to confirm that all samples came from a normal distribution. P-values were computed to compare each model to all others in the six regional analyses, and in the dataset of CPT case histories. These values, which indicate whether differences in model performance are statistically significant, are presented in Table 5. A significance level of 0.05 was adopted, such that P-values below 0.05 were deemed significant. All else being equal, small P-values can be expected when: (i) differences between two *AUC* values are large; or (ii) the uncertainties of *AUC* values are small; or (iii) distributions have high correlation. Using this criterion, Table 5 compares all model pairs and identifies which is significantly better. The model with better *AUC*, as reported in Table 4, is indicated in Table 5 via the cell shading. If the cell is shaded orange, the model in the left column is better, whereas if the model in the top row is better, the cell is shaded grey. The values given in each cell are the P-values; those less than 0.05 are highlighted via bold font and a red border. Table 5 can thus be used to determine whether differences in model performance, as first presented in Table 4, are statistically significant. Notable observations from Table 5 are: (i) in the initial series of six events at regional scale, the Ensemble model was significantly better than RB20 in two events (Mineral and Ridgecrest), RB20 was significantly better than the Ensemble model in one event (Northridge), and the two models were statistically indifferent in the remaining three (Loma Prieta, Nisqually, and San Simeon); and (ii) in the analysis of CPT case history sites, the Ensemble, ML, and DL models were significantly better than RB20. Collectively, these results suggest that the ML/DL models, which were trained on a modest dataset, predict ground failure with similar or better efficiency as RB20 and thus warrant further application, evaluation, and development. And, as is common in the prediction of ground motions, storm tracks, and other natural hazards, the proposed prediction models could be ensembled with other geospatial liquefaction models, thereby capturing the epistemic uncertainty of model development.

Table 5. [Color] P-value matrix to compare model performance. ML = machine learning model; DL = deep learning model; RB20 = Rashidian and Baise (2020); and ENS = ensemble of ML and DL models, as described in the text.

Statistically Better	↑	Nisqually			Loma Prieta			San Simeon			Northridge			Ridgecrest			Mineral			CPT Case Histories		
←		DL	RB20	ENS	DL	RB20	ENS	DL	RB20	ENS	DL	RB20	ENS	DL	RB20	ENS	DL	RB20	ENS	DL	RB20	ENS
Nisqually	ML	0.484	0.285	0.405																		
	DL		0.221	0.473																		
	RB20			0.334																		
Loma Prieta	ML				**0.004**	**0.002**	**0.000**															
	DL					0.138	0.596															
	RB20						0.278															
San Simeon	ML							**0.000**	0.370	0.341												
	DL								**0.000**	**0.000**												
	RB20									0.321												
Northridge	ML										0.440	0.055	0.613									
	DL											**0.003**	0.279									
	RB20												**0.036**									
Ridgecrest	ML													**0.001**	**0.017**	**0.002**						
	DL														**0.015**	0.341						
	RB20															**0.016**						
Mineral	ML																0.124	**0.001**	0.666			
	DL																	0.396	0.121			
	RB20																		**0.002**			
CPT Case Histories	ML																			**0.009**	**0.000**	0.057
	DL																				**0.000**	**0.004**
	RB20																					**0.000**

*Cell values are the P-values (i.e., probabilities) that *AUC* samples from two prediction models could have come from the same parent distribution (i.e., be statistically indifferent). The model with better *AUC*, as reported in Table 4, is indicated via the cell shading. Values less than 0.05 are deemed "significant" and are highlighted via bold font and a red border.

3.2 Software Implementation

Arguably, a limitation of any ML/DL model is the lack of a defined analytical expression easily ported and executed via hard copy. By corollary, simple interpretations of model structure and form are also generally lacking. While these detractions may be significant to traditionalists, it is clear the use of algorithmic learning will only grow in the field of geotechnics and geohazards, given its demonstrated capabilities when provided with large datasets. It is critical, however, that trained ML/DL models be provided as code, ideally in a format accessible to a broad userbase. Despite this necessity, enumerable ML/DL models have been published without code, meaning that while a model may be available for use by the respective developers, it is not easily accessed by the broader community, and is therefore not readily applied, tested, or improved upon by others.

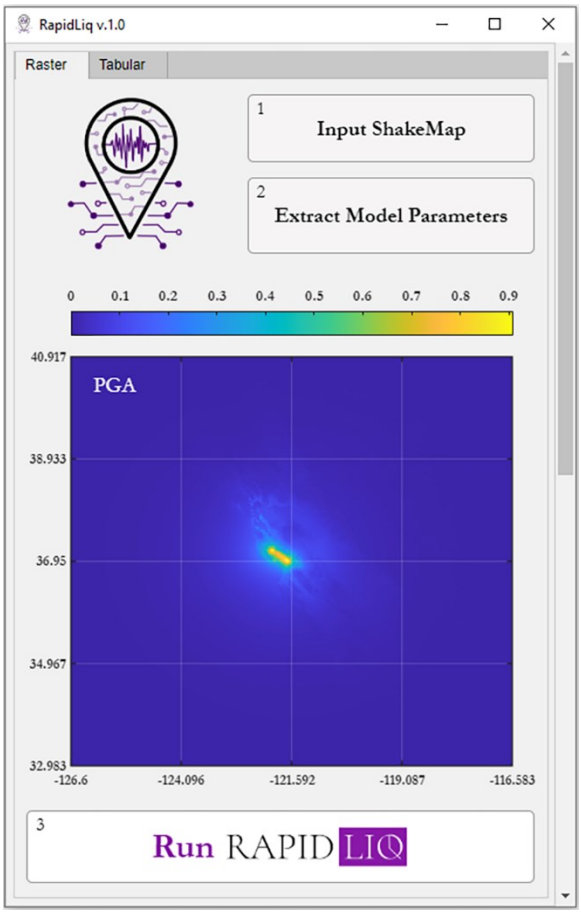

Fig. 12. User interface of *RapidLiq* (Geyin and Maurer, 2021b), which runs the ML, DL, Ensemble, and RB20 models. *RapidLiq* may be downloaded from: https://doi.org/10.17603/ds2-4bka-y039.

To facilitate user adoption and evaluation, the ML, DL, Ensemble, and RB20 models were programmed into *RapidLiq* (Geyin and Maurer, 2021b), a new Windows software program with a simple-to-use interface (Fig. 12). While the Rashidian and Baise (2020) model is widely referenced, it is not commonly implemented

by individual users, owing to the predictor variables that must first be compiled. These variables, and those of the proposed models, are compiled within *RapidLiq*, making user implementation trivial. The only required input is a ShakeMap of ground-motion parameters (i.e., *PGA*, *PGV*, M_w), either in Extensible Markup Language (.xml) or Geotagged Image File (.tiff) format. The first is easily downloaded from the USGS earthquake catalog (https://earthquake.usgs.gov/earthquakes/search/) minutes after an earthquake, or for numerous future scenario events. The second is a more general, flexible format, allowing for motions from various sources to be analyzed. The software then extracts predictor variables across the ShakeMap extents and outputs geotiff files mapping the probabilities of liquefaction-induced ground failure. These files may be viewed within the software or explored in greater detail using GIS or one of many free geotiff web explorers (e.g., http://app.geotiff.io/). The software also allows for tabular input, should a user wish to enter specific sites of interest and ground-motion parameters at those sites, rather than study the regional effects of an earthquake. At present, *RapidLiq* operates in the contiguous U.S. and completes predictions within 10 s for most events.

3.3 Limitations and Uncertainties

The geospatial models developed and tested herein are best suited for regional-scale applications where subsurface testing is infeasible (e.g., disaster simulation and loss estimation; planning and policy development; and emergency response and reconnaissance) or for preliminary site assessment in advance of subsurface testing. While such models have recently been adopted for a variety of uses, they are not intended to guide engineering design and do not replace the need for rigorous site-specific analyses of liquefaction hazard. In this regard, the proposed models predict liquefaction at a relatively coarse spatial resolution, given the resolutions of the geospatial predictor variables (see Table 2), and can thus easily fail to capture more localized, small-scale features that correlate to higher or lower liquefaction hazard.

Inherently, the findings presented herein are tied to the data analyzed. The applicability of these findings to other earthquakes elsewhere – particularly in regions underrepresented in model training – is unknown. Similarly, using the models beyond the range of the predictor variables studied herein (Table 2) could likewise introduce greater uncertainty. In addition, it should be emphasized that "ground failure," the ultimate prediction target, refers to free field liquefaction-induced surface settlement, cracking, and ejecta on ground that is generally level. Users should understand limitations of the *LPI* manifestation model to predict lateral spreading, which is a distinctly damaging expression influenced by complex subsurface and topographic features. Given that *LPI* and other similar manifestation models may be poor predictors of lateral spreading (e.g., Maurer et al., 2015b; Rashidian and Gillins, 2018), the proposed models may likewise predict it poorly. In this regard, the ground-failure datasets on which the models were tested might include lateral spreads, which could have the effect of reducing the measured model efficiency. Moreover, the proposed models do not explicitly predict damage to specific infrastructure types or assets, which would require detailed site and asset-

specific modeling. In this respect, liquefaction could trigger at depth and damage infrastructure without otherwise manifesting or could manifest without otherwise causing damage.

As discussed herein, the performance of any geospatial model is inherently tied to the resolution and accuracy of predictor variables, some of which are themselves predictions rather than measurements (e.g., the depth of groundwater). Inherently, the accuracy of liquefaction predictions is related to the accuracy of inputs, with some models having greater sensitivity to specific inputs. In the present effort, measurement and modeling uncertainties were not considered, and as such, the model outputs should be considered as median probabilities of ground failure. This should not be interpreted to mean that uncertainties do not exist. Among other uncertainties that could be considered in the future, ShakeMap IMs are uncertain; the prediction of *LPI* via geospatial variables is uncertain; and *LPI* is an uncertain predictor of ground failure. In the future, ML/DL techniques (e.g., Gaussian Process Regression) could be used to account for these uncertainties and make probabilistic predictions. Additionally, the most efficient geotechnical models for predicting liquefaction will inevitably change over time. In this regard, the proposed approach could be conditioned on models other than *LPI*, to include emergent mechanistic methods that may better capture the system-level response of soil profiles (e.g., Cubrinovski et al., 2019; Bassal and Boulanger, 2021; Hutabarat and Bray, 2021). For the present moment, the models proposed herein appear to perform as well as, and potentially better than, the current state-of-practice geospatial model (i.e., RB20), but were developed using an altogether different approach, and thus warrant further application. Ultimately, additional tests in past or future events are needed to confirm the findings presented herein and summarized below.

4. Conclusions

This paper presented a new approach to geospatial modeling of soil liquefaction that is driven by algorithmic learning but pinned to a mechanistic framework. In effect, subsurface measurements are predicted remotely within the framing of a popular model for probabilistically predicting ground failure. This merges a body of knowledge built over the last 50 years with the potential of machine and deep learning to predict subsurface conditions remotely. As hypothesized herein, this modeling approach has potential advantages over others used to date. Using this approach, three models termed ML, DL, and Ensemble were trained to predict *LPI* values in the absence of subsurface test data. This training utilized a modest dataset of 1,712 CPTs distributed across the US and a similarly modest set of twelve predictor variables. These consisted of two demand variables (i.e., *PGA* and M_w) and ten capacity variables (i.e., predicted ground water depth, measured distance to river and measured distance to coast, predicted depth to bedrock, measured annual precipitation, predicted V_{S30}, and the predicted (binomial) presence of unconsolidated soil, sandy soil, clayey soil, and silty soil). The capacity variables can be compiled at national scale in advance of model application. The demand variables are available at regional scale minutes after an earthquake or for various future earthquake scenarios.

Lastly, the predicted *LPI* values are transformed to probabilities of ground failure using an existing fragility function trained on all globally available liquefaction case histories.

The developed models were shown to provide efficient predictions in unbiased, forward application and were tested against the RB20 geospatial model. Collectively, these tests indicate that the proposed models warrant application and further evaluation. The proposed and RB20 models are available in *RapidLiq*, a free Windows program. Ultimately, significantly more in-situ geotechnical tests are available for model training, both in the U.S. and globally. Whereas ground-failure inventories are likely to grow slowly, the subsurface data needed to train the proposed approach exist in massive quantities. These data require compilation across different formats (some requiring digitization) and test types (e.g., CPTs and SPTs), as well as access from various entities, both public and private. However, community geotechnical datasets in New Zealand, Austria and Germany, Italy, and Washington State, for example, currently contain more than 40,000 CPTs. Similar datasets are likely to be created elsewhere. Likewise, the quantity of prospective geospatial predictor variables far exceeds that utilized in this study. As additional variables that more efficiently and sufficiently correlate to subsurface conditions are identified, the merits of the proposed modeling approach may be further realized. Expanding upon this approach, improved geospatial liquefaction models could thus potentially be developed for regional, national, or global applications. The limitations of geospatial models (e.g., the uncertainty, spatial resolution, and non-mechanistic nature of geospatial inputs) should be well understood by potential users.

5. Acknowledgements

This study is based on work supported by the National Science Foundation (NSF), US Geological Survey (USGS), Pacific Northwest Transportation Consortium (PacTrans), and Washington State Dept. of Transportation (WSDOT) under Grant Nos. CMMI-1751216, G18AP-00006, 69A3551747110, and T1461-74, respectively. However, any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and may not reflect the views of NSF, USGS, PacTrans, or WSDOT. In addition, we thank Janset Geyin of the Hacettepe University Interior Architecture and Environmental Design Program for their contributions toward design of the *RapidLiq* logos and user interface.

6. Data Availability

All data analyzed in this study is publicly available, as described and referenced in the text. *RapidLiq*, the software program developed to implement the models proposed herein, is available from the DesignSafe data depot at: https://doi.org/10.17603/ds2-4bka-y039.

7. References

Allstadt KE, Thompson EM, Jibson RW, Wald DJ, Hearne M, Hunter EJ, Fee J, Schovanec H, Slosky D, and Haynie KL (2021). The US Geological Survey ground failure product: Near-real-time estimates of earthquake-triggered landslides and liquefaction. *Earthquake Spectra*, 87552930211032685.

610 Anderson TW and Darling DA (1952) Asymptotic theory of certain" goodness of fit" criteria based on stochastic
611 processes. *The annals of mathematical statistics* 23(2): 193-212.
612 Auret L and Aldrich C (2011) Empirical comparison of tree ensemble variable importance measures. *Chemometrics and*
613 *Intelligent Laboratory Systems* 105(2): 157-170.
614 Bassal PC, and Boulanger RW (2021). System response of an interlayered deposit with spatially preferential liquefaction
615 manifestations. *Journal of Geotechnical and Geoenvironmental Engineering*, 147(12): 05021013.
616 Boulanger RW and Idriss IM (2014) *CPT and SPT Based Liquefaction Triggering Procedures, Report No. UCD/CGM-*
617 *14/01*, Center for Geotechnical Modeling, University of California, Davis, CA.
618 Box, GEP and Cox, DR (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B*. 26 (2):
619 211–252.
620 Breiman, L (1996) Bagging predictors. *Machine Learning* 24 (2): 123–140.
621 Cubrinovski M and Ishihara K (1998). State concept and modified elastoplasticity for sand modeling, *Soils and*
622 *Foundations* 38(4): 213-225.
623 Cubrinovski M, Rhodes A, Ntritsos N, and Van Ballegooy S (2019) System response of liquefiable deposits. *Soil*
624 *Dynamics and Earthquake Engineering*, 124: 212-229.
625 DeLong ER, DeLong DM and Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver
626 operating characteristic curves: a nonparametric approach. *Biometrics* 44: 837-845.
627 Durante MG and Rathje EM (2021) An exploration of the use of machine learning to predict lateral spreading. *Earthquake*
628 *Spectra*, 37(4): 2288–2314.
629 DWR (2020) Periodic groundwater measurements. *California Department of Water Resources*,
630 <https://sgma.water.ca.gov/webgis/?appid=SGMADataViewer#gwlevels> last accessed June 2020.
631 Elith J, Leathwick JR, & Hastie T (2008) A working guide to boosted regression trees. *Journal of animal ecology 77*(4):
632 802-813.
633 Fan Y, Li H, Miguez-Macho G (2013) Global patterns of groundwater table depth. *Science* 339 (6122): 940-943,
634 doi:10.1126/science.1229881
635 Fan Y, Li H, Miguez-Macho G (2020) Updated dataset for global patterns of groundwater table depth. <http://thredds-
636 gfnl.usc.es/thredds/catalog/GLOBALWTDFTP/catalog.html> last accessed July 2021.
637 Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861–874.
638 Federal Emergency Management Agency (FEMA) (2013). *Earthquake Model HAZUS-MH 2.1 Technical Manual*,
639 Washington, D.C.
640 Fick SE and Hijmans RJ (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land
641 areas. *International Journal of Climatology* 37(12): 4302-4315.
642 Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics*: 1189-1232.
643 Geyin M and Maurer BW (2020a). Fragility Functions for Liquefaction-Induced Ground Failure. *Journal of Geotechnical*
644 *and Geoenvironmental Engineering* 146(12): 04020142.
645 Geyin M and Maurer BW (2020b) Horizon: CPT-based liquefaction risk assessment and decision software. DesignSafe-
646 CI, doi: 10.17603/ds2-2fky-tm46.
647 Geyin M and Maurer BW (2021a) CPT-Based Liquefaction Case Histories from Global Earthquakes: A Digital Dataset
648 (Version 1). DesignSafe-CI. https://doi.org/10.17603/ds2-wftt-mv37.
649 Geyin M and Maurer BW (2021b) "RapidLiq: Software for Near-Real-Time Prediction of Soil Liquefaction."
650 DesignSafe-CI. https://doi.org/10.17603/ds2-4bka-y039.
651 Geyin M, Baird AJ and Maurer BW (2020) Field assessment of liquefaction prediction models based on geotechnical vs.
652 geospatial data, with lessons for each. *Earthquake Spectra* 36(3): 1386–1411.
653 Geyin M, Maurer BW, Bradley BA, Green RA, and van Ballegooy S. (2021). CPT-based liquefaction case histories
654 compiled from three earthquakes in Canterbury, New Zealand. *Earthquake Spectra*: 10.1177/8755293021996367.
655 Glorot, X. and Yoshua B. (2010) Understanding the difficulty of training deep feedforward neural networks. *Proceedings*
656 *of the thirteenth international conference on artificial intelligence and statistics*: 249–256.

Green RA and Bommer JJ (2019) What is the smallest earthquake magnitude that needs to be considered in assessing liquefaction hazard? *Earthquake Spectra* 35(3): 1441-1464.

Green RA, Lasley S, Carter MW, Munsey JW, Maurer BW, & Tuttle MP (2015) Geotechnical aspects in the epicentral region of the 2011 Mw 5.8 Mineral, Virginia, earthquake. *Geological Society of America, Special Paper 509*: 151-172.

Green RA, Bommer JJ, Rodriguez-Marek A, Maurer BW, Stafford PJ, Edwards B, Kruiver PP, De Lange G. and Van Elk, J, (2019). Addressing limitations in existing 'simplified' liquefaction triggering evaluation procedures: application to induced seismicity in the Groningen gas field. *Bulletin of Earthquake Eng* 17(8): 4539-4557.

Heath DC, Wald DJ, Worden CB, Thompson EM, & Smoczyk GM (2020) A global hybrid VS30 map with a topographic slope–based default and regional map insets. *Earthquake Spectra 36*(3): 1570-1584.

Ho TK (1998) The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8): 832–844.

Holzer TL, Noce TE, & Bennett MJ (2011). Liquefaction probability curves for surficial geologic deposits. *Environmental & Engineering Geoscience 17*(1): 1-21.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79(8): 2554–2558.

Horton JD, San Juan CA, and Stoeser DB (2017) The State Geologic Map Compilation (SGMC) geodatabase of the conterminous United States (ver. 1.1). *USGS Data Series 1052*, 46 p.

Hutabarat D and Bray JD (2021). Effective stress analysis of liquefiable sites to estimate the severity of sediment ejecta. *Journal of Geotechnical and Geoenvironmental Engineering* 147(5): 04021024.

Idriss IM, and Boulanger RW (2008) Soil liquefaction during earthquakes. *Monograph MNO-12* 2008; Earthquake Engineering Research Institute, Oakland, CA, 261 pp.

Iwasaki T, Tatsuoka F, Tokida K, and Yasuda S. (1978) A practical method for assessing soil liquefaction potential based on case studies at various sites in Japan. *2nd Intl Conf. Microzonation*.

Kayen R, Moss RES, Thompson EM, Seed RB, Cetin KO, Der Kiureghian A, Tanaka Y, and Tokimatsu K (2013) Shear-wave velocity–based probabilistic and deterministic assessment of seismic soil liquefaction potential. *Journal of Geotechnical and Geoenvironmental Engineering* 139(3): 407-419.

Kramer SL (2008). *Evaluation of Liquefaction Hazards in Washington State, Report No. WA-RD 668.1*, Washington State Transportation Center, Seattle, Washington.

Lai CG, Conca D, Famà A, Özcebe AG, Zuccolo E, Bozzoni F, Meisina C, Bonì R, Poggi V, and Cosentini RM (2019) Mapping the liquefaction hazard at different geographical scales. In *Earthquake Geotechnical Engineering for Protection and Development of Environment and Constructions*: 691-704, CRC Press.

Lehner B, Verdin K, and Jarvis A (2006) *HydroSHEDS Technical Documentation*. World Wildlife Fund US, Washington, D.C.

Lin A, Wotherspoon L, Bradley B, and Motha J (2021a). Evaluation and modification of geospatial liquefaction models using land damage observational data from the 2010–2011 Canterbury Earthquake Sequence. *Engineering Geology*, 287: 106099.

Lin A, Wotherspoon L, and Motha J (2021b) Evaluation of a geospatial liquefaction model using land damage data from the 2016 Kaikoura earthquake. *Bulletin of the New Zealand Society for Earthquake Engineering. In Review.*

Maurer BW, Green RA, Cubrinovski M, and Bradley BA (2014). Evaluation of the liquefaction potential index for assessing liquefaction hazard in Christchurch, New Zealand. *Journal of Geotechnical and Geoenvironmental Engineering* 140(7): 04014032

Maurer BW, Green RA and Taylor ODS (2015a) Moving towards an improved index for assessing liquefaction hazard: lessons from historical data. *Soils and Foundations* 55(4): 778-787.

Maurer BW, Green RA, Cubrinovski M, and Bradley B (2015b) Assessment of CPT-based methods for liquefaction evaluation in a liquefaction potential index framework. *Géotechnique* 65(5): 328-336**.**

MBIE (2017) Planning and engineering guidance for potentially liquefaction-prone land. New Zealand Ministry of Business, Innovation, and Employment (MBIE), Building System Performance Branch, ISBN 978-1-98-851770-4.

Moss RES, Seed RB, Kayen RE, Stewart JP, Der Kiureghian A, and Cetin KO (2006) CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. *Journal of Geotechnical and Geoenvironmental Engineering* 132(8):1032-1051.

NASA (2020) Distance to nearest coastline. *NASA Ocean Biology Processing Group (OBPG). < https://oceancolor.gsfc.nasa.gov/docs/distfromcoast/> last ccessed July 2021.*

National Research Council (NRC) (2016). *State of the Art and Practice in the Assessment of Earthquake-Induced Soil Liquefaction and its Consequences*, Committee on Earthquake Induced Soil Liquefaction Assessment (Edward Kavazanjian, Jr., Chair, Jose E. Andrade, Kandian "Arul" Arulmoli, Brian F. Atwater, John T. Christian, Russell A. Green, Steven L. Kramer, Lelio Mejia, James K. Mitchell, Ellen Rathje, James R. Rice, and Yumie Wang), The National Academies Press, Washington, DC.

Piryonesi SM & El-Diraby TE (2021) Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. *Journal of Infrastructure Systems* 27(2): 04021005.

Rashidian V and Gillins DT (2018) Modification of the liquefaction potential index to consider the topography in Christchurch, New Zealand. *Engineering Geology* 232: 68-81.

Rashidian V and Baise LG (2020) Regional efficacy of a global geospatial liquefaction model. *Engineering Geology*, 105644.

Rasmussen CE and Williams CKI (2006) *Gaussian Processes for Machine Learning*. MIT Press. Cambridge, Massachusetts, 2006.

Robertson PK and Wride CE (1998) Evaluating cyclic liquefaction potential using cone penetration test. *Canadian Geotechnical Journal* 35(3): 442-459.

Rokach L and Maimon O (2008) *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711.

Schmitt RG, Tanyas H, Jessee MAN, Zhu J, Biegel KM, Allstadt KE, ... & Knudsen KL (2017a) *An open repository of earthquake-triggered ground-failure inventories* (No. 1064). US Geological Survey.

Schmitt RG, Tanyas H, Jessee MAN, Zhu J, Biegel KM, Allstadt KE, ... & Knudsen KL (2017b) *An open repository of earthquake-triggered ground-failure inventories* (ver 2.0, September 2020), US Geological Survey data release collection, accessed Jan 1, 2021, at https://doi.org/10.5066/F7H70DB4.

Shangguan W, Hengl T, de Jesus JM, Yuan H, & Dai Y (2017) Mapping the global depth to bedrock for land surface modeling. *Journal of Advances in Modeling Earth Systems 9*(1): 65-88.

USGS (2021). "Map of CPT Data." *United States Geological Survey. < https://earthquake.usgs.gov/research/cpt/data/> Accessed July 2021.*

Van Ballegooy S, Malan P, Lacrosse V, Jacka ME, Cubrinovski M, Bray JD, O'Rourke TD, Crawford SA, and Cowan H (2014) Assessment of liquefaction-induced land damage for residential Christchurch. *Earthquake Spectra* 30(1): 31-55.

Vapnik V (1995) *The Nature of Statistical Learning Theory.* Springer, New York.

Wald DJ, Worden BC, Quitoriano V, & Pankow KL (2005) ShakeMap manual: technical manual, user's guide, and software guide (No. 12-A1).

Wald DJ and Allen TI (2007). Topographic slope as a proxy for seismic site conditions and amplification. *Bulletin of the Seismological Society of America* 97: 1379–1395.

Wald DJ, Earle PS, Allen TI, Jaiswal K, Porter K, and Hearne M (2008) Development of the U.S. Geological Survey's PAGER system (Prompt Assessment of Global Earthquakes for Response). *Proc. 14th World Conf. Earthq. Eng*., Beijing, China, 8 pp.

Youd TL and Hoose SN (1977) Liquefaction Susceptibility and Geologic Setting. *Proceedings of the Sixth World Conference on Earthquake Engineering* 3, 2189-2194.

Zhu J, Baise LG, and Thompson EM (2017) An updated geospatial liquefaction model for global application. *Bulletin of the Seismological Society of America* 107(3): 1365-1385.

751 Zimmaro P, Nweke CC, Hernandez JL, Hudson KS, Hudson MB, Ahdi SK, ... & Stewart JP (2020) Liquefaction and
752 related ground failure from July 2019 Ridgecrest earthquake sequence. *Bulletin of the Seismological Society of*
753 *America 110*(4): 1549-1566.
754 Ziotopoulou K and Boulanger RW (2016) Plasticity modeling of liquefaction effects under sloping ground and irregular
755 cyclic loading conditions. *Soil Dynamics and Earthquake Eng* 84: 269-283.